AUTOMATIC GENRE CLASSIFICATION OF ENGLISH STUDENTS' ARGUMENTATIVE ESSAYS USING SUPPORT VECTOR MACHINES

by

Sabrina Raaff

Dissertation submitted for the degree

Magister Artium in Linguistics and Literary Theory

at the

North-West University

Supervisor: Prof. A.J. van Rooy
Co-supervisor: Prof. G.B. van Huyssteen

2008
Potchefstroom Campus

ACKNOWLEDGEMENTS

My most sincere thanks go to my supervisor, Bertus van Rooy, for his consistent confidence in me, and understanding. I have really enjoyed working with you. And thanks go to my co-supervisor, Gerhard van Huyssteen, for all his constructive help and suggestions. Thank-you for checking over all my work so thoroughly.

Thank-you to the Linguistics Department at Rhodes University, which is very close to my heart, for my love of Linguistics. Especially, Ian Bekker, for inspiring me to come to the PUK and introducing me to Computational Linguistics.

I am especially grateful to André for his ever-enthusiastic support, for being truly interested in my explanations of machine learning, for staying up into the small hours of the morning and for his unwavering love; ek is verskriklik dankbaar vir alles.

I would like to thank my parents for their support, encouragement, and love in everything I have done. Thank-you for doing all the things for me that I did not have time to do and for providing a warm refuge in times of tiredness and stress.

To my family, you are too numerous to list, and I dare not risk leaving anybody out, you have all helped me in some way, I thank you all.

To my wonderful friends, thank-you: Alé, Aunty M & Colin, Candz, Carly, Claud, Gary, Gen, George, Hortense & Jacques, Megs C., Megs W., Melanie, The Fruitbowl: Hayls, Leapy, Lol, M, (and honorary members) Aniks, Linds, and Mel.

And thanks go to everybody else who inspired me, gave me readings for free and helped me understand all sorts of concepts essential to my research.

Sabrina Raaff

January 2008

ABSTRACT

Automatic text classification refers to the classification of texts according to topic. Similar to text classification is the automatic classification of texts based on stylistic aspect of texts, such as automatic genre classification, where texts are classified according to their genre. This is the classification task that concerns this research project.*

The project seeks to examine the genre of the argumentative essay, in order to develop a genre classifier, using an automatic genre classification approach, which will categorise prototypical and non-prototypical argumentative essays of student writers, into 'good' or 'bad' examples of the genre (binary classification). It is intended that this classifier will allow a senior marker (for example, a lecturer) to give student essays classified 'good' (those that require less feedback and volume of expert correction) to junior markers (for example, teaching assistants). This would afford the senior marker time to pay more attention to essays of a 'poorer' quality.

The corpus used for the research project is comprised of 346 argumentative essays drawn from a section of the British Academic Written English corpus and written by L1 English students. The data are composed of counts of linguistic features extracted from the texts. Once these features were extracted from the texts they were used to create four data sets: a raw data set, composed of raw feature frequencies, a data set composed of the feature set normalised for text length, a data set composed of inverse document frequency counts, and a data set composed of a logarithmic transformation of the feature frequencies. Various classifiers were built making use of these four data sets, using a machine learning approach. In this way, a classifier is trained on previous examples, in order to predict the class of future examples. The project uses support vector machines in STATISTICA's implementation of support vector machines, the STATISTICA Support Vector Machine module (Statsoft, 2006). Support vector machine learning is used because this technique has been shown to perform well in automatic genre classification studies and other classification tasks.

* Please note that *research project* or simply *project* is used solely to refer to the research that this dissertation reports on.

In light of the practical outcome of the project, the classifier's performance is evaluated in terms of the recall of 'bad' examples. The best results were obtained on the classifier built on the text-length normalised data set, using feature selection, a linear kernel and $C = 32$. The recall of 'bad' examples in the test set is 62.5 percent, the recall of 'good' examples in the test set is 74.5 percent, and training accuracy is 62.9 percent.

This study thus shows that argumentative essays can indeed be classified, using an automatic genre approach and that the differences between the prototypical and non-prototypical essays can be fairly adequately extracted, using linguistic features that are easy to compute. Furthermore, the study confirms good performance of support vector machines, especially if many features are used.

Keywords: Automatic Genre Classification/Recognition/Analysis, Automatic Text Classification, Information/Text Retrieval, Corpus Linguistics, Corpora, Computational Linguistics, Automatic Annotation, Machine Learning, Natural Language Processing.

OPSOMMING

Outomatiese teksklassifisering verwys na tematiese teksklassifisering. Dit is soortgelyk aan outomatiese teksklassifisering, gebaseer op stilistiese teksaspekte, soos outomatiese genre-klassifisering, waar tekste volgens genre geklassifiseer word.

Hierdie projek ondersoek die genre van navorsingsopstelle (ondersoekende tekste), ten einde 'n genre-klassifiseerder te ontwikkel, wat, deur van 'n outomatiese teksklassifiseringsbenadering gebruik te maak, prototipiese en nie-prototipiese navorsingsopstelle van studenteskrywers, as 'goeie' en 'swak' voorbeelde van die genre (binêre klassifisering) sal kategoriseer. Die doel is dat sodanige klassifiseerder 'n senior nasiener (byvoorbeeld 'n lektor/lektrise) sal toelaat om studentetekste wat as 'goed' geklassifiseer is (dus min terugvoering en deskundige insette vereis), aan junior nasieners (byvoorbeeld onderwysassistente) toe te vertrou. Die senior nasiener sal sodoende tyd beskikbaar hê om meer intensief aandag aan tekste van 'swakker' kwaliteit te skenk.

Die versameling geskrewe tekste wat vir hierdie projek gebruik is, bestaan uit 346 navorsingsopstelle uit 'n afdeling van die British Academic Written English Corpus, geskryf deur L1 Engelse studente. Die data is saamgestel uit 'n versameling linguistiese kenmerke, wat uit die tekste verkry is. Uit hierdie kenmerke is vervolgens vier stelle data geskep: 'n onverwerkte (rou) stel data, bestaande uit onverwerkte kenmerkfrekwensies; 'n stel data, bestaande uit 'n stel kenmerke, genormaliseer volgens tekslengte; 'n stel data, bestaande uit 'n versameling omgekeerde (teenoorgestelde) dokumentfrekwensies; en 'n stel data, bestaande uit 'n logaritmiese transformasie van die kenmerkfrekwensies. Die vier stelle data is gebruik om verskeie klassifiseerders te ontwikkel deur 'n masjinale (rekenaargebaseerde) leerbenadering gebruik te maak. Op hierdie wyse word die klassifiseerder volgens bestaande voorbeelde geprogrammeer, ten einde die klassifisering van toekomstige voorbeelde te voorspel. Hierdie projek maak gebruik van die ondersteuningsvektormasjien in STATISTICA se implementering van ondersteuningsvektormasjiene, naamlik die STATISTICA Ondersteunings-vektormasjienmodule (Statsoft, 2006). Die ondersteuningsvektormasjien leerproses is gebruik, aangesien hierdie tegniek reeds goeie resultate in outomatiese genre-klassifisering, asook ander klassifiseringstake gelewer het. In die lig van hierdie

praktiese uitkoms(te) van die projek word die klassifiseerder se prestasie ooreenkomstig die herroeping van 'swak' voorbeelde beoordeel. Die beste resultate is deur die klassifiseerder gelewer, wat met behulp van die genormaliseerde tekslengte datastel ontwikkel is, deur gebruik te maak van kenmerkseleksie, 'n liniêre kern en C=32. Die herroeping van 'swak' voorbeelde in die toetsstel is 62.5 persent, dié van 'goeie' voorbeelde, 74.5 persent en programmeringsakkuraatheid, 62.9 persent.

Hierdie studie bewys dat navorsingsopstelle inderdaad deur 'n outomatiese genrebenadering geklassifiseer kan word en dat die verskille tussen prototipiese en nie-prototipiese tekste redelik voldoende uit die maklik-rekenariseerbare linguistiese kenmerke geïdentifiseer kan word. Die studie bevestig verder goeie prestasie deur ondersteunigsvektormasjiene, indien van 'n verskeidenheid kenmerke gebruik gemaak word.

Sleutelwoorde: Outomatiese Genre-klassifisering/Erkenning/Analise, Outomatiese Teksklassifisering, Inligtings-/Teksherwinning, Versameling(s) van Linguistiese Tekste (Corpus/Corpora), Rekenaarlinguistiek, Outomatiese Annotasie, Masjinale (Rekenaargebaseerde) Leer, Natuurliketaalprosessering.

PREFACE

When I first began this project, I had taken only a few elementary courses in Computational Linguistics and otherwise had a solid linguistic background, but no knowledge or experience of Mathematics or Natural Language Processing. I, therefore, had a huge amount of catching up to do and simply drowned in the literature for a very long time.

In order to help me along the way, and to introduce me to concepts in context, rather than the decontextualised dictionary definitions I was reading up on at the time, I took an undergraduate Mathematics course for a semester. I also learnt how to use Linux, some Perl and came to love regular expressions.

Often, when I was stuck on something in a reading I found that it was because the authors assumed their readers knew as much as they did. I found this a major stumbling block especially as very few authors provided references to concepts that there was either no space to explain or which they assumed were known. As a result, I have tried to explain all concepts that may be alien to Linguists or else provided references in footnotes to background information and concepts that are important but not explained. In this way, this dissertation is written for Linguists with no computational or mathematical background, but the approach of this dissertation also addresses it to Computational Linguists.

For anyone who is 'lost in the literature', I recommend as a starting point an undergraduate course in Mathematics (to learn about matrices and vectors, complex numbers are also good to learn), Michael Oakes's (1998) *Statistics for corpus linguistics*, Neil Salkind's (2004) *Statistics for people who (think they) hate statistics*, and Tony Rietveld and Roeland van Hout's (2005) *Statistics in language research: analysis of variance*.

I would like to thank Amelia Nkosapantsi, Attie de Lange, Elsa van Tonder, Teresa Smit, and Wannie Carstens for all their support during my research period. As well as the National Research Foundation, without whom none of this research would have been funded.

Many thanks also go to Annick Griebenouw for showing me how to use SVMTool when I was still a Linux and Perl infant, and everybody at the Centre for Text Technology, especially Charlene Gentle, Jacques McDermid Heyns, Martin Puttkammer, and Suléne Pilon who provided lots of help, and patient explanations (and free software!).

Thank-you very much, to the wonderful ladies at the library, particularly Gerda van Rooyen for starting me off on the information hunt.

I am deeply grateful to the BAWE team for their data and helpful attitude, principally Siân Alsop and Jasper Holmes. Without them, I would not have had ANY data!

A huge thank-you to Jesus Giménez, for educating me about feature sets and encoding.

Thank-you also to Sarel Steele for giving me advice and dissertations when all other information sources ran dry.

Thanks also go to Liz Greyling for editing this dissertation.

And finally, thank-you so much to Google, without 'whom' I simply would never have managed to have the prerequisites to understanding the prerequisites of support vector machines.

Sabrina Raaff

January 2008

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

CHAPTER 1
Introduction

Composition, rhetoric and argumentation have traditionally played a key role in Western education, and argumentation in particular is valued highly for its associations with the concept of logical thinking, proofs and refutations
(English, 1999:17)

1.1 Introduction

Researchers distinguish between text categorisation and text classification (Jackson & Moulinier, 2002:119), where text categorisation generally refers to document sorting by content and topic[1] (Manning & Schütze, 1999:575), and text classification to any document classification not necessarily based on content, such as classification by author. In the literature, however, this distinction is normally established through explanation rather than terminology. In this research project, the automatic classification of texts according to topic is termed automatic text classification. Similar to text classification is the automatic classification of texts based on some stylistic aspect of texts; examples of such stylistic classification are authorship attribution studies (Mosteller & Wallace, 1964) and genre classification. This project is concerned with this latter classification, based on genre, which is referred to as automatic genre classification. *Genre* is defined at length in Chapter 2; in summary, it refers to a class of communicative events, in which the participants share some communicative purpose(s). This common purpose determines the discursive structure, style and content of a genre. Exemplar members of a genre thus demonstrate patterns of similarity with regard to structure, style, content and intended audience (Swales, 1990:58).

Automatic genre classification has its niche in effective webpage searching, seeking to create a web-search engine augmented with a genre identification module (Kwasnik, Crowston, Nilan, & Roussinov, 2000). Ultimately, users could specify genre type according to their information needs, which would ensure higher precision in retrieval and higher relevancy to the user. Automatic genre classification is thus largely associated with information retrieval. As a result, it is mainly used to distinguish between web-specific genres, such as FAQs (frequently asked questions), where the

[1] In this project *topic*, *subject*, and *domain* are used as synonyms.

corpora are collected from the web by the researchers. It has, however, also been applied with varying success to the classification of traditional[2] genres, which are generally drawn from existing corpora, such as the Lancaster-Oslo-Bergen[3] Corpus of British English (Johansson, Leech & Goodluck, 1978). For the most part, these studies are concerned with English corpora, but some work has been undertaken in German, Greek, Korean, Russian and Swedish (see for example Wastholm, Kusma, & Megyesi, 2005; also Stamatatos, Fakotakis, & Kokkinakis, 2000a). These studies vary widely in application, features and training methods, but have in common determining the best features for classifying genres. It is with these studies, which seek to automatically classify traditional genres, that this research project is concerned.

Chapter 1 serves as an introduction to the automatic genre classification task of this research project. The chapter commences with a brief overview of the research area and the contextualisation of this research project, in Section 1.2. Next, the problem statement is described in Section 1.3. Then the research questions are provided in Section 1.4. The research aims of this project follow in Section 1.5. Thereafter, the statement of the central hypotheses is presented in Section 1.6. Then an overview of the methodology is provided in Section 1.7. Finally, the chapter concludes with the chapter outline showing the structure of this dissertation, in Section 1.8.

1.2 Overview of the research area and contextualisation

A distinction is drawn between the genre classification of web-specific and traditional genres in Section 1.1. This project is concerned with traditional genres in particular. Automatic genre classification studies that are concerned with traditional genres are characterised by three main concerns: corpus, features and learning methodology.

The corpus provides the texts that are to be classified as well as the data, which are the basis of classification. The data are in the form of feature frequencies extracted from the texts. This feature set is predetermined based on the hypothesised characteristics of a particular genre and also on features that have been found useful in other studies. The

[2] These genres are also referred to as *paper* or *print genres*. These terms are not used here as they place too much emphasis on medium.
[3] Known as the LOB corpus.

choice of these features must be undertaken with care, in order to reduce the likelihood of many irrelevant features. This is because, in general, many learning methodologies are adversely affected by too many irrelevant features. In Chapter 4, it will be seen that this is something that the learning technique used in this project is fairly robust to.

The features are mainly of two types: lemmas (and sometimes, word-forms) and 'linguistic' features. The former set is more traditionally used in automatic text classification studies and is known as the bag-of-words (BOW) approach. The approach is hypothesised as and sometimes found to be less useful in automatic genre classification studies as it is too topic specific (Finn, 2002:75). That this is so is the main concern of many studies, which aim to show that 'linguistic' features are rather more useful in distinguishing genres and should rather be used. The word 'linguistic' is used because it is not altogether clear that the BOW approach is 'unlinguistic'. Indeed, it is evident in several studies (Argamon & Dodick, 2004a; also Santini, 2005b) that both the feature sets can be used well together, and moreover, that the BOW approach can be used successfully, but with more careful word selection.

Once the features have been determined, they are extracted from the texts using various techniques, for example, regular expressions to match search terms and extract frequency counts. Then the hypothesised differences between genre classes are sometimes explored in terms of these feature counts. This is done, in order to estimate a more accurate idea of the discriminatory ability of particular features and thus, their subsequent use in training a genre classifier. Such intermediate exploration of features is also useful in the removal of irrelevant features before training.

Once the final feature set has been determined, a classifier is developed on a training set from which it learns to make future classifications. Phrased statistically, the classifier is trained on multiple independent variables (the features), which predict the dependent variable (the genre class). The classifier is then tested on a set of texts that it has not observed before. It is desirable to achieve high accuracy on the training set because based on this accuracy it can be deduced that the features used adequately extract the differences between particular genre classes. Additionally, it is desirable to achieve high accuracy on the test set because this shows how well the classifier performs on unseen

data and how well it can be expected to perform on new test sets. This is referred to as the classifier's *generalisation ability*. If the classifier is to be re-usable it must have a high generalisation ability. An accuracy that is too high on the training set can, however, indicate that, although the training set can be perfectly or near perfectly classified, the classifier is too attuned to the idiosyncrasies of the training set. As a result, it fits the training set very well but does not generalise well. This situation is referred to as *overfitting* because the classifier overfits the training set. The opposite, and equally undesirable situation is called *underfitting*, as the classifier underfits the training set.

Various learning methodologies are used, such as factor analysis (Biber, 1988), discriminant analysis (Karlgren & Cutting, 1994), *k*-nearest neighbour (Wolters & Kirsten, 1999), multiple regression (Stamatatos, Fakotakis, & Kokkinakis, 2000b), logistic regression (Boese, 2005), decision-tree learning (Finn, 2002), Naïve Bayes (Santini, 2004a), and support vector machines (Argamon & Dodick, 2004a). These techniques are all explained and reviewed in context, in Chapter 2.

1.3 Problem statement

The norms and expectations of academic discourse introduce concerns for both students and lecturers. These concerns revolve around elucidating precisely what the attributes of academic writing and the discourse ideals are, for informing assessment and for informing student academic writing; the former is the main concern of this project. Academic writing is regarded as an essential means of communication in tertiary education and determines students' success at tertiary institutions. Hyland (2004a:5) defines successful academic writing as "the ability of writers to offer a credible representation of themselves and their work, by claiming solidarity with readers, evaluating their material and acknowledging alternative views". There are three main approaches to academic writing: the skills-based approach, the acculturation approach and the practice-based approach (Lea & Street, 1997, cited in Lea & Street, 2000). The first approach considers that there is a set of skills applicable to all academic disciplines, and that can be learned and transferred to any academic context (Lea & Street, 2000:34). The second approach Lea and Street (1997, cited in Lea & Street, 2000) term the "academic socialisation approach", in which the task of the lecturer is viewed as one

of socialising students into a new 'culture' (Lea & Street, 2000:34). The third approach is referred to as the "academic literacies approach" (Lea & Street, 1997, cited in Lea & Street, 2000). In contrast to the first two approaches, the third approach does more than merely acknowledge disciplinary and departmental differences in academic literacy practices. This approach views academic institutions as sites of "discourse and power" (Lea & Street, 2000:35) and academic literacies as social practices. It thus views academic literacy as encompassing a variety of communicative practices, which includes different fields and genres. Furthermore, it sees each communicative practice in context where social meanings and identities are evoked.

This research project takes this third approach in that it acknowledges the differences between communicative practices, in particular those of various academic genres. Moreover, this project views student writing, as academic writing, in terms of meaning-making and ideological conflicts (Davidson & Tomic, 1999; Turner, 1999; also Ivanič, Clark & Rimmershaw, 2000). As a result (as will be shown in Section 2.2) 'good' and 'bad' examples of the argumentative essay are labelled as such not because they indicate skills or a deficit of skills but rather because they are not in keeping with the discourse ideals of the gatekeepers, as can be seen from the grade awarded them.

This research project seeks to examine the genre of the argumentative essay. These essays are written by students within an academic context. According to Van de Poel (2006:17), a particular academic context, in which academic writing takes place is constructed from:

(A) a limited repertoire of text genres;
(B) an author who is defined as an academic in some way, e.g. a lecturer or student;
(C) a main goal that is to render a point of view about an academic topic;
(D) an objective and argumentative way of writing; and
(E) a set of conventions regarding referencing and layout.

Furthermore, an academic text bears the following characteristics (Van de Poel, 2006:18):

(A) It is well embedded in an academic context.
(B) Its point of departure is a thesis or a research question.
(C) It intends to persuade the ideal audience.
(D) It delivers the author's personal view with respect to the central tenet of the text.

(E) It is written by an author who is not necessarily made prominent.
(F) It contains standardised formal characteristics.

This research project contests the last point as it implies that all academic texts have the same formal characteristics. The project extracts various formal characteristics of one genre of academic writing, in order to determine whether linguistic features can be used to classify texts, even within one genre.

Texts representative of this genre, argumentative essays, serve to confirm or reject a thesis statement, or to persuade the reader of the writer's point of view, and as such are defined as instances of argumentative writing (Van den Poel, 2006:75). This ability to argue based on facts and examples, reason and consequence, authority, subjective judgement and deliberation of pro and cons (Van den Poel, 2006:80) is considered valuable in Western education (English, 1999:17). Therefore, it is essential that students learn to argue in writing, in order to succeed in many academic discourse communities. Intuitively, it follows that evaluative feedback plays an important role in acquiring this knowledge.

An automated feedback system would provide an opportunity for lecturers to provide more detailed feedback in a shorter period of time. A starting point of this type of evaluation is an automated means of determining the standard of students' essays. To this end, a program that can analyse the presence of features indicative of proficient academic writing would provide a means for lecturers to pay more attention and time to students who struggle in their writing. An example of a program with similar aims is Trushkina (2006), which automatically detects lower-level language errors in L2 English learners' argumentative essays, in order to allow lecturers time to focus on higher-level phenomena (see also Louw, 2006). In addition, such a program could inform lecturers as to the particular attributes of the genre that require attention on the part of both the learner and lecturer.

This research project is mainly concerned with the former goal of such a program, but also sheds some light on features of the genre at hand. Such a task is one of binary classification, where the essays are grouped into two classes: 'good' or 'bad' examples of the argumentative essay genre. Essays that output 'good' were considered indicative

of a student who has successfully acquired the norms of academic writing within the genre of argumentative essays.

Such a system would separate essays needing less feedback and volume of expert correction (classified 'good') from those needing more attention (classified 'bad'). In this way, the system would allow a senior marker (for example, a lecturer) to give student essays classified as 'good' examples of the genre to junior markers (for example, teaching assistants). This would afford the senior marker time to pay more attention to essays of a 'poorer' quality. This classifier could even be biased to classify texts as 'bad', rather than 'good', in cases of uncertainty to ensure that essays labelled 'bad' examples are not given to junior markers who may not be able to provide the kind or volume of feedback required.

Determining the approach to this type of feedback system requires some reframing of the problem at hand. This involves putting forth some hypotheses regarding the nature of the classification task. These hypotheses are detailed in Section 1.6. The approach this research project takes, is one of automatic genre classification. Major studies in this field (to be reviewed in Chapter 2) reveal that this approach has not been applied to so subtle a genre class as argumentative essays.

Much of the work in pedagogy within corpus linguistics compares non-native speaker corpora with native speaker corpora, in order to compare patterns of use of lexis and grammatical structures (Flowerdew, 2002:98). Examples are Granger and Rayson (1998), who compare word frequency profiles from the International Corpus of Learner English (ICLE), a corpus of argumentative essay writing by advanced non-native learners, to a control corpus from the Louvain Corpus of Native English Essays (LOCNESS), as well as Hyland and Milton (1997), who investigated native speaker and non-native speaker high school students' argumentative academic writing in terms of expression of doubt and certainty.

This project, however, does not seek to compare the differences in argumentation between non-native speakers and native speakers but rather to determine 'good' and 'poor' examples of argumentative essays within a group. The project is concerned with

the argumentative writing of native speakers, because then it is likely that there will be fewer minor errors (spelling and morphological errors, see Trushkina, 2006:155), making it easier to extract linguistic information, such as part-of-speech (POS) tags, which rely on correct language structure to achieve high accuracy (such errors characterise, for example, the Tswana Learner English Corpus compiled at North-West University, South Africa).

In addition to the automatic genre classification approach, there are other natural language processing approaches to the problem addressed by this research project (see for example, Teufel & Moens, 1999; also Buckingham Shum, Uren, Li, Domingue & Motta, 2002). Two relevant examples of such an approach are Moreale and Vargas-Vera's (2003) automated argument extraction tool, and Burstein, Marcu, Andreyev, and Chodorow's (2001) thesis statement classifier. Moreale and Vargas-Vera's (2003) automated argument extraction tool (similar to this research project) is concerned with argumentation in students' essays. This tool is not of a classificatory nature, rather it seeks to categorise and highlight argumentative strategies in students' essays. Thus, the output of the tool is intended to assist students in evaluating their own work (formative) and as a supplementary tool for marking (summative). Burstein, Marcu, Andreyev, and Chodorow's (2001) thesis statement classifier seeks to identify the thesis statement in essays. Unlike Moreale and Vargas-Vera's (2003) tool, this classifier is not an end-product, but the creators suggest that the features of a particular essay's thesis statement could be of evaluative use to the writer of the essay (Burstein *et al*, 2001:98).

As argumentation in academic writing is valued, it determines students' success at tertiary institutions. It is thus essential for students to learn to argue in writing, in order to succeed in many academic discourse communities. Feedback plays an important role in acquiring this knowledge. This project aims to develop a classifier that will ease the workload of senior markers, in order to allow them additional marking time, thereby allowing them to provide higher quality feedback.

1.4 Research questions

The following research questions arise from the preceding discussion:

1. What are the most discriminating linguistic features between 'good' and 'bad' examples of the argumentative essay genre?

2. Can these linguistic features be easily computed and extracted?

3. Can an automatic genre classification approach be used to develop a classifier, which will categorise prototypical and non-prototypical argumentative essays of student writers, into 'good' or 'bad' examples of the genre?

4. Will support vector machines (SVMs), as a machine learning technique, provide good generalisability, especially across domains, while requiring the least amount of human effort?

1.5 Research aims

In response to the research questions, this project aims to:

1. Establish the most discriminating features between 'good' and 'bad' examples of the argumentative essay.

2. Determine whether these features can be easily computed and extracted.

3. Develop a classifier using an automatic genre classification approach, which will categorise prototypical and non-prototypical argumentative essays of student writers, into two classes: 'good' or 'bad' examples of the genre.

4. Determine whether SVMs will provide good generalisability, especially across domains, while requiring the least amount of human effort.

1.6 Statement of the central hypotheses

This research project posits several hypotheses regarding the approach towards the classification task. According to Grabe and Biber (1987, cited in Biber, 1988:204), student essays use the surface form of academic prose, but are relatively non-informational and extremely persuasive. They therefore deduced that student essays "do not have a well-defined discourse norm in English" (Grabe & Biber, 1987, cited in Biber, 1988:204).

1. The first hypothesis is in reaction to this. It is hypothesised that there *are* computationally extractable differences between argumentative essays in a higher-grade band ('good') and those in a lower-grade band ('bad') that can be used to predict the classes of new essays.

2. Second, these class differences can be adequately represented by linguistic features that are easy to compute and extract.

3. Third, the differences between essays that place them in a higher- or lower-grade band are indicative of the prototypicality of the essays; therefore, 'good' essays can be viewed as prototypical and 'bad' essays as non-prototypical.

4. Fourth, this prototypicality of argumentative essays can be extended to a genre class so that this classification task can be viewed as one of genre classification. In this case, 'good' essays are instances of the argumentative essay genre, while 'bad' essays, although still examples of the genre are poor instances of the genre.

5. Accordingly, it is then hypothesised that previous automatic genre classification studies can be used to inform this project in terms of features and methodology. This is supported by the fact that the feature set in major automatic genre classification studies remains fairly constant; possibly because they have their origins in Biber's (1988) language variation study (see Chapter 2). This is a positive indicator for this project as it implies that features that have worked well in other projects can be used with some confidence in this research project. It can also be deduced that other aspects of these studies can be used to guide this research project, such as evaluation metrics.

1.7 Overview of methodology

Initially, a literature review of the field of automatic genre classification was conducted, in order to determine the standards of practice in terms of applications, corpora and genres, learning techniques, features, and evaluation metrics. Thereafter, nine main steps were followed, in order to develop the classifier. The **first step** was to select the machine learning technique (the algorithm). The best machine learning technique was determined by the literature review on automatic genre classification as well as a review of machine learning. The **second step** was to identify and acquire a corpus from which to extract the features. The **third step** was to choose the features that were to be used. The features were chosen based on the literature review of automatic genre classification and two well-known grammar books: Biber, Johansson, Leech, Conrad, and Finegan (1999); and Quirk, Greenbaum, Leech, and Svartvik (1985). The **fourth step** was to prepare the texts before the features were extracted. This preparation included the removal of formatting, essay questions, essay titles, bibliography, appendices, headings, footnotes, graphs, illustrations, tables, some of the punctuation and equations. It further entailed character set conversion,[4] the standardisation of apostrophes and quotation marks, and tokenisation. The **fifth step** was to mark-up sentences, paragraphs, quotations, references, punctuation marks, nominalisations, two- and three-word complex prepositions, two-word adverbs, and multi-word conjuncts using XML tags, with part-of-speech (POS) tags. The **sixth step** was to extract the features using STATISTICA's text miner, STATISTICA Text Mining and Document Retrieval module (Statsoft, 2006). The **seventh step** was to standardise the essays' grades (the dependent variable), to remove multiple occurrences of features from the data set (data cleaning) and to transform the data in three ways. This step also entailed reducing the feature set using feature selection tests. In order to determine which feature selection tests to use the features were assessed for normality using four descriptive methods. The **eighth step** was to train the classifier using STATISTICA SVM (Statsoft, 2006).[5] The **final step** was to test the SVM classifier. The process of developing the classifier is illustrated in Figure 1.1 below.

[4] The texts were converted from Unicode to ISO/IEC 8859-1.
[5] This tool will be described in Section 3.7.

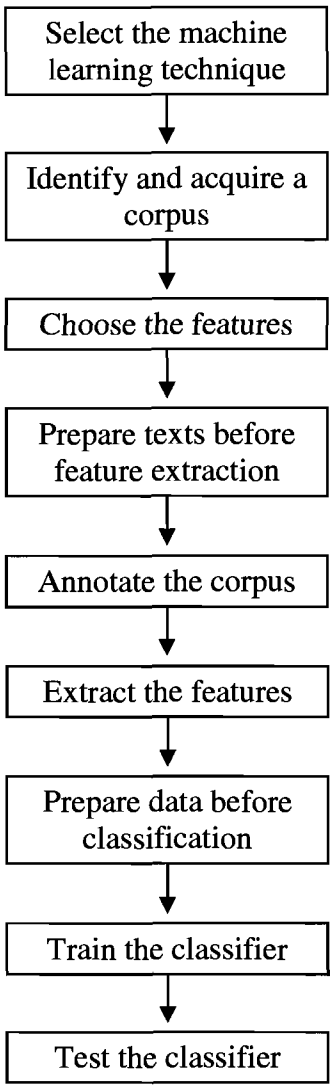

Figure 1.1: Process of developing the classifier

This project made use of SVMs for classification not only because this technique has shown good performance in a variety of pattern classification problems (Burges, 1998:121, see also Schölkopf & Smola, 2002:22), but also because it has been shown to have good performance in automatic genre classification studies (this is discussed in Chapter 2). It thus seems reasonable to assume that SVM learning is one technique that can be expected to perform well for the problem posed by this research project.

In addition to selecting a technique that is not necessarily generally the best technique for various problems, but at least one of the better techniques for this problem, it is also important to determine how good performance is to be determined and measured (Hand, 1997:3), that is, what is meant by *good* for this project? In light of the practical outcome

of this classification project, as reviewed in Section 1.3, the classifier's performance is evaluated in terms of the recall of 'bad' examples. This metric provides a measure of the number of 'bad' examples that are correctly labelled 'bad' (see Chapter 4 for a detailed review of various evaluation metrics). Measuring performance by the number of 'good' examples that are correctly labelled 'good' is not as important, because 'good' essays being incorrectly labelled 'bad' would not be as detrimental as 'bad' essays being incorrectly labelled 'good'.

1.8 Chapter outline

In Chapter 2, the notions of machine learning and supervised learning will be defined, and the basic notation for the machine learning process used in this project is introduced. These concepts will be placed in the framework of automatic genre classification and the task of genre classification for this project further explained. Next, the use of *genre* in this project will be defined with particular reference to the genre of this project. Thereafter, a review of automatic genre classification will be presented, in order to detail the background to the features and methods used in this research project. Where possible, comparisons will be made between these projects, and each statistical technique used in these studies explained. Furthermore, these studies will be critically assessed in terms of the validity of pre-defined genre classes, results, evaluation measures and the features used for genre extraction. This will be done to determine the potential value of features and methodology for application to this research project. The literature review will first review seminal works in the field. Thereafter, contemporary automatic genre classification studies will be reviewed with detailed reference to projects that are relevant to this research project, with regard to application, corpus, features, or method. Finally, studies that use SVMs for the purposes of genre classification will be reviewed.

Chapter 3 will provide detailed background on the data and learning methodology used to develop the genre classifier. This chapter will discuss all the features deemed potentially relevant as good predictors of prototypical or non-prototypical examples of argumentative essays. Thereafter, text preparation before feature extraction will be detailed. Next, the annotation of the features will be described. Thereafter, data

preparation before classification will be detailed. Lastly, SVMs for the linearly separable and non-linearly separable case will be presented.[6]

Chapter 4 will detail how the data were used in training the SVM classifier, and the method used in the selection of the training parameters. The potential data concerns of imbalanced data sets, differing misclassification costs, and the normal distribution assumption of the data set will be raised and addressed in terms of this research project. Next, various evaluation indicators and metrics will be presented, and the most suitable accuracy measure for this project will be discussed. Thereafter, the results of the various classifiers built on different data and feature sets, using C- and ν-SV classification, and two kernels will be reported. This chapter will also address various hypotheses, some of which are raised in Chapter 3. Finally, the best classifier's performance will be analysed and seven potential reasons put forth for the results.

In Chapter 5, the dissertation will be concluded with a summary of the preceding chapters. Furthermore, the results and findings of this study will be reviewed, with reference to the hypotheses postulated in Chapters 3 and 4. Thereafter, recommendations for future research will be made.

1.9 Summary

This chapter provided an overview of the background to this research project. First, an introduction to automatic genre classification and the research area was provided. Thereafter, the problem statement was described and the central hypothesis stated. Then, the research questions arising from the problem statement and the corresponding research aims of this project were delineated. Next, an overview of the methodology of this research project was outlined. Finally, the chapter outline showing the structure of the following sections was sketched.

Chapter 2 will define *genre* and present a review of major studies in the field of automatic genre classification.

[6] SVMs for data that can be separated in space by a line, and data that cannot be separated by a line.

CHAPTER 2

Literature review

The word [genre] is highly attractive — even to the Parisian timbre of its normal pronunciation — but extremely slippery
(Swales, 1990:33)

2.1 Introduction

Machine learning is a technique used to 'teach' a program (referred to as the learner) the features of the classes it must learn to classify. The goal is for the program to be able to extend the 'knowledge' gained by training to unseen data, in order to classify this new data into the classes defined during training. Vapnik (2000:19–20) represents this kind of learning by a "model of learning from examples". This model is illustrated in Figure 2.1 below; where G is the generator of the data, S is the target operator or supervisor's operator, and LM is the learning machine.

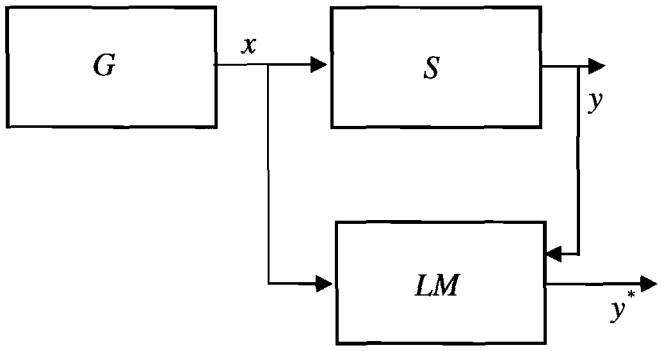

Figure 2.1: A model of learning from examples (Vapnik, 2000:20)

During learning, the LM observes the training set, pairs (x, y). Once the LM has been trained it must be able to return a value y^* for any given x. It is intended that such a y^* value approximates S's y response. For this genre classification task x represents the essay and y the classification label, 'good' (example) or 'bad' (example).

Naturally, in order for a learner to learn, there must be a 'teacher'. 'Teaching' is referred to as supervision. There are differing degrees of supervision, ranging from supervised to unsupervised learning, in which the amount of human intervention involved is minimal.

Essentially, supervision refers to the degree and types of annotation of data, as well as the amount of information (in the form of instructions) the computer is given regarding the classification task; such as what data must be classified, how the data must be classified and into which classes the data must be classified.

Machine learning can, of course, be used for many other learning problems that do not require explicit classification as an end-product. For the purposes of this research project, however, the introduction provided above depicts the type of machine learning this project is concerned with; classification of students' essays into prototypical ('good') and non-prototypical ('bad') examples of argumentative essays.

This chapter defines such automatic genre classification in Section 2.2. This section does not provide a detailed overview of the different uses of *genre*, but rather outlines the background to the definition of *genre* assumed in this study, and further elucidates what is meant by *genre*. It should be noted that not all researchers in the field of automatic genre classification assume the same definition of *genre*. This is further clarified in Section 2.3, which provides an overview of the state-of-the-art in automatic genre classification, with particular emphasis on genre classification studies that make use of SVMs. This section reviews the features, methods, and results of the automatic genre classification systems of previous work in the field. It also briefly discusses the types of problems that can be solved, using the approach and techniques of automatic genre classification.

2.2 Defining genre

Originally, *genre* referred to a kind of picture, which depicted a scene from ordinary domestic life, and became extended in usage to refer to classes of articles (Swales, 1990:33). An overview of the term's development in folklore, literary studies, and rhetoric is provided by Swales (1990, see also Hyland, 2004b:25–50). This section, however, is concerned with the use of the term in linguistics. This usage is similar to the meaning Swales (1990) intends when using the term *genre*, as does the ethnographer, Saville-Troike (1982) who lists *greetings*, *lectures* and *jokes* as some examples of genre types (Swales, 1990:39).

The use of *genre* adopted by this project follows that of Swales and other Hallidayean linguists. In order to explain this use of *genre* some reference must be made to *register*. *Register* is analysed according to field, tenor and mode. Field refers to the content and type of activity involved; tenor refers to the role, relationships, and status of the participants; and mode refers to the channel of communication (Swales, 1990:40). Collectively, field, tenor and mode act as "determinants of the text through their specification of register" (Halliday, 1978:122).

In this way, according to Martin (1985), genres are realised through registers, and registers themselves are realised through language (see also Lee, 2001:46). Martin (1985) provides similar examples to those of Saville-Troike of genre types: *lectures*, *seminars*, *poems*, *narratives* and *manuals* (Martin, 1985:250). *Genre* thus determines the way field, tenor and mode can be combined in any linguistic situation, in any particular culture (Swales, 1990:41). This last remark is important as genre types are not the same in all cultures. Therefore, deconstructing the norms of genre types can be helpful for cross-cultural awareness and education of, for example, students learning the rules and structure of argumentative essays.

Furthermore, Martin's (1985) view of *genre* leads to an analysis of discourse structure, which looks at the beginning, middle and ending of a text. These stages of development also separate *register* from *genre* in that *register* can be identified at the sentence-level, whereas *genre* can only be realised in completed texts. Accordingly, *genre* determines "the conditions for beginning, continuing and ending a text" (Couture, 1986:82). As examples of *genre*, Couture (1986:87) offers the *research report* and *business report*, and as examples of *register*, the language of scientific reporting and the language of newspaper reporting. In the case of this study, the register being used (or rather the target register) is the language of academic writing and the genre of the argumentative essay.

Genres and registers are often complementary and, according to Couture (1986:86), successful textual communication may require demonstration of the appropriate relationship between the genre and register systems. In this research project, it is assumed that for the students to acquire a 'good' mark for their essays, they will need to

demonstrate their acquisition of the norms of the language of academic writing, and the structural rules of the genre: the argumentative essay.

The usefulness of genre analysis and classification has at times been questioned, and accused of leading to "heavy prescription and slavish imitation" (Swales, 1990:38). After reviewing the attitudes towards and the use of *genre* in the disciplines of folklore, literary studies, linguistics and rhetoric, Swales (1990) demonstrates some commonalities in the stance of academics in these disciplines. From this, he deduces that contrary to what he terms "ancient misapprehensions" (Swales, 1990:37), genre theory can indeed be useful for educating students without resorting to "narrow prescriptivism" (Swales, 1990:45). Moreover, educating students about genres can illuminate reflections upon linguistic and rhetoric choices for students as writers, rather than deny them such opportunities of choice in structuring their writing (Swales, 1990:45).

In attempting to establish a working definition of *genre*, Swales discusses genre membership. This leads to the questioning of what it is that determines membership of any particular genre. He proposes two ways of determining the answer to this: the definitional approach and family-resemblance approach (Swales, 1990:49). The definitional approach requires drawing up a limited set of simple properties that would define all and only the members a particular genre from anything else (Swales, 1990:49). He provides many counters, with examples, to this approach, which will not be detailed here, the essence of which is that this approach is often difficult to accomplish in the case of genre types.

The next approach, family resemblance, is concerned with similarities and relationships between members of a group as opposed to a set of limited properties. The family-resemblance approach, as proposed by Wittgenstein (1953:31), lead to prototype theory. Prototype theory is associated with Rosch (1975); it examines members of classes along a continuum of least typical to most typical. The member that is established as most typical is the prototype of that class. In terms of this project this means that although the essays are instances of the argumentative essay genre, not all are typical members. Rather, some essays are most typical members and thus, characterise the genre the most,

while other essays are least typical (peripheral) members. This is illustrated in Figure 2.2, below.

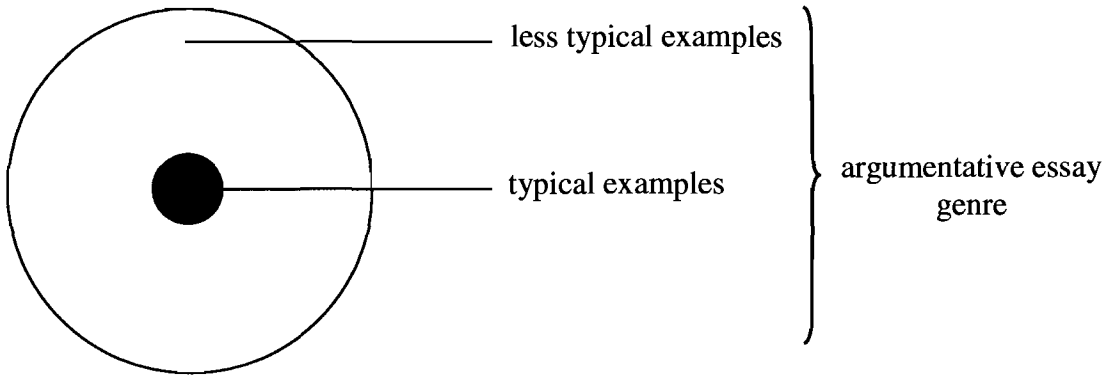

Figure 2.2: Framing the genre classification task in terms of prototype

After establishing how genre type membership is determined, Swales turns to a short, but considered definition of *genre*, which is adopted in this research project: *genre* refers to a class of communicative events, in which the participants share some communicative purpose(s). This amounts to the rationale of the *genre*, which determines the discursive structure, style and content. Exemplar members of a genre demonstrate patterns of similarity, with regard to structure, style, content and intended audience (Swales, 1990:58). Such exemplars are generally considered prototypical by members of the discourse community.

As mentioned earlier, the task of this project is to label texts as prototypical and non-prototypical instances of the genre of argumentative essays. Determining which essays are prototypical and which are not and in addition, which features make prototypical essays prototypical, is not as unbiased as it may seem. This is because prototypicality is determined by the discourse community, in this case the markers of the essays. The writers of the essays are still in training to become members of the discourse community — Swales (1990:53) suggests the term "apprentice members" — and in this discourse community, similar to many others, there are gatekeepers. The educators and markers (often the same people) seek to teach the students the norms of the discourse community, and therefore, essentially help preserve these norms and keep the non-compliant out (by giving their essays 'poor' marks). It therefore seems reasonable to assume that 'good' marks are indicative of prototypical essays and *vice versa*.

According to Swales (1990:52), communicative purpose, form, structure and audience are properties that determine the prototypicality of a member of a particular genre. It can therefore be argued that argumentative essays written within the sphere of academic discourse can be viewed as a genre type based on similarity of communicative purpose and audience. It is then assumed that this similarity must hold for form and structure too. The features relating to the form and structure of the students' essays used in this project are thus very important in classifying the texts.

The features of such prototypical essays can, undoubtedly, be determined through detailed micro-linguistic analysis. Such detailed analysis, however, would not suit the purposes of this project, which seeks to make fairly quick classifications (even if some accuracies must be lost). Yet, determining (prescribing) the features of the genre *a priori* would potentially limit the accuracy of such classification. Therefore, a large selection of linguistic features is used to classify texts as prototypical or non-prototypical. The background to these features is discussed in Section 2.3 below, in a comprehensive review of the features and methods used in automatic genre classification, and the features used to classify the texts in the corpus will be discussed in detail in Section 3.3.

2.3 Overview of previous automatic genre classification studies

This section reviews previous relevant work in the field of automatic genre classification, and aims to provide a review of the state-of-the-art in automatic genre classification studies.[7] First, background work relating to genre classification and seminal works in the field are discussed; thereafter research projects that are relevant and have a similar purpose to this research project are reviewed in more detail. Finally, those studies that use SVMs for machine learning, for the purposes of genre classification, are discussed in detail.

In the main, each study is discussed separately, because there is much variation between data, features, application and rationale of each project. It should be noted that as this

[7] The most recent review of the state-of-the-art in this field was conducted by Santini (2004b). It is a very complete review but has a strong interest in web-specific classification, thus it tends to refer only briefly to studies relating to more traditional genres.

project is only concerned with linguistic features of genres, genre classification of web-specific genres is for the most part not discussed as many of the features used, relate to layout and HTML encoding. Moreover, such genre classification studies are not directly relevant to this project as they have an entirely different purpose; they mainly classify electronic genres that are very different from traditional genres for purposes of information retrieval in web searches. They also often address matters particularly relevant to electronic genres such as genre evolution.[8]

2.3.1 Overview of seminal works in genre classification studies

Biber's (1988) seminal work provides the background to genre classification studies and has become a classic in this field (see Johannesson & Wallström, 1999, for a study that makes direct use of Biber's features). Moreover, it has influenced the Expert Advisory Group on Language Engineering Standards' guidelines on text typology (EAGLES, 1996:23–25).

His work (Biber, 1988) in language variation sought to determine the dimensions upon which spoken and written varieties differ linguistically. The data he used were drawn from two corpora: the LOB Corpus of British English and the London-Lund Corpus of Spoken English (Svartvik & Quirk, 1980). In order to compensate for the lack of non-published written texts in the corpora, an additional collection of personal and professional letters were added to the two corpora.

In order to establish the underlying dimensions upon which these varieties differ, he analysed twenty-three spoken and written 'genres' (Biber, 1988). Such a wide variety of 'genres' was covered in an attempt to make use of data that cover the complete range of situational variation. It should be noted that he makes use of the groupings already used in the corpora, and does not create his own additional groupings, this would seem to imply that he agrees with the labelling of such groupings as 'genre'. Indeed, he goes on to define his use of *genre*, using it to refer to "text categorisations made on the basis of external criteria relating to author/speaker purpose" (Biber, 1988:68). Furthermore, he

[8] See for example, Santini (2005a); Crowston and Kwasnik (2004); Shepard, Waters and Kennedy (2004); Rehm (2002); and Roussinov, Crowston, Nilan, Kwasnik, Cai, and Liu (2001).

considers *text-type* to refer to texts grouped according to similarity in linguistic form (Biber, 1998:70).

It has already been established in Section 2.2 that form and communicative purpose are both considered *genre*-defining in this project. Thus Biber's (1988) referring to these groupings of texts, as 'genres' is not quite what is meant here. For example, the grouping *biographies* would be considered a genre type according to the meaning this research project assumes, but many of the other groupings: *religion*, *academic prose*, and *humour*, would not. Biber was, however, not actually seeking to classify genre types in his study, but rather, as previously mentioned, to determine the linguistic variations between spoken and written varieties. Therefore, the relevance of his use of *genre* is less important. Rather, it is his methodology and linguistic features that were deemed essential to informing this project.

Biber (1988:71–72) made use of sixty-seven linguistic features in his study, which were identified from a survey on previous studies of spoken and written variation. Similar to the present study, he selected the largest possible range of potentially salient features and made no *a priori* decisions regarding their importance. He grouped these features into sixteen grammatical categories (Biber, 1988:72):

(A) tense and aspect markers;
(B) place and time adverbials;
(C) pronouns and pro-verbs;
(D) questions;
(E) nominal forms;
(F) passives;
(G) stative forms;
(H) subordination features;
(I) prepositional phrases, adjectives, and adverbs;
(J) lexical specificity;
(K) lexical classes;
(L) modals;
(M) specialised verb classes;
(N) reduced forms and dispreferred structures;
(O) coordination; and
(P) negation.

He then determined the frequencies of each of these linguistic features in all the genres, in order to study co-occurrence patterns among the features. These co-occurrence patterns indicate functions or dimensions underlying the variation between varieties. In

order to establish the underlying dimensions, Biber (1988) made use of factor analysis. This type of multivariate statistical analysis derives a reduced set of variables from a large set of original variables. In this case, the original variables were the frequencies of the linguistic features, which were reduced to a set of factors. Thus, each factor represents a group of linguistic features that had a high frequency of co-occurrence.

In factor analysis, first the correlations between all features are established and displayed in a matrix (correlation matrix). Second, the size of the correlations are compared, for example, a large negative correlation indicates that the presence of the first variable correlates with the absence of the second variable. Similarly, for a large positive correlation, the presence of the first variable correlates with the presence of the second variable. The correlation coefficient, if squared, indicates the statistical significance of the relationship between variables by measuring the percentage of variance between them. This procedure is described in more detail by Biber (1988:80–97). Using this technique, he (1988:115) determined seven factors:

1. informational versus involved production;
2. narrative versus non-narrative concerns;
3. explicit versus situation-dependent reference;
4. overt expression of persuasion;
5. abstract versus non-abstract information;
6. on-line informational elaboration; and
7. factor 7, indicating academic hedging, but unlabelled due to under-representation.

The factors established by Biber (1988) are not used in this research project. Their mention is relevant, however, because they represent much of the variation between genres, albeit with the focus on written and spoken varieties. Moreover, because these factors were determined using the sixty-seven features, mentioned above, it is plausible to assume that the features can be potentially useful discriminators for the current project. Indeed, many of the features that will be discussed in Section 3.3 are derived from Biber (1988). This feature set used by Biber (1988) has been substantially enlarged in recent work; for example, Biber, Conrad, Reppen, Byrd, Helt, Clark, Cortes, Csomay and Urzua (2004). Several of the features used in this later work are also used in this project, as will be discussed in Chapter 3.

Karlgren and Cutting (1994) took Biber's work (1988; 1989) as a starting point for their research. They make use of similar features to those of Biber's study (1988; 1989), paying more attention to those that can be (readily) automatically computed. They made use of frequency counts of the following features (Karlgren & Cutting, 1994:1072):

A) Parts-of-speech
   1. nouns;
   2. present participles;
   3. present tense verbs;
   4. prepositions;
   5. adverbs;
   6. first person pronouns; and
   7. second person pronouns.

B) Lexical words
   1. *it*;
   2. *me*;
   3. *that*;
   4. *therefore*; and
   5. *which*.

C) Ratios and lexical information
   1. average number of words per sentence;
   2. average number of characters per sentence;
   3. type/token ratio;
   4. average number of characters per word;
   5. total sentence count;
   6. total character count; and
   7. long words (longer than six characters).

The frequencies were computed for each of the texts, which were taken from the Brown University Corpus of Written American English (Francis & Kučera, 1982). Karlgren and Cutting (1994) then made use of discriminant analysis on these texts using the computed frequencies. This type of statistical analysis determines a set of discriminating functions, which can discriminate (to varying degrees of accuracy) between the classes

under examination. Once these functions have been established, new and unseen texts can be categorised according to their feature counts.

Karlgren and Cutting (1994) conducted three classification experiments, and similar to Biber (1988), they did so using the categories provided by the corpus. The categories have three levels: the first level is comprised of the classes *informative* and *imaginative*, and the second level is made up of the classes *press*, *fiction*, *non-fiction* and *miscellaneous*. The third level has fifteen groups, for example, *reportage*, *popular lore*, and *skills and hobbies* (Karlgren & Cutting, 1994:1071). As LOB is modelled on the Brown Corpus, the groupings at level three are the same.

The three experiments were conducted using the different groupings of each level. The experiments entailed training (the reason for this is explained in Karlgren & Cutting, 1994:1073). Categorisation at the first level resulted in correct classification of 478 cases out of 500 (approximately 96 percent). At the second level grouping, 366 out of 500 cases were correctly classified (approximately 73 percent). The texts that proved the most difficult to classify, were from the subset labelled *learned/humanities*, which were largely misclassified as *miscellaneous* instead of *non-fiction*. These texts were instances of academic prose written within disciplines in the humanities faculty (Karlgren & Cutting, 1994:1072). This finding is relevant to this project, which seeks to classify texts drawn from various disciplines across most of the faculties (see Section 3.2 for actual faculties). Thus, this project is set quite a challenge.

As the subset *learned/humanities* is relevant to this project some reasons are posited for its misclassification. The reasons are difficult to determine. Possibly, the four categories in level two are somewhat meaningless, and should therefore not be used for groupings; or else, the four level-three categories under *miscellaneous* (*religion*, *skills and hobbies*, *popular law*, *belles lettres*) are similar in feature frequency counts to those in *learned/humanities*. It seems unlikely that the texts in these five groups are similar in structure, perhaps then, more (either in number or in ability) discriminating features need to be used.[9]

[9] See Sigley (1997) for more on corpora categories and their validity.

That this is a likely explanation is given further credence when regarding the classification results at the third level: 258 texts out of 500 were correctly classified (52 percent). And again, the group with the lowest accuracy level was that of *learned/humanities*. The texts in this group were misclassified chiefly as *religion* and *belles lettres*, which are grouped under *miscellaneous* (Karlgren & Cutting, 1994:1073).

The 'genre' classes used in Karlgren and Cutting's (1994) study, similar to that of Biber's (1988), indicate a dissimilar definition of *genre* to what is intended by the term in this research project. Firstly, the labels given for level one are indicative of rhetorical classes rather than genre types. Secondly, of the categories at level two, only *press,* perhaps, may be considered a discourse type, while *fiction* and *non-fiction* are categories based on content rather than genre, and *miscellaneous* is void of any meaning as a category at any level. This somewhat contrived classification of texts may be an explanation for the poor classification results for the fifteen categories in experiment three (see also Kessler, Nunberg & Schütze, 1997:33).

In contrast with Biber (1988) and Karlgren and Cutting (1994), Kessler *et al.* (1997) use *genre* to refer to texts that are grouped according to similar communication purpose, which in turn is connected to the formal properties of the texts. Accordingly, they term attributes indicative of genre type *generic facets* (Kessler *et al.*, 1997:33). These facets, they explain, refer to the practical function and communication (often indicative of rhetorical strategies) of texts in a class, and are associated with a set of linguistic properties, which they term *generic cues* (Kessler *et al.*, 1997:33).

They claim three advantages for generic facets. Firstly, these facets provide a framework for understanding genres. Secondly, some applications requiring text classification, especially those in an information retrieval context, may find categorisation according to facet rather than genre advantageous. Thirdly, using facets as part of the genre classification solution rather than genre type classes, will allow for labelling an unknown genre category in terms of its facets. Classifiers that are trained to recognise genre classes only will misclassify any unknown genre type.

As previously mentioned, Kessler *et al.*, (1997:34) assert that these facets can be identified by *generic cues*, which are observable surface cues. In their study, they make use of fifty-five generic cues, represented by four groups: structural, lexical, character-level, and derivative cues. Structural cues are passives, nominalisations, syntactic categories of words and topicalised sentences. Examples of lexical cues are terms of address and words used to express dates. Character-level cues are punctuation marks, such as question marks and exclamation marks, capitalised and hyphenated words, and acronyms. Derivative cues are variation measures (such as standard deviation) and ratios derived from counts of lexical and character-level cues, such as characters per word. The ratios were combined (to form nearly three-thousand different ratios) and represented as natural logarithms. For example (Kessler *et al.*, 1997:34):

$$\alpha \log \frac{W+1}{S+1} + \beta \log \frac{C+1}{W+1} + \gamma \log \frac{W+1}{T+1} = (\alpha - \beta + \gamma)\log(W+1) - \alpha \log(S+1) + \beta \log(C+1) - \gamma \log(T+1), \quad (2.1)$$

where:

W = tokens,

S = sentences,

C = characters, and

T = tokens.

Kessler *et al.* (1997), similar to Karlgren and Cutting (1994), make use of the Brown Corpus. In contrast to Biber (1988) and Karlgren and Cutting (1994), however, they make their own class distinctions, using three generic facets, which they label *brow*, *narrative* and *genre*. *Brow* relates to the intellectual background of the intended reader, and has the levels of *popular*, *middle*, *upper-middle* and *high*. *Narrative* is a binary facet indicating whether the text is written in a narrative style or not. The *genre* facet has the values of *reportage*, *editorial*, *scitech*, *legal*, *non-fiction* and *fiction* (Kessler *et al.*, 1997:34).

For their study, Kessler *et al.* (1997) made use of logistic regression and neural networks. The first classification method they used, logistic regression, is a statistical technique in which the transformed values of the dependent variable (in this case, the facet) are predicted by a linear combination of the predictor variables (in this case, the

cues). The transformation of the dependent variable is called the link function; the logit transformation is used for logistic regression (for more on this see Statsoft, 2004). The second classification method that Kessler *et al.* (1997) used, neural networks, is a type of machine learning technique that was originally motivated by biological learning systems (for computational details, see Mitchell, 1997:81–127).

For the logistic regression method, accuracies of 78 percent for the narrative facet, 61 percent for the genre facet and 44 percent for the brow facet were obtained. The neural networks method performed better overall, with accuracies of 82 percent for the narrative facet, 75 percent for the genre facet and 47 percent for the brow facet. The detailed results for each subcategory, for both logistic regression and neural networks can be viewed in Kessler *et al.* (1997:37). It is relevant to note that lower accuracies were obtained for *scitech*, *non-fiction*, *editorial* and *legal* than for *reportage* and *fiction*.

Kessler *et al.* (1997:36) suggest several possible reasons for lower accuracies. Sparse training examples are put forth as an explanation for poorer performance on the *legal* and *scitech* classes. For the *non-fiction* and *editorial* classes, lower accuracy was, in the main, a result of *non-fiction* being misclassified as *editorial*. As was remarked earlier, *non-fiction* is a category based on content, not genre. Moreover, *non-fiction* is at a higher level of analysis than *editorial*, which is a genre type. It seems then, that *non-fiction* is an unwise choice as a genre category. Indeed, Kessler *et al.* (1997:36) propose making *editorial* a subcategory of *non-fiction* in future studies. They also suggest further decomposition into facets labelled *opinion* and *institutional author*.

The above three studies (Biber, 1988; Karlgren & Cutting, 1994 and Kessler *et al.*, 1997) form the background to current studies in the field of automatic genre classification. The next section discusses contemporary genre classification studies that are relevant to this research project.

2.3.2 Overview of contemporary genre classification studies

This section reviews major studies in automatic genre classification that are relevant to this research project, with regard to application, corpus, features, and method. The studies reviewed here are presented in chronological order, in order to trace the

improvements on features and findings by different studies at different times on the same features. It should be noted, however, that equal comparisons between studies can often not be made due to a lack of standards in automatic genre classification, regarding a benchmark corpus for training and testing, evaluation metrics, baseline performance or indeed, training methodology. Therefore, each study is discussed separately, and similarities between studies at any level are highlighted.

In their study, Wolters and Kirsten (1999) make use of content words, function words, POS frequencies[10] tagged with the German Stuttgart-Tübingen Tagset (STTS) of 54 tags and punctuation. They used a German corpus, LIMAS, which was modelled on the Brown Corpus and therefore has similar 'genre' classes (Wolters & Kirsten, 1999:143).

They provide in-depth details on feature sets and feature distributions, which they analysed before undertaking any classification tests, in order to determine if indeed the documents in their corpus differed according to their groupings (Wolters & Kirsten, 1999:144–145). What is most noteworthy is that they used decision trees as an exploratory device (and not for classification purposes). In Chapter 3, it will be shown that pre-classification exploration of features can help remove irrelevant features before training. In genre classification, decision trees are one of the most used classification techniques (see for example Finn, 2002; and Dewdney, VanEss-Dykema & MacMillan, 2001). Wolters and Kirsten (1999) use three *k*-nearest neighbour algorithms (KNN): RIBL,[11] learning vector quantisation (LVQ)[12]and IBL1(-IG).[13]

The KNN algorithm is a type of instance-based learning in which training examples are stored and used to classify new instances. This technique, unlike many machine learning techniques, can construct a different approximation to the target function for each new query, rather than attempting to approximate the target function for the hypothesis space only once (Mitchell, 1997:230–231). In order to classify new instances, the KNN algorithm assigns a particular classification to a new instance, based on the classification of its nearest neighbours, defined by standard Euclidean distance (Mitchell, 1997:231–232).

[10] The corpus was tagged with the MALAGA system. See Beutel (1998) for more on the system.
[11] See Emde and Wettschereek (1996) for more on RIBL.
[12] For more on learning vector quantisation see Kohonen, Kangas, Laaksonen, and Torkkola (1996).
[13] See Daelemans, Van den Bosch and Weijters (1997) for more on IBL1.

Wolters and Kirsten (1999) test their algorithms using different feature sets, composed of various combinations of content words, function words, POS and punctuation. Their combining of the various features, in order to determine each feature set's usefulness for genre classification, encouraged the approach to feature combination used in this research project.

Their results (Wolters & Kirsten, 1999) are somewhat difficult to interpret for two reasons. Firstly, they conduct tests on classes that appear to be more subject than genre-related. They claim that they realise this distinction and are examining both genres (*press texts* and *fiction*) and domains; however, from the two 'genre'-only categories they provide this is not clear. Secondly, they report on precision and recall only, on particular tasks, which they do not specify in sufficient detail. Nevertheless, they appear to have excellent results (using 10-fold cross-validation).[14] The most relevant results for this research project are on their task A, which contains forty-five academic texts drawn from the fields of humanities (H), and science and technology (S). For these academic texts, they achieve an average recall of 99.67 percent and precision of 100 percent on their test set, using content words (500 lemmas) and POS tags, training with the LVQ algorithm (Wolters & Kirsten, 1999:147). These results must be interpreted with caution because it appears that genre and topic classification are thrown together here. Genre and topic are not necessarily orthogonal to one another. In terms of classification, some overlap is therefore to be expected. It is not the overlap that is problematic here, but rather the lack of clear definition and identification of the target of classification. A clear definition of *genre* and the genre classes to be classified is required for results to be interpreted clearly and the findings of the study to be extended to other studies. The poor definition of *genre* and genre classes is a critique emphasised throughout this project.

In addition to Wolters and Kirsten's (1999) results, they report on classification on a larger data set, with differing results (they report only on precision in this experiment, using LVQ). Task H (109 documents) achieves its lowest precision at 19.6 percent and its highest precision at 100 percent; task S (72 documents) achieves its lowest precision

[14] For an explanation of cross-validation see Chapter 4.

at 17.3 percent and its highest precision at 74.4 percent. This experiment thus confirms that POS tags are good discriminators for academic texts.

In view of tasks H and S, Wolters and Kirsten (1999) provide strong evidence that different word types (function and content words, versus function words only) are not equally discriminatory for texts from the humanities and those representing the field of science and technology. This implies that cross-domain classification may not be as successful as task A may indicate. As they do not expound on the genre categories in task A, it is rather difficult to judge whether their classifier performed good genre classification across domains. But certainly, if they classify across domains, this can be favourably compared to Finn's (2002, see below) cross-domain classification accuracies, which are lower. Thus, it would appear that cross-domain genre classification has been successful to varying degrees. This is a factor that this research project considers and will report on in Chapter 4.

In contrast to the traditional BOW approach to text classification tasks, Stamatatos *et al.* (2000a), make use of the frequencies of the most used words not in the training corpus but rather in a corpus representative of the written language of English: the British National Corpus (BNC; Aston & Burnard, 1998). In addition to these frequencies, they make use of the most frequent punctuation marks: full stop, comma, colon, semicolon, quotation marks, round brackets, question mark, and hyphen (Stamatatos *et al*, 2000a:812). They drew their data from a section of Wall Street Journal (WSJ): *editorials* (40 documents), *letters to the editor* (40 documents), *reportage* (40 documents), and *spot news* (40 documents).[15]

Following Karlgren and Cutting (1994), Stamatatos *et al* (2000a) make use of discriminant analysis for their genre classifier. They report on accuracies based on a 50 percent training set and 50 percent testing set, achieving 97.5 percent, using the top thirty words from their list derived from the BNC (Stamatatos *et al*, 2000a:811). Their findings indicate that using the top fifty words leads to overfitting. Their results inform the features and results of this project, which are reported in Chapters 3 and 4 respectively. Stamatatos *et al.* (2000a:811) provide further evidence that the traditional

[15] Documents containing *What's News* or *Who's News* (Stamatatos *et al.*, 2000a:810).

BOW approach does not discriminate as successfully in genre classification as in topic text classification.

Combining words from their BNC list and the top eight punctuation marks results in even better accuracy; depending on the number of words used, they can achieve 100 percent accuracy (Stamatatos *et al.*, 2000a:812). Moreover, this result stays far more stable when training with much smaller sets of data. This would seem to imply that punctuation is a good discriminator of genre and can thus be reliably used with corpora of limited size.

Stamatatos, Fakotakis, and Kokkinakis (2000b) make use of twenty-two features derived from a text-processing tool for Modern Greek, called the sentence and chunk boundaries detector (SCBD). The features are grouped according to token level from the output of the sentence boundary detector, phrase level based on the output of the chunk boundary detector and analysis level from the text analysis of SCBD. These features are given below (Stamatatos *et al.*, 2000b:477–478):

(A) Token level

1. detected sentences/words;
2. punctuation marks/words; and
3. detected sentences/potential sentence boundaries.

(B) Phrase level

1. detected NPs (noun phrases)/total detected chunks;
2. detected VPs (verb phrases)/total detected chunks;
3. detected APs (adverbial phrases)/total detected chunks;
4. detected PPs (prepositional phrases)/total detected chunks;
5. detected CONs (conjunctions)/total detected chunks;
6. words included in NPs/detected NPs;
7. words included in VPs/detected VPs;
8. words included in APs/detected APs;
9. words included in PPs/detected PPs; and
10. words included in CONs/detected CONs.

(C) Analysis level

1. detected keywords/words;

2. special words/words;
3. assigned morphological descriptions/words;
4. chunks' morphological descriptions/total detected chunks;
5. words remaining unanalysed after pass 1/words;
6. words remaining unanalysed after pass 2/words;
7. words remaining unanalysed after pass 3/words;
8. words remaining unanalysed after pass 4/words; and
9. words remaining unanalysed after pass 5/words.

Using the twenty-two features listed above, Stamatatos *et al.* (2000b) trained their classifier on a corpus of Modern Greek (250 documents, equally divided into ten genres), which they compiled themselves from Internet sources. The corpus is comprised of the genres of *press editorial*, *press reportage*, *academic prose*, *official documents*, *literature*, *recipes*, *curricula vitae*, *interviews*, *planned speeches*, and scripted *broadcast news* (Stamatatos *et al.*, 2000b:481). It is clear from this division that although they provide some criticism of the corpora typically used in genre classification experiments, such as the Brown Corpus[16], for its poorly defined genres, some of their genre groupings are also fuzzy. They are blurred both in terms of what is meant by *genre* in this project and in terms of their definition of genre (Stamatatos *et al.*, 2000b:480). The most relevant example of this is *academic prose*. This project considers *academic prose* a style of writing associated with different genres written in an academic context; *recipes*, in comparison, is what is meant by genre here.

Making use of the twenty-two-variable feature vectors for each document Stamatatos *et al.* (2000b:479–480) train their classifier, using the statistical techniques of discriminant analysis and multiple regression. The latter technique is used to predict the value of a dependent variable from two or more independent variables, with each independent variable being assigned relative weight. This weight is determined by the relative contribution of each independent variable in determining the value of the dependent variable. Using the mean values of the variables, standard deviation and correlation coefficients of the variables can be derived, in order to determine the slope of a straight line that will pass through the majority of the data points used. This line is known as the

[16] See Section 2.3.1 for a discussion of these genres, in the review of Karlgren and Cutting (1994).

regression line and it approximates to the data, allowing new data points to be classified (Oakes, 1998:33–36).

Stamatatos *et al.* (2000b) report accuracies based on a 50 percent training set and a 50 percent testing set, which results in ten documents per genre for training and ten documents per genre for testing. Overall accuracy on both discriminant analysis and multiple regression is reported as an identification error rate of 0.18, which is the number of incorrectly classified texts divided by the total number of texts (Stamatatos *et al.*, 2000b:482–483). The most significant result for this research project is presented in a confusion matrix of genre classification over all ten genres, showing that the error rate for *academic prose* is zero. This success, however, must be interpreted cautiously as the classifier is tested on only ten documents, representing a rather small sample, and therefore it is likely that the sample may not be representative of the population. This may result in random error when attempting to generalise this model to other data (Biber, 1993a:219–220). Indeed, the learning curve Stamatatos *et al.* (2000b:491) present of seven to fifteen documents indicates that the classification accuracy does indeed improve marginally (from 0.18 to 0.15 identification error rate) with more training data.

Biber (1993b:243) states that linguistic representativeness depends on sample size among other factors, such as a thorough definition of what the target is. He presents an equation that can be used to determine the required sample size for single features (Biber, 1993b:253). This equation shows that the required sample size is a function of the standard deviation relative to the mean of a feature (Biber, 1993b:254). In his example, the feature with the highest mean score is nouns, which is the most frequent word class (Biber, Johansson, Leech, Conrad & Finegan, 1999:65). The sample size required for this class is 59.8 texts. For any features that are less frequent than this a larger sample size will be required. This would seem to indicate that sample size cannot be determined but that ten texts are likely to be too few to be linguistically representative. The problem of training sample size is discussed in more detail in Chapter 4.

In addition to this successful classification of *academic prose*, the discriminatory ability of individual features of Stamatatos *et al.*'s (2000b) overall genre classifier is significant. They determine the overall contribution of each feature to the classification of their genres using *t*-tests, and find that the most significant indicators of genre are punctuation marks/words, words included in PPs/detected PPs and special words/words (Stamatatos *et al.*, 2000b:491). Again, this must be interpreted with caution as *t*-tests assume normal distribution, which is often not the case with linguistic data (Karlgren, 1999:153; also Santini, 2004a:3). Rather non-parametric tests should be used, such as the Mann-Whitney *U* rank sum test (Oakes, 1998:11). Normal distribution, parametric and non-parametric tests will be discussed further in Chapter 3.

The former of the two facets suggested by Kessler *et al.* (1997:36), in Section 2.3.1, *opinion*, is one that Finn (2002) applied to his own study.[17] He does not explicitly use the term *generic facet*, but rather *genre class*. Such a term can be misleading, as it appears to refer to *genre* types. It is not immediately apparent but it seems that he uses the term to refer to aspects of a document by which it can be classified in terms of genre, rather than referring to *genre* types (see also Santini, 2006a for commentary on this). Finn (2002:3) cites the readability and level of technical detail of a document as examples.

Finn's study (2002:11–15) draws texts from the web in the form of news articles (796 documents from three domains), and movie and restaurant reviews (1354 documents from two domains).[18] Details of the websites from which these documents originated, and the pre-classification of such documents can be found in Finn (2002:69–71).

His study focuses on the genre classes of *opinion/fact* and *positive/negative*. The first genre class, *opinion/fact*, is a subjectivity classification and is studied by investigating whether news articles report facts or the author's opinion. Finn (2002) mentions that this is a distinction often made within the discourse of press reportage, and furthermore that documents offering the author's opinion are often editorials. The latter remark is relevant to the suggestion provided by Kessler *et al.* (1997:36), as discussed earlier, with regard to their classifier's poor performance on *editorials*. The second genre class,

[17] See also Finn, Kushmerick and Smith (2002).
[18] The corpus used by Finn (2002) is available for free download at http://www.smi.ucd.ie/hyppia.html.

positive/negative, seeks to establish whether movie and restaurant reviews are positive or negative.

In addition to classifying texts according to these genre classes, Finn (2002) examines the accuracy of such classification across subject domains. All the studies discussed above make use of corpora drawn from a variety of subject domains, but none test or at least report any statistics on the classification accuracy across domains. Finn (2002) explores classification in the domains of football, politics, finance, movies, and restaurants by training the classifier on one domain and testing in another, thereby determining the cross-domain accuracy. This focus on cross-domain accuracies is relevant to this research project where all the data are drawn from and classified across domains.

Finn's (2002) classifier was trained using three feature sets: BOW, POS and a set of features relating to lexical information, which Finn (2002:15) terms "hand-crafted text statistics". The hand-crafted (HC) set is composed of the following (Finn, 2002:107–108):

A) Lexical information
   1. average number of sentences beginning with *I*;
   2. average sentence length;
   3. average word length;
   4. long words (longer than five characters); and
   5. number of words in document.

B) Lexical words
   1. stopwords:[19] *certainly*, *her*, *highly*, *him*, *his*, *it*, *large*, *little*, *me*, *mine*, *mostly*, *much*, *my*, *our*, *probably*, *that*, *they*, *us*, *very*, *we*, *where*, *which*, *you*, *yours*;
   2. keywords: *absolutely*, *altogether*, *archive*, *article*, *column*, *completely*, *doesn't*, *editorial*, *enormously*, *entirely*, *extremely*, *fact*, *feature*, *fully*, *greatly*, *highly*, *intensely*, *isn't*, *news*, *opinion*, *perfectly*, *report*, *strongly*, *thoroughly*, *totally*, *utterly*, and *very*.

C) Punctuation symbols: *! " $ % & ' ( ) * + , - . : ; = .* and *?*.

[19] Stopwords are usually excluded in text classification that takes a BOW approach because of their high frequency.

The above set of features was constructed specifically for the *opinion/fact* genre class and was gathered by asking people for their intuitions on good indicators of this class. Such an approach seems rather 'unlinguistic', which indeed it is, but it must be considered in light of the purpose of the classifier. This kind of 'genre' classifier is intended to be helpful to users of document retrieval systems. Therefore, it seems sensible to collect data from such users regarding their needs, judgements of relevance of documents and the factors upon which the judgements are based. Many studies in the field of genre classification of a similar purpose to Finn's (2002) project also acquire information for their task from users.[20] Moreover, these studies conduct user surveys, in order to generate genre taxonomies.[21]

Finn (2002) makes use of each feature set separately, in order to compare their genre prediction accuracies across and within domains. He trains the classifier on these features using decision-tree learning.[22] This technique is a type of inductive machine learning in which the learned function is represented by a decision tree. Decision trees classify instances in a tree structure, sorting instances down the tree from root to leaf nodes, with leaf nodes providing final classifications. Each node tests an attribute of an instance under classification, with each branch corresponding to a value of the possible values for a particular attribute. Such attribute testing iterates, so that classification continues down the tree (Mitchell, 1997: 52–53).

For the first genre class, *opinion/fact*, the following accuracies (defined as the average percentage of correct class predictions on 10-fold cross-validation) in single domain experiments are reported: the POS feature set results in an average accuracy of 84.7 percent, the BOW feature set achieves an average accuracy of 87.2 percent, and the HC set results in an average accuracy of 88.3 percent (Finn, 2002:74). The accuracies in the domain transfer experiments are reported as follows: the BOW feature set results in an average accuracy of 67 percent, the HC feature set achieves an average accuracy of 71.8 percent and the POS feature set results in an average accuracy of 81.5 percent (Finn, 2002:74).

[20] See for example, Santini (2006a), Meyer zu Eissen and Stein (2004); also Roussinov *et al.* (2000).
[21] See Rosso (2005) for more on the role of users in generating genre taxonomies.
[22] Using C4.5 (Quinlan, 1993), a learning algorithm, which implements the decision-tree learning technique.

Finn's (2002) examination of the accuracies and analysis of the performance of these particular features are of relevance to this research project. He suggests that the POS feature set is a more generalisable indicator of genre class then the BOW feature set, which is domain specific (Finn, 2002:75). Furthermore, he deduces that the POS feature set and the shallow text statistics of the HC feature set perform the best overall for genre classification of *opinion/fact.*

For the second genre class, *positive/negative*, the following accuracies in single domain experiments are reported: the POS feature set results in an average accuracy of 61.3 percent, and the BOW feature set achieves an average accuracy of 82.7 percent (Finn, 2002:78). This result is not surprising, as *positive/negative* appears to be a content-based rather than genre-based class. The accuracies in the domain transfer experiments were reported as follows: the POS feature set achieved an average accuracy of 47.1 percent and the BOW feature set resulted in an average accuracy of 47.8 percent (Finn, 2002:79). Intuitively, it is therefore to be expected that content words will have better discriminatory strengths than POS features. In addition, this provides further evidence against using BOW only as a feature for building classifiers that generalise well; it is domain-based, thus it cannot be expected to generalise well across domains.

Finn (2002:76) concludes that the choice of discriminatory feature sets depends on the genre classification task at hand and that it is likely that a combination of feature sets will produce the most favourable results. This conclusion is significant for this research project and provides the motivation for the mixed feature set that was used (see Chapter 3 for the features used).

In their study, Lee and Myaeng (2002) have an unusual approach regarding features they consider to be genre-revealing. Instead of linguistic measures, they use *term frequency* (*tf*) and *inverse document frequency* (*idf*).[23] *Tf* was first suggested by Luhn (1957), in order to establish relative term importance in a document. The general premise of *tf* is that terms that occur frequently in a document are more representative of

[23] In their follow-up study, Lee and Myaeng (2004) make use of some linguistic features such as nouns, pronouns, exclamations, verb endings, person names and special symbols (punctuation marks, currency and mathematic symbols). However, the focus of their study is using genre classification to assist in topic classification; as a result, their reported results and methodology are not sufficiently detailed to review meaningfully here.

that document in terms of its meaning than terms that occur infrequently in that document (Jurafsky & Martin, 2000:651). Similarly, *idf* (first defined by Spärck Jones, 1972) is based on the hypothesis that terms that occur in specific documents rather than having equal distribution over all the documents in a collection are more likely to be discriminatory (Jurafsky & Martin, 2000:651 & 653).

They determine the *tf*s and the *idf*s for each document and their terms in their corpus, which is comprised of English and Korean documents (7828 and 7615 documents respectively) collected from the web. These documents represent the following seven genres: *reportage*, *editorial*, *research articles*, *critical reviews*, *personal homepage*, *Q&A* (questions and answers), and *product specification*. Using the *tf*s and the *idf*s in various ratios they deduce an equation they term the "similarity-based approach" (Lee & Myaeng, 2002:146–148), which calculates the genre class of a new document.

The results of training (for the English document set only), using the similarity-based approach on a training set of 50 percent and a testing set of 50 percent, is reported, using micro-average recall/precision. The highest result of their approach is 0.87 (Lee & Myaeng, 2002:148). The result most relevant to this project is the accuracy obtained on the *research articles* genre using 130 terms: micro-average recall/precision is reported as 0.99, and accuracy (deduced from the confusion matrix) is 97.7 percent (Lee & Myaeng, 2002:148). This accuracy is reported on a test set of 600 documents, which removes any concern regarding sample size. Certainly, it would also be necessary to gain more knowledge about the homogeneity in terms of topics, faculties in which the papers were written, and structure, in order to further interpret this high accuracy.

In their study, Kelih, Antić, Grzybek, and Stadlober (2005) analyse the word length of 190 Russian texts, composed of 95 letters and 95 poems. They seek to determine to what extent word length contributes towards distinguishing texts according to author and genre (Kelih *et al.*, 2005:498). They claim that to measure word length in characters, as genre classification studies have tended to do, is not a suitable unit of measurement for many languages (Kelih *et al.*, 2005:499). Rather, they suggest measuring word length in the number of syllables per word; thus redefining a word as an orthographical-phonological unit.

They obtain six variables, all representing word length frequency distributions (Kelih *et al.*, 2005:501). Using discriminant analysis and experimenting on various combinations of these six variables they can achieve up to 89.5 percent accuracy (Kelih *et al.*, 2005:504. As with Karlgren & Cutting (1994), this is on a training set only). The accuracy is obtained, using only two variables: the relative proportion of four-syllable words and the quotient of dispersion. The former of the two variables is strongly discriminatory, and used on its own, can achieve 76.3 percent accuracy (Kelih *et al.*, 2005:504). This compelling evidence for word length being a powerful discriminator of genres, if defined as syllables per word, provides motivation for inclusion as a feature in future projects (this will be discussed further in Chapter 5). The evidence for word length as a good discriminator also provided some motivation for the inclusion of readability scores as features, as word length (in characters and syllables) forms part of the readability scores detailed in Chapter 3.

Santini (2004a) examines ten genres of fifteen documents each from the BNC: *conversation*, *interview*, *public debate*, *planned speech*, *academic prose*, *advert*, *biography*, *instructional*, *popular lore*, and *reportage*. Her approach to genre classification is to investigate the discriminatory ability of syntactic analysis. In genre classification, syntax is currently only partially employed in POS. This approach, however, is a rather shallow one. In following Argamon *et al.* (1998), she suggests using POS trigrams as indicators of deeper syntactic features.

Santini (2004a) explores this in her study, using four sets of features; 835 POS[24] trigrams excluding punctuation, 1033 POS trigrams including punctuation, 65 POS trigrams derived from the first set and 74 POS trigrams derived from the second set. Using these four feature sets Santini (2004a) trains her genre classifier using a Naïve Bayes classifier.[25] Naïve Bayes is a type of Bayesian learning based on Bayes' theorem (Jurafsky & Martin, 2000:148–149). It provides a way to calculate the probability of a hypothesis (in this case of a document belonging to a particular genre), based on its prior probability and the actual probabilities observed in data (Mitchell, 1997:156).

[24] The corpus was tagged using the Constituent Likelihood Automatic Word-tagging System (CLAWS5 tagger; Garside & Smith, 1997).
[25] The classifier used is from Weka open source machine learning software, available at http://www.cs.waikato.ac.nz/ ml/weka/index.html.

On stratified random sampling[26] 10-fold cross-validation she reports the following average accuracies: 78.9 percent on feature set one, 78.6 percent on feature set two, 88.9 percent on feature set three, and 84.1 percent on feature set four (Santini, 2004a:4). These accuracies are given for the six written genres only, as they are the most relevant for this research project.

After testing accuracy on POS trigrams, she conducts the same experiment using POS bigrams and unigrams, in order to establish whether performance using POS trigrams does indeed result in significantly improved classification. Analogous to the POS trigram experiment, the test on bigrams uses 451 POS bigrams without punctuation, 568 POS bigrams with punctuation, 36 POS bigrams derived from set one and 41 POS bigrams derived from set two. The classification accuracy on the written genres using POS bigrams reports 76.1 percent accuracy on set one and 70.5 percent accuracy on set two. When some selection of bigrams is made, accuracy improves to 86.8 percent on both sets three and four (Santini, 2004a:5).

In the POS unigram test, only two feature sets are used due to lower feature frequencies. The first set is comprised of twenty unigrams excluding punctuation, and the second set is composed of twenty-four unigrams including punctuation. Accuracy on the former set is reported as 72.2 percent and on the latter set as 74.4 percent of the six written genres (Santini, 2004a:5). This overall good accuracy of POS influenced the use of POS in this project, which will be discussed in Chapter 3.

Wastholm *et al.* (2005) conduct a genre classification study using the Stockholm-Umeå Corpus (SUC, 1997) of 500 Swedish texts, which, as with the German LIMAS corpus used in Wolters and Kirsten (1999), follows a division akin to the Brown Corpus.[27] This corpus is tagged with the Preparatory Action for Linguistic Resources Organization for Language Engineering (PAROLE) tagset.[28] Similar to Santini (2004), they use a Naïve Bayes classifier[29] trained using lemmas, POS, POS plus subcategories (for example,

[26] See Chapter 4, for a brief explanation of stratified random sampling.
[27] For the complete categories, see Forsbom (2005:17).
[28] Available at http://spraakbanken.gu.se/lb/parole/.
[29] The Perl script for this classifier can be downloaded at CPAN: http://search.cpan.org/_kwilliams/Algorithm-NaiveBayes-0.03/.

proper noun) and complete PAROLE word classifications (for example, gender) (Wastholm *et al.*, 2005:2).

They run their experiments on nine upper-level genres (*press reportage*, *press editorials*, *press reviews*, *skills, trades and hobbies*, *popular lore*, *biographies and essays*, *miscellaneous*, *learned and scientific writing*, *imaginative prose*), using unigrams, bigrams and trigrams of each feature. Sentence markers were included for the bigrams and trigrams. Ten-fold cross-validation is used to refine their classifier, whilst fifteen percent of the texts are kept aside as testing data. On this test set the best reported results are overall error rates of 40 percent for the POS trigram set and 38.7 percent for the POS plus subcategories bigram set. The best performance for the lemmas set is a 53.3 percent error rate, which provides even more evidence for using linguistic data, rather than a BOW approach, for genre classification (Wastholm *et al.*, 2005:3). Similar to Finn (2002), they suggest combining all feature sets for better classification performance (Wastholm *et al.*, 2005:3).

In addition to such overall testing results, Wastholm *et al.* (2005) provide precision and recall measures for the nine genres. The most relevant of these for this project are 33.3 percent precision and 25 percent recall on *biographies/essays*, and 63.6 percent precision and 77.8 percent recall on *learned and scientific writing* (Wastholm *et al.*, 2005:3).

The results on *learned and scientific writing* are in keeping with other studies but the results on *biographies/essays* are markedly poor. There are some potential reasons for this. The majority of the studies in automatic genre classification seek to make classifications of English texts; it can thus be assumed that the types of features discussed in these studies may not be equally successful if applied to other languages. It can also, however, be seen from some of the studies reviewed in this section that these features have been equally successfully applied to Greek and German. Language can thus not be put forth as a likely explanation for poor performance. Other potential explanations for performance are genre groupings, the sample size, and training and testing methodology. The last two can almost immediately be dismissed as being the most likely explanations. Training and testing methodology appear unlikely to explain

poor performance as training, testing and validation are conducted scientifically, using the Naïve Bayes algorithm. This algorithm has been used with good results by Santini (2004a), and it is therefore not probable that the training algorithm used can account for the poor performance. Size, too, is unlikely to explain poor performance because the corpus used by Wastholm *et al.* (2005) is composed of 500 texts, which is an adequate sample size. The most likely explanation is genre groupings, because the corpus follows the divisions of the Brown corpus. That these divisions are not always indicative of 'genre' has been discussed previously. This is a concern that Wastholm *et al.* (2005:3) address, which indicates that they are at least aware of this problem. Karlgren and Cutting (1994) were also rather unsuccessful in their classification of biographies, with an error rate of 65 percent. This provides further evidence in favour of clear definition of genre classes for automatic genre classification, as without such clarity the target of genre classification and desired outcomes are blurred.

Boese (2005)[30] focuses on web-genre classification, which has previously been mentioned as not relevant to this research project. However, as she examines many electronic versions of traditional genres, and uses an extensive list of linguistic features, her study is considered appropriate for review here.

She examines ten genres (343 documents), collected from the web, in terms of three sets of features, which represent style, form and content respectively. Similar to Meyer zu Eissen and Stein's (2004) presentation-related features, form represents document layout and web-specific features, such as HTML tags, and is therefore not discussed here (Boese, 2005:29). The style feature set is (Boese, 2005:26):

(A) Readability statistics[31]

1. Kincaid;
2. ARI;
3. Coleman-Liau;
4. Flesch Index;
5. Fog Index;
6. Lix; and
7. SMOG-Grading.

[30] See also Boese and Howe (2005).
[31] ARI and the Flesch readability statistics will be discussed further in Chapter 3.

(B) Sentence information (counts)

1. characters;
2. words;
3. sentences;
4. paragraphs;
5. syllables;
6. questions (counts and percent);
7. passive sentences (counts and percent);
8. sentences with at most 13 words (counts and percent);
9. sentences with at least 28 words (counts and percent);
10. average length in words;
11. longest sentence length;
12. shortest sentence length;
13. average length of words (in characters); and
14. average length of paragraphs (in sentences).

(C) Word usage (counts and percent)

1. conjunctions;
2. pronouns;
3. prepositions;
4. nominalisations; and
5. verb types (counts only).

(D) Sentence beginnings (counts)

1. pronouns;
2. interrogative pronouns;
3. articles;
4. subordinating conjunctions;
5. conjunctions; and
6. prepositions.

Many of the content features relate to web-specific genres and formats. Content features relevant to this study are (Boese, 2005:31):

(A) the top fifty common words in the BNC (derived from Stamatatos *et al.*, 2000a);

(B) the most frequent words across corpus;

(C) the most frequent words from each genre;
(D) a stopword list of 430 words; and
(E) twenty-six punctuation types.

Using different combinations of these features, Boese (2005) trained a genre classifier using the LogitBoost algorithm, which makes use of additive logistic regression (see Kessler *et al.*, 1997). On stratified 10-fold cross-validation, the best overall accuracy reported is 92.1 percent, making use of a set including all features except term frequency (Boese, 2005:50). On this same set, the classifier performs very well, achieving 100 percent accuracy for the genre most significant to this project, *technical paper* (33 documents), which refers, among others, to scholarly articles (Boese, 2005:52).

Such high accuracies are certainly impressive. However, it must be noted that form features, such as HTML tags, play a large discriminatory role for web classification (Boese, 2005:50). For this research project, which is concerned with a traditional genre, it is rather more useful to consider the accuracies achieved on relevant style and content features. The overall accuracy obtained on the twenty-six punctuation types is 44.6 percent, on the style features is 55.4 percent, and on the top fifty common words in the BNC is 62.4 percent (Boese, 2005:50). These results indicate that punctuation can be a useful discriminator if used in combination with other features (see also Stamatatos *et al.*, 2000a), and further, confirm the findings of Stamatatos *et al.* (2000a) regarding the effectiveness of the most common words of a language.

This section has reviewed the most relevant genre classification studies to this project, in order to establish common practice regarding corpus, linguistic features and training methodology. The review has covered studies on Greek, German, Korean, Russian and Swedish, but it is clear that automatic genre classification studies focus on English in the main. This is favourable for this project, which is also concerned with English, in that comparisons can be more equally made between such studies. Therefore, their features, methodology, and findings can be extended to this research project.

Previously, in this section, it was noted that there is a lack of standards in automatic genre classification regarding, among others, a benchmark corpus for training and testing. This much is clear from the studies reviewed; however, some trends regarding corpora are evident. An example of such a corpus-related trend is that similar kinds of 'genre' divisions occur in various projects. From this, it would seem that one of the main problems of automatic genre classification is that target genre classes are not clearly defined. This is partly indicative of *genre* itself not being clearly defined from the offset. Clarity in terms of both *genre* and target genre classes is required to define the genre classification task, to ensure that the results of the task can be sensibly interpreted. This *a priori* definition also ensures that the classification task is indeed one of genre classification and not of, for example, content-based classification or register classification.

The main trend is to make use of features that are easily extracted or easily computed from the corpus. This results in many features, which are mainly POS, function words (often derived from stopword-lists), keywords, punctuation, readability statistics, and lexical information, such as: average sentence length (in words), average word length (in characters), type/token ratio, total sentence count, total character count, and long words. Behind these features is a preoccupation with proving POS and other linguistic features more discriminatory for genre classification than a BOW approach. There is some evidence in favour of deeper-level features (see Stamatatos, Fakotakis & Kokkinakis, 2000b; also Santini, 2004a) but as these are manually expensive, and because surface features perform well, they are still not widely used. Similar to the need for the *a priori* definition of *genre* and target genre classes, features need to be examined before classification in terms of their potential discriminative ability, in order to identify and remove irrelevant features, which can be unhelpful in training. This too does not appear to be widely practiced (see Wolters & Kirsten, 1999; and Stamatatos *et al.*, 2000b), possibly because it is time-consuming.

The methodologies used in training classifiers on the basis of selected features differ widely: factor analysis (Biber, 1988), discriminant analysis (Karlgren & Cutting, 1994; Stamatatos *et al.*, 2000a; Stamatatos *et al.*, 2000b; Kelih *et al.*, 2005), *k*-nearest neighbour (Wolters & Kirsten, 1999), multiple regression (Stamatatos *et al.*, 2000b),

logistic regression (Boese, 2005), decision-tree learning (Finn, 2002), and Naïve Bayes (Santini, 2004a; Wastholm *et al.*, 2005). From this list it can be seen that the most popular techniques used in these studies are discriminant analysis and Naïve Bayes.

SVMs are excluded from this list, as studies that make use of SVMs are reviewed in the next section. These studies are reviewed separately because they make use of the machine learning approach adopted in this research project and are therefore to be considered on their own.

2.3.3 Overview of genre classification studies that make use of support vector machines

This section discusses five genre classification studies, relevant to this research project in terms of both linguistic features and, more important, methodology. In a similar manner to Section 2.3.1 and Section 2.3.2, studies are reviewed individually and in chronological order.

Dewdney *et al.* (2001) conduct a genre classification study of seven web-specific genres, using the Carnegie Mellon University (CMU) genre corpus (9705 documents). These genres are not relevant to this research project, and are thus not discussed here, but it is important to note that Dewdney *et al.* (2001) define their intended meaning of *genre*. They define *genre* as "a set of conventions in the way in which information is presented" (Dewdney *et al.*, 2001:1). Their notion of *genre* is too vague to agree entirely with what is meant with *genre* for this project. Despite their genre types not being relevant to this project, the features and machine learning techniques used are relevant.

Dewdney *et al.* (2001) use two feature sets: the first is a set of eighty-nine linguistic and layout features, and the second is a set of 323 words. Unfortunately, they do not detail all their features, but provide a broad outline (Dewdney *et al.* 2001:4). From this outline, it can be seen that they make use of the following linguistic features, among others (Dewdney *et al.* 2001:4):

(A) POS;

(B) closed-word sets, such as days of the week, months of the year, and signs of the zodiac; punctuation;
(C) the mean and variance of sentence length and word length;
(D) combined sentence length, word length and syllable estimates to provide measures of sentence complexity;
(E) Flesch readability metric; and
(F) mean word length divided by the mean sentence length.

As previously mentioned, this project is concerned with linguistic features, thus the layout features (for example, whitespace, line spacing, and tabulation) used by Dewdney *et al.* (2001) are not relevant to this project.

They train three classifiers, using three different kinds of machine learning algorithms: a Naïve Bayes classifier, decision-tree classifier[32], and a SVM (Radial Basis Function kernel) classifier.[33] SVMs as a machine learning technique will be described and discussed in Chapter 3. Each classifier is trained using each feature set in isolation and then a combination of both feature sets.

The classifier trained on SVMs performs the best on the two sets and the combination set (Dewdney *et al.*, 2001:1). Since none of the genre classes is directly relevant to this research project, the results are reported in terms of overall accuracy. The following results on 10-fold cross-validation are reported, as they are most significant for this research project. On the set of linguistic and layout features, an accuracy of 83.4 percent, a recall of 83.6 percent, a precision of 90.1 percent and a F1 Metric of 86.7 percent are achieved. On the word frequency features an accuracy of 81.7 percent, a recall of 81.8 percent, a precision of 88.4 percent and a F1 Metric 85 percent are achieved. On the combination set, the SVM classifier performs with an accuracy of 83.6 percent, a recall of 84 percent, a precision of 94.9 percent and a F1 Metric of 89.1 percent (Dewdney *et al.*, 2001:7).

From these results it can be seen that the classifier performs the best if the words and the linguistic features are used in combination. As mentioned in Section 2.3.2, Finn

[32] Quinlan's (1993) C4.5 decision-tree learner was used.
[33] The implementation used is SVM^{light} (Joachims, 1999a).

(2002:76) concludes that a combination of feature sets is likely to produce the most favourable results. This has been demonstrated by Dewdney *et al.* (2001) and provided further motivation for combining the features used in this project, which are representative of linguistic features and the BOW approach.

Argamon and Dodick (2004a)[34] undertake a genre classification study, more akin to this research project, from a more clearly defined linguistic approach than any other genre classification study. They approach genre classification from an SFL (systemic functional linguistics) perspective, examining systemic preferences across what they term "scientific genres" (Argamon & Dodick, 2004a:2). In particular, they wish to examine writer's stance towards assertions in the text and cohesive strategies in two journal genres: two paleontological journals, *Palaios* and *Quaternary Research* (222 documents); and two physical chemistry journals, *Journal of Physical Chemistry A* and *Journal of Physical Chemistry B* (238 documents).

At first glance, these genres may perhaps seem contrived and relating to content rather than genre (see the examples of genres presented in Section 2.2; also Couture, 1986:87). Argamon and Dodick (2004a), however, demonstrate the contrary. These two journal genres are indeed distinguishable, based on 101 features. The features are frequency counts of keywords and phrases of a functional nature, which were assembled independently of the corpus and therefore, less likely to overfit the training data. Argamon and Dodick (2004a:3–4) examine these features in terms of systems and subsystems: conjunction (elaboration, extension, and enhancement), modality (type, value, orientation, and manifestation), and comment (admissive, assertive, presumptive, desiderative, tentative, validative, evaluative, and predictive).

Argamon and Dodick (2004a) use a SVM (linear kernel)[35] to train the classifier.[36] They report accuracies (on 20-fold cross-validation) of between 83 and 91 percent in distinguishing between the two journal genres. These good accuracies have two implications for this research project. Firstly, the study demonstrates the good performance of SVMs in genre classification and thus motivates the use of this machine

[34] See also Argamon and Dodick (2004b).
[35] See Chapter 3 for more on kernels.
[36] The classifier was trained using SMO learning algorithm (Platt, 1999) implemented in the Weka package (Witten & Frank, 1999).

learning technique. Secondly, the good discriminatory ability of the features used by Argamon and Dodick (2004a) provide additional encouragement for the inclusion of many of these words and phrases as features (see Chapter 3 for more on features used in this project and their motivation). Argamon and Dodick (2004a) provide proof that rhetorical differences can be extracted, using easily computed linguistic features. They also show that the features used for automatic genre classification need not be dichotomous, that is, BOW versus purely linguistic features, but that words and phrases indicating underlying structure can be used to great success.

Meyer zu Eissen and Stein (2004) conduct a genre classification study, using eight genres, seven of which are web-specific (800 documents). They do not provide a clear definition of what they mean by *genre* but give examples of genres, such as *letter* and *editorial*, that agree with the usage of the term in this research project. Yet, it appears that they do not use *genre* in the same way as is used for this research project. This is because their web-'genres' seem to rather be webpage types, with the exception of *help* and *articles*. Even, *articles*, although superficially appearing to be a genre, is not a single genre because it includes research articles, reviews, technical reports, and book chapters. Nevertheless, this 'genre' is the one that is the most applicable to this research project. *Articles* consists of 100 documents that were used for training and testing.

Meyer zu Eissen and Stein (2004) distinguish their 'genres' based on four feature types (mainly normalised counts) that they combine into two sets according to computational ease. They are (Meyer zu Eissen & Stein, 2004:264–265):

(A) Set one

1. Presentation-related features, which are HTML-specific and represent the appearance of a document.[37]
2. Closed-word sets
    a. average word frequency class;
    b. currency symbols;
    c. help symbols;
    d. shop symbols;
    e. date symbols;

[37] Presentation features are not detailed here, as they are not relevant to this project; see Meyer zu Eissen and Stein (2004:264) for more on these features.

f. first names;
g. surnames; and
h. words that do not appear in Webster's dictionary.

3. Text statistics
   a. punctuation: question marks, colons, semicolons, dots, commas and exclamation marks;
   b. letters; and
   c. digits.

(B) Set two

1. POS[38]
   a. nouns;
   b. verbs;
   c. relative pronouns;
   d. prepositions;
   e. adverbs;
   f. articles;
   g. pronouns;
   h. modals;
   i. adjectives; and
   j. alphanumeric words[39].

They use the two feature sets separately in each experiment, in order to establish whether the discriminatory gain of Set Two outweighs its computational (and time) expense (Meyer zu Eissen & Stein, 2004:264). They initially explored the classification performance of these features using discriminant analysis (Meyer zu Eissen & Stein, 2004:266).

Their genre classifier was trained, using neural networks and SVMs. On average, for one-against-all classification,[40] they report an accuracy of 70.7 percent for Set One and an accuracy of 72.7 percent for Set Two. These scores are for the classifier trained,

[38] The corpus was tagged using the part-of-speech tagger of the University of Stuttgart available at http://www.ims.uni-stuttgart.de.
[39] This is also not clearly related to POS but is grouped as such in the original.
[40] This is a technique used for multiclass SVMs. Refer to Schölkopf and Smola (2002:211–212) for more on this.

using SVMs only, as they do not report accuracies for the neural networks learner, noting only that SVMs have the more accurate performance (Meyer zu Eissen & Stein, 2004:266). This conclusion and the superior performance of Set Two motivated the use of SVMs and POS features in this research project.

Aires, Aluísio, and Santos (2005) explore their genre classification, using the Brazilian Portuguese Lácio-Ref corpus (4278 documents).[41] This corpus is divided into five 'genres' and thirty 'text-types'. Their use of *text-type* corresponds to the meaning of *genre* in this project, while their use of *genre* does not; for example, they consider *poem* and *dissertation* as instances of text-types, and *scientific* and *instructional* as instances of genres.

They use forty-six features, derived from Biber (1988) and Karlgren (2000), detailed below (Aires *et al.*, 2005:2–3).

(A) Word-based statistics
1. type/token ratio;
2. capital type/token ratio;
3. average word length in characters; and
4. long words (more than six characters) count.

(B) Text-based statistics
1. character count;
2. average sentence length in characters;
3. average sentence length in words;
4. sentence count; and
5. text length in words.

(C) Other statistics
1. subjective markers;
2. Portuguese specific words, such as *que* and *se*;
3. discourse markers;
4. *wh*-questions;
5. amplifiers;
6. downtoners;

[41] For more on this corpus see Aluísio, Pinheiro, Finger, Nunes and Tagnin (2003).

7. emphatics;
8. verbs: suasive, private and public;[42]
9. articles: definite and indefinite;
10. pronouns: first person, second person, third person, demonstrative, indefinite, and pronominal expressions;
11. prepositions;
12. adverbials: place and time;
13. adverbs;
14. interjections;
15. contractions;
16. conjuncts; and
17. conjunctions: causative, final, proportional, temporal, concessive, conditional, conformative, comparative, and consecutive.

Using these features, Aires *et al.* (2005) train three classifiers using three different learning techniques: SVMs[43], decision trees[44] and logistic regression[45]. The most relevant results for this research project are those reported on the SVM classifier on the 'text-types', because this corresponds more closely to *genre* in this project. On 10-fold cross-validation of these 'text-types', Aires *et al.* (2005:3) report a 55 percent precision, a 91 percent recall and a 69 percent F-measure, using the SVM classifier. In Chapter 4, it will be seen that the most relevant evaluation measure for this research project is recall. Comparison to the decision-tree classifier, which achieves a 67 percent recall, and the logistic regression classifier, which achieves a 74 percent recall, shows that the SVM classifier outperforms these two classifiers.

Aires *et al.* (2005) clearly demonstrate the usefulness of linguistic features for genre classification, not only for English, as many studies have done, but also for Portuguese. In addition to this, the good performance of SVMs as a technique for training a classifier is confirmed.

[42] These verbs will be explained in Chapter 3.
[43] The SVM used is the SMO algorithm, and like Argamon and Dodick (2004a) it is implemented in Weka.
[44] They use J48, a Weka implementation of C4.5.
[45] The LMT algorithm is used (Witten & Frank, 1999).

In a follow-up study to Santini (2004a), reviewed in Section 2.3.2, Santini (2005b) further investigates the possibility of deeper syntactic structures as good discriminatory features for genre classification. In this study, she clearly defines what she means by *genre* and *text-type*; moreover, she uses both these terms in agreement with what is meant in this research project (Santini, 2005b:3). She uses two types of features, which she terms "linguistic facets" (Santini, 2005b:3). They are functional and syntactic cues, which, as with those of the majority of automatic genre classification studies, can all be automatically extracted.

The functional cues are extracted using the Connexor parser (Tapanainen & Järvinen, 1997). These functional cues are (Santini, 2005b:8–18):[46]

(A) predicators;
(B) complex NPs;
(C) nominals;
(D) pronouns: first person, second person, third person, and third person singular inanimate;
(E) present tense group;
(F) past tense group;
(G) imperatives;
(H) active;
(I) passive;
(J) negative particles;
(K) existential *there*;
(L) expressiveness;
(M) markers: time, location, instrument, probability, necessity, and manner;
(N) verbs: activity, communication, mental, causative, occurrence, existence, and aspectual; and
(O) connectives: enumerative, equative, reinforcing, summative, appositive, resultative, inferential, reformulatory, replacive, antithetic, concessive, discoursal, and temporal.

[46] Her work is very comprehensive, demonstrating each cue type with an example.

The twenty-five types of syntactic cues represent syntactic patterns created by finding common patterns from the parser output, and then writing an algorithm to detect these patterns, using regular expressions (Santini, 2005b:19–22).[47] They are (Santini, 2005b:18–38):

(A) Adverbial clauses: concession clause, conditional clause, contrast clause, exception clause, purpose clause, reason clause, result clause, similarity manner comparison clause, space clause, and time clause.

(B) Complement/nominal clauses

1. verb+*that* clause;
2. adjective+*that* clause;
3. *that* omission;
4. *wh*-clause;
5. verb+*to* clause;
6. adjective+*to* clause;
7. verb+*ing* clause;
8. comparative clause; and
9. relative clause.

(C) Simple sentences: phenomenon registering, action recording, phenomenon identifying, phenomenon linking, quality attributing, and action demanding.

Using eighty-four linguistic facets of the types listed above, as well as 211 POS trigrams, Santini (2005b) trains two SVM[48] classifiers to distinguish between four web-specific genres of 200 webpages each.[49] The average accuracy she reports over ten test sets, and using ten different seeds is 84.28 percent on the linguistic facet set and 86.50 percent on the POS trigrams set (Santini, 2005b:38–39).

Santini (2005b), similar to Argamon and Dodick (2004a) is linguistically very thorough and shows that surface words can be indicative of deeper structure. This study (Santini, 2005b) provided two strong influences for this project. Firstly, the high accuracy on set one provides good evidence for linguistic features as genre indicators; especially considering that the set used was a small feature set (Santini, 2005b:39). Secondly, the

[47] See Chapter 3 for more on the use of regular expressions in this research project. See also Friedl (1997) for more on regular expressions.
[48] Again, the training is done using the Weka package (Witten & Frank, 2000).
[49] No more detail is provided regarding the corpus.

performance of Santini's (2005b) classifiers gives definitive proof of the good performance of SVMs as a machine learning technique.

This section has discussed five automatic genre classification studies that use SVMs to develop genre classifiers. These studies show that when compared with classifiers trained on the same data but using different learning techniques, SVMs outperform other learning techniques. Furthermore, two findings are relevant for this project. First, Dewdney *et al.* (2001) demonstrate that words, that is, the BOW approach, and linguistic features can produce better results when used in combination than when used on their own. Second, Argamon and Dodick (2004a), and Santini (2005b) show that the features used for automatic genre classification need not be dichotomous, that is, BOW versus purely linguistic features, but that words and phrases can be indicative of deeper linguistic structure and as such can be used with much success.

2.4 Summary

This chapter commenced with an illustrated explanation of the notions of machine learning and supervised learning, and introduced the basic notation used in this project for learning by examples. These concepts were placed in the framework of automatic genre classification and the task of genre classification for this project explained.

Thereafter, the use of *genre* in this project was defined and the genre examined in this project identified in terms of prototypicality. Next, a review of automatic genre classification was presented, in order to detail the background to the features and methods used in this project. Each study was presented separately, for two reasons: firstly, because of variation between data, features, application, and rationale; and secondly, because of the lack of standards in automatic genre classification, regarding a benchmark corpus, evaluation metrics, baseline performance and methodology. However, where possible, comparisons were made between the projects reviewed. In addition, each statistical technique used in the reviewed studies was explained, and the value of the studies emphasised with particular focus on the value of the linguistic features used for this project. Moreover, these studies were critiqued in terms of validity of pre-defined genre classes, results, evaluation measures, and features.

In this review, previous relevant work in automatic genre classification and seminal works in the field were first reviewed. The most important study reviewed was Biber's (1988) seminal text variation work that forms the foundation from which many automatic genre classification studies have sprung. Thereafter, more current automatic genre classification studies were reviewed, with detailed reference to projects that are relevant and have a similar purpose to this research project, regarding application, corpus, features, and method. Finally, those studies that use SVMs for machine learning, for the purposes of genre classification, were detailed.

In spite of a lack of standards in automatic genre classification regarding, among others, a benchmark corpus for training and testing, some trends regarding corpora are evident. For example, the same kinds of 'genre' divisions occur in various projects, mainly because the same corpora are used (such as the Brown Corpus). These divisions are often not based on genre distinctions but rather on topic or discipline distinctions. It can be seen that poorly defined genre classes stem from an unclear definition of *genre* at the offset. At times, this results in low classification accuracies, and this also makes it difficult to deduce the reasons for low accuracy (for example, Karlgren & Cutting, 1994). It is thus important to clearly define *genre* before classification, and furthermore, to have clear target genre classes. This will mean that the object of classification is elucidated, and that results can be interpreted in light of the genre class. Such *a priori* clarification of the genre task ensures that the classification task is indeed one of genre classification and not of, for example, content-based classification.

The studies reviewed in this chapter mainly use features that are easily computed and extracted from the corpus. These features are mainly POS, function words (often derived from stopword lists), keywords, punctuation, readability statistics, and text statistics; such as, average sentence length (in words), average word length (in characters), type/token ratio, total sentence count, total character count, and long words. A concern with proving POS and other linguistic features better features for genre classification than a BOW approach underlies these features. There is also some evidence in favour of deeper-level features (see Stamatatos, Fakotakis & Kokkinakis, 2000b) but these are still not widely used, because they are manually expensive. Wolters and Kirsten (1999) and Stamatatos *et al.* (2000b) show that features need to be

examined before classification to determine their discriminative ability. Such examination needs to be done, in order to remove irrelevant features, which can be unhelpful in training. Pre-classification exploration does not appear to be widely practiced, possibly because it is time-consuming.

A variety of learning methodologies are used in training classifiers, the most popular techniques of which are discriminant analysis (Karlgren & Cutting, 1994; Stamatatos *et al.*, 2000a; Stamatatos *et al.*, 2000b; Kelih *et al.*, 2005), SVMs (Dewdney *et al.*, 2001; Argamon & Dodick, 2004a; Meyer zu Eissen & Stein, 2004; Aires *et al.*, 2005; Santini, 2005b) and Naïve Bayes (Santini, 2004a; Wastholm *et al.*, 2005).

The following concerns, relevant to this project were raised in the chapter: cross-domain accuracy, training sample size, evaluation of the discriminant abilities of features, normal distribution, parametric tests, and non-parametric tests. These will all be addressed in terms of this project in Chapters 3 and 4. Furthermore, based on the results obtained on SVMs and on various linguistic features in the studies that were reviewed, particular linguistic features were included in this project. The features that were particularly emphasised in this chapter were POS, the top fifty words of the BNC, punctuation, and various function words.

These features will be discussed further in Chapter 3, which will detail the steps followed in developing the SVM classifier, regarding the training and testing corpus, feature selection, feature extraction and the learning methodology used to train the genre classifier.

CHAPTER 3
Developing the classifier

Features do not randomly co-occur in texts. If certain features consistently co-occur, then it is reasonable to look for an underlying functional influence that encourages their use
(Biber, 1988:13)

3.1 Introduction

The aim of this project is develop a classifier that distinguishes between 'good' and 'bad' examples of argumentative essays. It must learn to do so, using the assessment of each essay, the dependent variable, according to human evaluators (the original markers of the essays), and linguistic features, the independent variables, (henceforth, only features) of the text.

In general, certain steps need to be followed, in order to develop such a classifier. **First**, a learning method must be selected by which to train the classifier (the *LM*), to distinguish between 'good' and 'bad' examples of the argumentative essay genre. **Second**, a corpus, which will provide the essays and thus the data for the classifier, is required. **Third**, the salient and relevant features that will allow the classifier to categorise 'good' and 'bad' examples of argumentative essays must be selected. **Fourth**, the corpus must be prepared in various ways, for example, the essays must be converted to text format. **Fifth**, the features that are to be extracted must be annotated. **Sixth**, the features must be extracted from the essays, in order to gather frequency counts on all the features in each text. **Seventh**, the data must be prepared before classification in different ways, for example, the data set must be transformed. **Lastly**, the classifier needs to be trained and tested, using the features and learning method selected.

Following steps one to eight, this chapter details all the information relating to the corpus, feature selection and extraction, and learning methodology. The chapter commences with Section 3.2, which discusses the corpus and provides the background to the selection of the essays used. In Section 3.3, the features used for training the classifier are described. In Chapter 2, many of these features were already described in the context of previous research. Hence, only some necessary details as to the reasoning

for the selection of each feature and examples of studies, which make use of each particular feature are given. Thereafter, Section 3.4 delineates the processes of text preparation before feature extraction. Next, Section 3.5 discusses the annotation of various features of the essays, in terms of POS tagging and XML mark-up. After providing a description of annotation, the preparation of the data before classification is described in Section 3.6, in terms of the standardisation of dependent variable, the removal of multiple occurrences of terms, data transformation, and feature selection. Lastly, the machine learning technique used to train the classifier is presented and explained in Section 3.7.

3.2 The corpus

The corpus is made up of 346 (813987 words) untimed essays from a section of the British Academic Written English (BAWE) corpus (Nesi, Sharpling & Ganobcsik-Williams, 2004),[50] which are freely available to researchers, but cannot be reproduced here.[51] All the essays are written in Standard British English by native speakers. Furthermore, these essays are all instances of argumentative essays written by students within an academic context. They are drawn from across the faculties and represent essays from many of the disciplines, except Engineering; and from most subjects, except Chemistry, as argumentative writing was not present among these essays/assignments. [52] For purposes of further research, a list, identifying which essays were used in the research project and their respective grading, is available in Appendix 1.

In this section, the necessary background to the corpus used for this research project has been delineated. The next section discusses the features used in training the genre classifier.

[50] The corpus was developed at the Universities of Oxford Brookes, Reading and Warwick under the directorship of Hilary Nesi (Warwick). Corpus development was assisted by funding from the ESRC (RES-000-23-0800).
[51] More details can be found at http://www.warwick.ac.uk/go/bawe/.
[52] For a complete list of departments/courses please see Appendix 1.

3.3 The features[53]

The features used to train the learner to distinguish between 'good' and 'bad' examples of the argumentative essay genre were derived from previous automatic genre classification studies (particularly Biber, 1988), reviewed in Chapter 2. The feature selection was also informed by two well-known English grammar books: Biber *et al.* (1999), and Quirk *et al.* (1985). Furthermore, many of the features were selected from the literature relating to academic discourse in general, chiefly Biber *et al.* (2004).

The features were extracted, using STATISTICA Text Mining and Document Retrieval (Statsoft, 2006). The text miner indexes all words found in the input documents, and computes a table of documents and words, a frequency matrix enumerating the number of times each extracted word occurs in each document. If all the words in a document are not to be indexed, it has an option for specifying the words and phrases to be extracted. It also provides an option for only those words that occur in a selected percentage of files to be extracted, and provides a stemming function[54] that is particularly helpful in extracting the different word forms of verbs. Once the raw word frequencies have been computed, STATISTICA Text Mining and Document Retrieval module (Statsoft, 2006) offers three transformations: logarithmic frequencies, binary frequencies and inverse document frequencies.

A list of all the extracted features, as well as some examples of each feature type, follows below.

3.3.1 Parts-of-speech

Words that are alike in grammatical function are categorised into groups such as verbs, nouns, adjectives, adverbs, and prepositions. These word classes are traditionally referred to as parts-of-speech (Halliday & Matthiessen, 2004:50–51). Parts-of-speech have been used in many automatic genre classification projects with good results; see for example, Finn (2002); Biber (1988); Karlgren and Cutting (1994); also Kessler *et al.* (1997). More information on POS tagging and the tagset used is given in Section 3.5.

[53] All complete feature lists can be found in Appendix 2.
[54] *Stemming* refers to the reduction of words to their roots, in order to count the different word forms of the words to be extracted, as instances of those words (Statsoft, 2004).

3.3.2 Punctuation marks

Terminal punctuation (full stops, question marks, and exclamation marks) and clausal punctuation (semi-colons and colons) were counted. The exclamation mark is a surface cue indicating emotive sentences and imperatives. Similarly, full stops indicate declarative sentences, and question marks indicate interrogative sentences.

Interrogatives may prove to be a relevant feature for academic writing because, according to Hyland (2002a:530), direct questions are used by writers of academic prose to engage readers and draw them into their argumentation and line of reasoning. Little attention has been paid to direct questions in academic writing, because they are perceived to be rarely used in this discourse (Hyland, 2002a:530). Hyland (2002a:530) however, finds that despite their relatively low frequency in comparison to their frequency in conversation, they do occur in academic writing.

In addition, semi-colons and colons may prove to be of relevance as they have been found to be indicative of complex sentences in previous research (Rauber & Müller-Kögler, 2001:5). As with parts-of-speech, punctuation has also been used in many automatic genre classification studies. Examples of these studies include Boese and Howe, (2005), Forsbom (2005), Lee and Myaeng (2004) and Stamatatos *et al.* (2000b).

In punctuation mark-up, commas were excluded because some manipulation of commas was required during sentence mark-up (see the mark-up scripts on the accompanying CD). Punctuation symbols such as $ and £ were not considered as they did not occur frequently in the essays being examined. More details regarding the mark-up of punctuation marks is given in Section 3.5.

3.3.3 Quotations

Many automatic genre classification studies that use punctuation as features make use of counts on single and double quotation marks as punctuation marks, such as, Lee and Myaeng (2004) and Bisant (2005). Others, for example, White, Cardie, and Ng (2002) gather counts on the presence of quotations. It appears that no automatic genre classification studies accord quotations special status.

In this study, however, more information on quotations was gathered: quotations were categorised and counted as two different types: integrated and non-integrated (following Nelson, 2002). Integrated quotations were regarded as those which were incorporated into the quoting author's sentence (refer to Section 3.3.5.5 for a definition of *sentence*). Those, generally longer, quotations that were not incorporated into the quoting author's sentence, and which, therefore, made up a sentence or sometimes a paragraph on their own were considered non-integrated quotations.

This study considered quotations and the verbs associated with them (see reporting verbs in Section 3.3.6.4) as potentially helpful features for distinguishing between prototypical and non-prototypical argumentative essays because, according to Hyland (2002b:116), reporting is an important convention of academic writing. Reporting shows the following principal quoting structures: direct quotation, paraphrase, summary, and generalisation (Hyland, 2002b:116). As this project is concerned with features that are easy to extract, direct quotations and references are extracted. It is assumed that references are indicative of the remaining three quoting structures given by Hyland (2002b:116): paraphrase, summary, and generalisation.

3.3.4 Nominalisations

According to Halliday & Matthiessen (2004:656), nominalisation is the "single most powerful resource for creating grammatical metaphor". Grammatical metaphor is "associated with the discourses of education and science" (Halliday & Matthiessen, 2004:636). It refers to one type of process (mental, behavioural, verbal, relational, and existential) being presented as another type of process. Grammatical metaphor can also refer to two or more types of processes being represented as a single process, which is not of the same type of either of the original processes (Halliday, 1987:76). Grammatical metaphor serves to background information, and is important for establishing context at the discourse level (Halliday, 1987:78). Its widespread use in writing thus deemed it a potentially relevant feature for this project. Only nominalisations representative of grammatical metaphor were considered here because they can be extracted automatically. More specifically, nominalisations are a kind of ideational metaphor where, for example, processes and qualities are represented as if

they were entities (Halliday & Matthiessen, 2004:637; see also Biber *et al*, 1999:322 for more examples), for example *rationalisation* and *demonstration*.

In addition to the role of nominalisations as grammatical metaphor, Chafe and Danielewicz (1987:98) claim that nominalisations are a linguistic device used to enlarge the size of intonation units. Along with nominalisations, prepositions and attributive adjectives, two linguistic devices that also effect intonational unit size, are "unusually frequent" in academic writing. These features are similar to Chafe's (1982) notion of integration (*integration/fragmentation dimension*), which is characterised by features that "pack information into a text" (Biber, 1988:21), such as nominalisations, prepositional phrases and attributive adjectives. In addition to this dimension, Biber (1988:21) proposes the *detachment/involvement dimension* (based on Chafe, 1982), in which detachment, intuitively expected to characterise writing in an academic context,[55] is also typically marked by nominalisations.

The above arguments for the important role of nominalisations, both at the sentence and discourse level, deemed nominalisations a potentially genre-revealing feature. Counts on nominalisations were extracted by matching nouns with suffixes *-tion*, *-ment*, *-ness*, and *-ity* (Biber *et al*, 1999:323). These particular nominalisations were selected for two reasons: firstly, because of their high occurrence in academic discourse in the Longman Spoken and Written English (LSWE) corpus (Biber *et al.*, 1999:322); and secondly, because they are suffixes that are, in the main, used to create abstract nouns. Extracting abstractness is relevant for this project as it is one defining characteristic of written discourse (Biber, 1988:47).[56] Written academic discourse, in particular, has been found to be highly abstract (Biber, 1986). Biber (1986:393, 395) finds that nominalisations, prepositions and passives[57] tended to co-occur and interprets these as representative of highly abstract texts (see Biber, 1986:395 for an intuitively attractive discussion of the causal relationship between these three constructions).

[55] Chafe (1982) describes academic writing as being characterised as detached and integrated.
[56] See Blankenship (1974) and Chafe and Danielewicz (1986).
[57] Passives are not used as features here, because they could not be easily automatically extracted (see for example Biber, 1988).

3.3.5 Text statistics

3.3.5.1 Word count

Word counts were extracted, making use of Oxford WordSmith Tools 4.0 (Scott, 2004). What constitutes a word was mainly determined by SVMTagger[58], as the number of POS tags and the token count of each document had to be kept the same. Therefore, numbers were considered and counted as words; amounts given with their currency in numerals, for example *£2*, were considered one word; and hyphenated words were considered one word. An exception to keeping POS tags and tokens equal in one document were apostrophes. SVMTagger tags apostrophes indicating possession separately, but not apostrophes indicating omission, for example, *the student's essay — student NN 's POS*. Therefore, apostrophes were allowed in words and at the end of words.

This feature was not used directly but rather as a calculation of sentence length and, as is shown in Section 3.6, for normalising feature counts to text length.

3.3.5.2 Word length

Word length is one of the most commonly used features in automatic genre classification (see for example Aires *et al.*, 2005; Bisant, 2005; Braslavski & Tselischev, 2005; Dewdney *et al.*, 2001; Finn, 2002; Karlgren & Cutting, 2004; Kessler *et al.*, 1997; and Santini, 2006b). The average word length in characters of each essay was extracted, making use of Oxford WordSmith Tools 4.0 (Scott, 2004).

3.3.5.3 Long words

Biber (1988) asserts that long words have "more specific, specialised meanings" than shorter words. Long words, which are defined in the literature as words longer than five, six or seven characters, have been found useful in automatic genre classification studies, for instance Aires *et al.* (2005), Finn (2002), Karlgren and Cutting (2004). This study considered those words longer than six characters. The raw frequencies of long words in

[58] This POS tagger is discussed in Section 3.5.1.

each essay were extracted and counted, using Oxford WordSmith Tools 4.0 (Scott, 2004).

3.3.5.4 Type/token ratios (TTR)

TTRs provide a measure of the number of different types of words in a text and thus potentially measure the "richness of vocabulary" (Chafe & Danielewicz, 1987:91) of texts. Since academic writing displays high lexical diversity, this measure can prove to be a particularly salient measure of prototypical and non-prototypical argumentative essays. They are widely used as a feature in the studies reviewed; Kessler *et al.* (1997) and Stamatatos *et al.* (2000b) provide some discussion of the TTR as a relevant feature.

The TTR is only helpful if texts of equal sizes are being compared, as the relation between the number of different lexical items and the total number of words in a particular text is not linear. This means that a large number of different lexical items that occur in the first 100 words of a text will be repeated in the next 100-word chunk, thus the different types will be reduced, as the length of the text increases (Biber, 1988:239). Therefore, if the TTR of a very short text is compared to that of a very long text it will appear that the TTR of the short text is much higher. A solution to this is to standardise the TTR to make for an informative and comparable measure. As this research project makes use of full texts and not abstracts, the texts are of varying lengths. Thus, the TTR of each text was standardised, using Oxford WordSmith Tools 4.0 (Scott, 2004), which calculates the number of types to tokens every n words and then computes a running average over each time it recalculates according to the size of n. It should be noted that this is just one method among others of calculating the standardised TTR (STTR) (for another method see Tuldava, 1995:131–150). TTRs were extracted, using Oxford WordSmith Tools 4.0 (Scott, 2004) and standardised for 800 words. The shortest texts were just over 800 words long, therefore $n = 800$ was selected. If the default, $n = 1000$ were used, these shorter texts would get $STTR = 0$, which would, of course, not be helpful. In this case, where $n = 800$ the TTR is calculated anew with every 800 words, and the average TTR over all the calculations is given as the STTR.

3.3.5.5 Sentence count

Defining sentences is a troublesome problem in Linguistics (see Roberts, 1960), but it did not suit the purposes of this project to try to incorporate many definitions of sentences. It rather seemed that what was most important was that the chosen definition be applied consistently throughout the texts. In addition, considering the computational nature of this project, a definition, which would allow easy automatic retrieval, seemed the most sensible. Sentences were thus defined as beginning with a capital letter and ending with a terminal punctuation mark: full stop (.), exclamation mark (*!*) and question mark (*?*). Gimén̈ez and Màrquez (2006:16), the creators of SVMTagger also take terminal punctuation to be "unambiguous sentence separators". This further confirmed the definition of *sentence* for this project (see also Halliday & Matthiessen, 2004:6; and Lebart, Salem & Berry, 1998:36).

Sentences were first marked up and then extracted from the texts (see Section 3.5 below for details regarding sentence mark-up).

3.3.5.6 Sentence length in words

As with many of the other lexical ratios, that Kessler *et al.* (1997:34) term "derivational cues", sentence length is commonly used as a feature in automatic genre classification (Kessler *et al.*, 1997, and Karlgren & Cutting, 1994). The average sentence length per document was calculated by simply dividing the number of tokens by the number of sentences in a text.

3.3.5.7 Paragraph count

As the next level in the graphological hierarchy of language (Halliday & Matthiessen, 2004:6–7), paragraphs seem an intuitive feature for inclusion, yet paragraph counts were not a widely used feature in automatic genre classification. It is interesting to note that besides paragraphs being graphological units in texts, they can be viewed as a kind of punctuation marker (Nystrand, 1987:209). This viewpoint extends back to the origin of a paragraph; it was a symbol noted in the margin "to indicate conceptual, narrative, and other shifts in the flow of discourse" (Nystrand, 1987:209). More information on

how paragraphs were distinguished and marked-up can be found in Section 3.4 and Section 3.5.

3.3.5.8 Paragraph length in sentences

In a similar manner to the calculation of sentence length, average paragraph length was calculated by dividing the number of sentences by the number of paragraphs in each text.

3.3.5.9 Readability scores

Readability tests are intended to provide an indication of the reading level of a text. Some readability scores, such as the Flesch–Kincaid Grade Level Score, give an indication of the reading level in terms of the grade level score, which is useful in secondary educational contexts. Other readability scores indicate the reading level according to a difficulty/ease scale. This type of readability score is suitable for the purpose of this project. Grade level scores are not helpful for essays at tertiary level, as the essays are all at a high reading level and would thus be equally scored.

Two widely used measures of readability for English were selected: the Automated Readability Index (ARI) and the Flesch Reading Ease score. The ARI (Smith & Kincaid, 1970) formula is given by:

$$((tc/tw \times 4.71) + (tw/tse \times 0.5)) - 21.43, \quad (3.1)$$

where *tc* is the total amount of characters, *tw* is the total amount of words, and *tse* is the total amount of sentences.

The Flesch Reading Ease (Flesch, 1948) formula is given by:

$$206.835 - 1.015\ (tw/tse) - 84.6\ (tsy/tw), \quad (3.2)$$

where *tw* is the total amount of words, *tse* is the total amount of sentences, and *tsy* is the total amount of syllables. Both these scores were calculated for each text (see Dewdney *et al.*, 2001; also Boese, 2005 for examples of automatic genre classification studies making use of readability statistics).

3.3.6 Word lists

3.3.6.1 Key function words

Keywords are typically used in text classification in general and in automatic genre classification. Words are considered as keywords not simply because they occur frequently in a text, but rather because they occur with unusual frequency in a text (or corpus) as compared to another text (or corpus). This means, that in order to establish keyness, a suitable comparison corpus must be identified and a word list extracted from it for comparison.

The corpus used in this project was measured against the written section[59] of the BNC, in order to establish keyness. The BNC was chosen as the comparison corpus because it represents the writing of native speakers of British English, which is the language of the project's corpus. Furthermore, the BNC represents texts from a variety of genres and across a broad spectrum of topics. It was thus deemed representative of the language being examined, and suitable for comparison.

The key function words were taken from the top thousand keywords in the corpus and they were then extracted, using Oxford WordSmith Tools 4.0 (Scott, 2004). Function words were defined as all words other than nouns and verbs. This project took the top key function words because it seemed more likely that function words would be better indicators of style and genre than, for example nouns, which are more expressive of content.

3.3.6.2 The top fifty most frequent words in the BNC

These words were selected based on Stamatatos *et al.* (2000a), who, in their study, found the most common words of the BNC to be good discriminators of genre. The list was recalculated from the unlemmatised word frequency list, selected from the written section of the BNC for this research project, using Oxford WordSmith Tools 4.0 (Scott, 2004). As Stamatatos *et al.* (2000a) considered *'s* and *n't*, which this project did not include, some minor differences in the final word list resulted.

[59] The entire spoken section of the BNC was excluded, and written scripts, which were to be read, for example, news scripts were removed.

3.3.6.3 Prepositions

As mentioned earlier, prepositions, as with nominalisations, are linguistic devices used to increase the size of intonation units. It was also mentioned previously that prepositions occur with unusual frequency in academic prose (Chafe & Danielewicz, 1987:98). Biber (1988:237) further reinforces the importance of prepositions in academic discourse, claiming that they are linguistic devices "packing high amounts of information" into academic writing.

It can thus be expected that a large number of prepositions would characterise prototypical argumentative essays. For their potentially significant role in distinguishing between prototypical and non-prototypical essays, prepositions were extracted as features of the essays. Three lists of prepositions were used: simple prepositions (Biber, 1988:236–237), two-word complex prepositions, and three-word complex prepositions (Quirk *et al.*, 1985:669–671). The particulars of extracting two-word and three-word prepositions are given in Section 3.5.

3.3.6.4 Reporting verbs

As this project considers quotations to be potentially helpful discriminators, reporting verbs that are associated with these quotations were also considered. According to Hyland (2002b:116), reporting verbs indicate a significant rhetorical choice as they allow writers "to convey both the kind of activity reported and whether the claims are to be taken as accepted or not". For example, *demonstrate* and *show* can be used to indicate the writer's agreement, whilst hedges, such as *suggest* and *imply* indicate a more hesitant rhetorical choice.

Following Biber (1988) and Aires *et al.* (2005), reporting verbs, which Biber (1988) terms *specialised verb classes*, are used in this project. They are derived from Quirk *et al.* (1985:1180), who define two main categories of superordinate verbs, the first of which they term *factual*, as it accompanies an indicative verb and introduces propositional information. They further divide factual verbs into "public" and "private" (Quirk *et al.*, 1985:1180). This division is based on intellectual states that are observable

(public) and not observable (private), examples of public verbs are *affirm* and *mention*, while examples of private verbs are *assume* and *establish*.

Quirk *et al.* (1985:1180) term the second category *suasive*. Verbs of this category "imply intentions to bring about" a future change of some kind (Quirk *et al.*, 1985:1180), examples are *intend* and *propose*. A list of factual and suasive verbs, derived from Quirk *et al.* (1985:1181–1182) associated with quotations (allowance was made for literary extracts and quotations as well) and paraphrasing of citations, was compiled. The category of perception verbs from Quirk *et al.* (1985:1033) is also included here as they are indicative of academic hedging (Biber, 1988:242), examples are *seem* and *appear*, see Section 3.3.6.6 below.

3.3.6.5 Conjunction

In English, cohesion is created by conjunction, reference, ellipsis, and lexical organisation (Halliday & Matthiessen, 2004:533). This means that conjunctions have an important function in providing cohesion to texts. In fact, cohesive relationships are one of the three areas that English (1999:22–24) regards as essential in her analytic framework of assessing student academic writing.

This research project considers three kinds of conjunction: conjunctive adjuncts, coordinating conjunctions, and subordinating conjunctions (delineated below).

3.3.6.5.1 Conjunctive adjuncts (conjuncts)

Conjunctive adjuncts, also known as linking adverbials (Biber *et al*, 1999:558–560) relate clauses to each other. Establishing relationships between clauses is a component of the system of cohesion. Moreover, Biber (1986) finds that they occur with high frequency in academic discourse. This renders them a potentially relevant indicator of prototypical argumentative essays. Henning (2006) provides further evidence for this feature as essential to the coherence and cohesion of student argumentative essays.

The word list used is derived from Quirk *et al.* (1985:631–640) and Biber (1988:239). Similar lists can be found in Halliday and Matthiessen (2004:82). Examples are *in comparison*, *notwithstanding*, *altogether*, and *similarly*.

3.3.6.5.2 Coordinating conjunctions

Coordinating conjunctions, also known as coordinators, join independent clauses, and are similar in function to linking adverbials in that they too link clauses. They differ, however, in that they can only occur at the clause boundary. For the sake of easy automatic extraction, no distinction was made between phrasal and clausal coordination. Furthermore, only simple coordinators were considered because they are the most frequent coordinators, according to Biber *et al.* (1999:79–83): *and*, *but*, and *or*. The coordinators *and* and *or* are particularly more frequent in academic prose than in the other registers they examine (Biber *et al.*, 1999:81).

3.3.6.5.3 Subordinating conjunctions

Subordinating conjunctions, also called subordinators, introduce dependent clauses. The list of subordinators was drawn from Biber *et al.* (1999:85) and Quirk *et al.* (1985:1078–1079). Examples are *although*, *since*, *when*, and *whilst*.

3.3.6.6 Hedges

Hedging, a term originally introduced by Lakoff (1973), is claimed by Hyland (1998:6) to be "an essential element of academic argument" (see also Flowerdew, 2002:98). Hedges are linguistic devices used by authors to express tentativeness in their propositions and argument. Such devices can be detected and extracted by surface phenomena (Hyland, 1998:3), which suits the purposes of this research project that aims to use only those features that are easily extracted.

Thus, although hedging is found in many linguistic forms, at both the lexical (for example, epistemic adverbs) and phrase levels, only features indicating hedging in the form of single words were used in this project. As mentioned in Section 3.3.6.4, hedges overlap with other word lists, such as reporting verbs (perception verbs and others),

likelihood stance adverbs, and downtoners. Therefore, hedges were extracted as part of the groups above, detailed in Sections 3.3.6.4, 3.3.6.7, and 3.3.6.8.

3.3.6.7 Downtoners

These adverbs are also known as diminishers as they are used to diminish the effect of a modified item. Their selection was based on Biber (1988:240) and Quirk *et al.* (1985:597–598). Examples are *partially*, *barely*, *virtually*, and *hardly*.

3.3.6.8 Stance adverbs

Stance adverbs overlap with Halliday and Matthiessen's (2004:125–131) comment adjuncts, which express the author's attitude towards the content of the text or message. The selection of these adverbs is derived from Biber *et al.* (1999:557–558, 853–874) and Biber *et al.* (2004:32).

1. Non-factual adverbs

Non-factual adverbs are used to express the author's feeling regarding the quality of certain information (Biber *et al.*, 2004:32). Examples are *confidentially*, *typically*, *strictly*, and *frankly*.

2. Factual adverbs

These adverbs are used to comment on the reality of a proposition (Biber *et al.*, 1999:557). Examples are *always*, *undoubtedly*, *indeed*, and *certainly*.

3. Likelihood adverbs

Such adverbs are used to show propositional limitations and demonstrate evidence for a proposition whilst not being specific as to the source (Biber *et al.*, 1999:557). Examples are *predictably*, *roughly*, *apparently*, and *perhaps*.

4. Attitudinal adverbs

These types of adverbs indicate the attitude of a writer towards a proposition (Biber *et al.*, 1999:558). Examples are *fortunately*, *surprisingly*, *astonishingly*, and *curiously*.

3.3.6.9 Stance adjectives

3.3.6.9.1 Stance adjectives controlling *that*-clauses

According to Biber *et al.* (1999:671–674), adjectives controlling *that*-clauses all control stance. The following features were selected based on Biber *et al.* (2004:34–35).

1. Attitudinal adjectives

Such adjectives demonstrate a writer's evaluation of a proposition or event. Examples are *amused*, *necessary*, *unusual*, and *incidental*.

2. Certainty/factual adjectives

These adjectives indicate certainty on part of the writer regarding particular propositions (Biber *et al*, 1999:671). Examples are *false*, *true*, *correct*, and *inevitable*.

3. Likelihood adjectives

Examples are *probable*, *possible*, *doubtful*, and *likely*.

3.3.6.9.2 Stance adjectives controlling *to*-clauses

The following list represents the major semantic domains in which adjectives that control *to*-clauses occur (Biber *et al.*, 1999:716–721). Again, these adjectives were selected according to Biber *et al.* (2004:35).

1. Certainty adjectives

Examples are *unlikely*, *guaranteed*, *apt*, and *liable*.

2. Ability/ willingness adjectives

Examples are *careful*, *fit*, *unable*, and *able*.

3. Personal affective adjectives

Examples are *perturbed*, *glad*, *surprised*, and *astonished*.

4. Ease/difficulty adjectives

Examples are *pleasant*, *easier*, *difficult*, and *tough*.

5. Evaluative adjectives

Examples are *desirable*, *better*, *wonderful*, and *useless*.

3.3.6.10 Nouns

3.3.6.10.1 Stance nouns taking *that*-clauses

According to Biber *et al.* (1999:648), *that*-clauses complements of nouns are "one of the primary devices used to mark stance in academic prose". The list of stance nouns below was derived from Biber *et al.* (2004:33) and Biber *et al.* (1999:648–651).

1. Factual nouns

Examples are *knowledge*, *realisation*, *assertion*, and *fact*.

2. Likelihood nouns

Examples are *impression*, *idea*, *suspicion*, and *contention*.

3. Non-factual nouns

Examples are *news*, *report*, *proposition*, and *comment*.

4. Attitudinal nouns

Examples are *hope*, *reason*, *view*, and *thought*.

3.3.6.10.2 Nouns taking *to*-clauses

According to Biber *et al.* (1999:652), these nouns occur frequently in academic discourse. Thus, in following Biber *et al.* (2004), the controlling nouns were used as features. Examples are *proposal*, *intention*, *potential*, and *authority*.

An overview of all the features used as attributes of texts in the corpus has been provided above. Many of these features, especially those that are part of word lists, were simply extracted, using STATISTICA Text Mining and Document Retrieval module (Statsoft, 2006) with no annotation required. However, other features required annotation; what these features are, how they have been annotated and the reasons for their annotation are discussed in Section 3.5. The next section provides all the details relating to preparing the texts for mark-up and subsequent feature extraction.

3.4 Text preparation before feature extraction

This section provides all the information relating to the preparation of all the texts before feature extraction. Since this section details some preliminaries prior to mark-up there is some reference, and occasional (but logically necessary) overlap with Section

3.5, which details mark-up of the text for features that cannot be automatically extracted. The first part of text preparation presents all the necessary information, relating to the cleaning and preparation of the data before feature extraction. It should be noted that unless stipulated, this preparation was done, using regular expressions. The scripts were written in Perl, and compiled and run on Linux. The full scripts can be found on the accompanying CD.

The texts were prepared by first converting them to text format. This means that all formatting, such as bolding, italics and varying text size, was removed. The texts were originally encoded in Unicode but this character encoding was problematic for processing by the SVMTagger,[60] as the English models used for training are in Latin-1, known more formally as ISO/IEC 8859-1 (Giménez, 2007). All the texts were thus converted to Latin-1, using Utrac-0.3.0, a command-line tool that converts character sets (Calando, 2004). [61] The converted texts were also used for the CLAWS4 tagger that takes plain text ASCII files as input.[62]

Next, the essay questions, essay titles, bibliography, appendices, headings, and footnotes were removed manually. As the essays are written by students from the broad spectrum of faculties, a mixture of graphs, illustrations, tables, and equations occur. These were not considered as part of the data and were removed from the essays, by first marking them up, using regular expressions and then removing them (along with their tags) automatically. In order to facilitate paragraph mark-up, all indents indicating paragraphs were removed, and paragraphs were separated by a blank line.

Punctuation marks that were used as features are detailed in Section 3.5. It is necessary to note here, however, that brackets (square and angle) were removed automatically because content in these brackets were mainly equations. Other common punctuation marks and symbols (%, =, <, >[63] and &) were converted to words to allow them to be tagged as words rather than as symbols (the tag SYM, used for symbols is not reliably tagged, see Appendix 2). The ellipsis character, if inserted as a special character in

[60] Section 3.5 discusses this tool, which was used for part-of-speech tagging.
[61] The tool is freely available at http://utrac.sourceforge.net/download/utrac-0.3.0.tar.gz.
[62] Section 3.5 also discusses this tool, which was used for part-of-speech tagging.
[63] This does not refer to angle brackets used as brackets but rather to the mathematical symbols *bigger than* and *less than*.

Microsoft Word, yields a strange symbol in text format. This practice was very common among the writers of the essays, thus this symbol was converted to the usual three ellipsis dots (...).

Non-standard usage of the apostrophe, according to Standard British English writing conventions (Truss, 2003:46) in, for example *1960's* if indicating a plural, not possession, was removed. This was done for the taggers, which otherwise would tag all instances of such occurrences as possessive.

In cases where long quotations (an instance of non-integrated quotations) were not marked with quotation marks, quotation marks were inserted manually. Such quotations were then also incorporated into the paragraph preceding them, if the quotations were introduced by a colon or any other clausal punctuation (including a comma). All 'smart' quotation marks (“,”, ‘ and ’) were replaced with straight quotation marks (' and "). In addition, all quotation marks were standardised as double straight quotation marks ("). Cases of play titles and book titles were standardised to single straight quotation marks ('). Apostrophes for both possession and omission were retained in 'smart' apostrophe form during XML mark-up, which, as previously mentioned, is detailed in Section 3.5. This was done to allow for easy distinction between single straight quotation marks and apostrophes.

For the set of texts used for submission to the taggers, all non-integrated quotations were removed because they were regarded as extra-corpus material, as they are long sections of text not written by the author of the essay and thus considered not characteristic of the essay itself (this is in following Nelson, 2002:10). This is also true of integrated quotations. However, these quotations form part of the sentence and could not be removed without detrimental effect to the POS tagging results. Unfortunately, removal of the non-integrated quotations could not be done without first marking-up the texts with non-integrated quotation tags, in order to flag and remove the quotation itself. The mark-up script did all the marking-up simultaneously; therefore, all other tags had to be removed again for the texts used for POS tagging. In addition, all 'smart' apostrophes were converted to straight apostrophes, as 'smart' apostrophes are not

included in the Latin-1 character set. The final step was then tokenising the texts, so that the input texts had one word (a token), punctuation mark, or number per line.

After text preparation and before feature extraction, annotation of various features was first required, in order to extract them. Throughout this section, some mention has been made of annotation. The next section provides all the details relating to the mark-up of features.

3.5 Annotation of the corpus

This section describes the process of annotating the features of the essays that could not be easily extracted without annotation, using Oxford WordSmith Tools 4.0 (Scott, 2004) or STATISTICA Text Mining and Document Retrieval module (Statsoft, 2006). These features are tagged with two types of tags: POS tags and XML tags. The section first discusses POS tagging and then XML tagging.

3.5.1 POS tagging

The texts were grammatically annotated for parts-of-speech, using a software program called a POS tagger, which assigns to each word in a text its POS in context (Garside & Smith, 1997:103). POS taggers can take three approaches: a stochastic approach, a rule-based approach, or a hybrid approach (a combination of both the stochastic and the rule-based approaches). Stochastic taggers select the preferred tag of a word based on probabilities, which in turn are based on the tags in the training corpus.[64] Such probabilities are calculated according to the frequency of the particular tag as well as the probability of the tag based on its immediate neighbours (physical context) in terms of both their words and associated tags (Garside & Smith, 1997: 102). Rule-based taggers make use of linguistic knowledge to tag words according to syntactic correctness (Jackson & Moulinier, 2002:13),[65] for example, such taggers make use of the phrase-structure rules of a particular language.

[64] See, for example, Brants (2000) for a stochastic tagger.
[65] See, for example, Brill (1992) for a rule-based tagger.

There were two requirements for selecting a POS tagger for this project:

1. it had to have a high degree of accuracy; and

2. it had to require no training, that is, have already been trained and evaluated on English POS tagging.

Stochastic taggers require less handcrafting, and therefore, less manual annotation. Thus, with the requirements for a POS tagger, for this project, in mind, the SVMTagger from SVMTool was selected.[66] According to Smith (1997:138), the benchmark error rate for POS taggers is 3 to 5 percent. SVMTool's reported accuracy is higher than any other tagger, according to the literature, with a reported accuracy of 97.16 percent for English on the WSJ corpus (Giménez & Màrquez, 2004). It makes use of the Penn Treebank tagset (Marcus, Santorini & Marcinkiewicz, 1993), which is made up of 36 tags (the complete list of POS tags is available in Appendix 2). This is the tagset, which English POS taggers, in the main, are trained on. Initially, the relatively small size of this tagset appeared to be favourable for this research project where the feature set was already very large.

For the purposes of this project, the English models based on the WSJ corpus were used.[67] These models are the ones on which SVMTagger is trained, which suited the purposes of this project that required a trained tagger. The corpus was tagged, using the Perl version of SVMTool v1.3. The specific usage and options used can be found in Appendix 3.

Later in this study, it was **hypothesised** that a larger tagset might highlight linguistic subtleties that would not be noticed with the Penn Treebank tagset. Furthermore, it was postulated that a tagset, which extracted finer linguistic detail may extract potentially useful discriminatory features, and thus a classifier built on this tagset would perform more accurately than one built on the Penn Treebank tagset (this hypothesis will be discussed further in Chapter 4). An option would have been to train SVMTool, using an expanded tagset. However, as has been mentioned, a trained tagger was desired for this project. For this reason, the texts were tagged, using the CLAWS4 (Constituent Likelihood Automatic Word-tagging System) tagger, which is also a stochastic tagger

[66] Freely downloadable from http://www.lsi.upc.edu/~nlp/SVMTool/.

[67] The models are also downloadable from http://www.lsi.upc.edu/~nlp/SVMTool/.

like SVMTool.[68] The tagger was developed at the universities of Lancaster, Oslo, and Bergen by Roger Garside (Garside, 1987:30) and is written in C. This version of the tagger uses the CLAWS7 tagset, which is a modification of the original CLAWS1 tagset. The CLAWS7 tagset consists of 140 tags. Similar to SVMTool, CLAWS also performed well in previous projects, and consistently achieved a 96 – 97 percent accuracy. The version of the tagger that was used for this research project was used to tag *circa* 100 million words of the BNC, for which an error rate of 1.5 percent, with *circa* 3.3 percent ambiguities unresolved, was achieved (UCREL, 2006). Classifiers were thus trained, using the Penn Treebank tags and the CLAWS7 tags separately, in order to test the hypothesis presented here.

3.5.2 XML tagging

In addition to POS tags, XML mark-up was used for sentences, paragraphs, quotations, references, punctuation marks, nominalisations (identified by suffix), two- and three-word complex prepositions, two-word adverbs, and multi-word conjuncts. The XML tags were inserted by matching characters, using regular expressions. The scripts were written in Perl, and compiled and run on Linux. Full scripts can be found on the accompanying CD, and the XML tags that were subsequently extracted as data for the classifier, can be viewed in Appendix 2. The resulting tags, such as the POS tags and unannotated features, were then extracted, using STATISTICA Text Mining and Document Retrieval module (Statsoft, 2006).

The scripts used to do the sentence mark-up were mostly error free. However, some manual proof checking was required to correct errors such as abbreviations with full stops, that were sometimes incorrectly marked-up with end-of-sentence tags, and capitalised words, that were sometimes incorrectly marked-up with beginning-of-sentence tags.

In addition to sentence mark-up, paragraphs, quotations, and references were marked up and their tags extracted. Quotations were marked-up, either as integrated or non-integrated quotations. This distinction and separate tagging of integrated and non-

[68] The CLAWS4 tagger is not free but it can be freely sampled, for more, see http://www.comp.lancs.ac.uk/ucrel/claws/.

integrated quotations follows the mark-up used for the International Corpus of English (Nelson, 2002). Sentences within non-integrated quotations were also marked-up separately. These were marked up to extract counts on the features, and to flag them for removal before POS tagging.

Sentence, quotation, and reference mark-up were necessary because sentence, reference, or quotation distinctions needing to be retrieved could not be automatically retrieved through feature extraction. This was not so for punctuation marks, nominalisations, two- and three-word complex prepositions, two-word adverbs, or multi-word conjuncts. They can be automatically extracted from the text by simple pattern matching. However, STATISTICA Text Mining and Document Retrieval module (Statsoft, 2006) cannot extract punctuation marks, nor is it capable of fuzzy text matching, which was necessary for extracting nominalisations. It was also somewhat unreliable in extracting phrases. Therefore, punctuation marks, nominalisations, three-word complex prepositions, two-word complex prepositions, two-word likelihood stance adverbs, two-word factual stance adverbs, and multi-word conjuncts were also marked-up with XML tags, which could be matched and then extracted using STATISTICA Text Mining and Document Retrieval module (Statsoft, 2006).

This section has described the annotation of essays used in training the classifier. The next section details the preparation of the data before submission to the classifier.

3.6 Data preparation before classification

The preceding section has discussed the mark-up of the texts, using POS tags and XML tags. This section discusses the preparation of the data before it was submitted to the SVM classifier.

Once all the features discussed in Section 3.3 were extracted from the texts, some data processing was necessary. Firstly, a dependent variable needed to be calculated. The dependent variable is the variable that depends on the predictor variables to determine its value. In the case of this project, the dependent variable is the variable, which indicates the classification of the essays: 'good' or 'bad' examples of the argumentative essay genre.

In order to calculate the values of the 'good'/'bad' variable for each essay, the original grades of all the essays are required. As all the essays from the BAWE were written for real assignments in various departments, this means that the original grade assigned to each was already available. No additional grading by hand was therefore necessary. However, three values for grading are found in the BAWE: typical percentage grading (for example, *60 percent*), class grading (*2:1* and *1:0*) and evaluative grading (either *satisfactory* or *excellent*). Percentage grading was selected as the more usual value, and so it was decided to standardise the other two types of grading to percentage grading. Of course, this choice is arbitrary. The only requirement is that only one value was selected to represent the dependent variable, *viz.* grade.

The standardisation of grading type was done as follows. Only two types of class grading are given in the BAWE: those representing first degrees (*1:0*) and upper second degrees (*2:1*). First degrees, which fall within 70 percent and above, were converted to 85 percent (Holmes, 2007b), while upper second degrees, which fall within 60 to 69 percent, were converted to 65 percent (Holmes, 2007b).

Similarly, evaluative grading in the BAWE is only of two types: *excellent* and *satisfactory*. The first evaluative grade, *excellent*, represents a first degree and the second evaluative grade, *satisfactory*, an upper second degree (Holmes, 2007a). Thus, following the conventions for first and second degrees stated above, the first evaluative grade was taken to represent 85 percent, and the second evaluative grade 65 percent.

The standardised grading available for every essay could now be used to compute the 'good'/'bad' variable. This was done by calculating the median of the grades, which was 68 percent. Therefore, any essay with a mark of 68 percent or above was considered 'good', and likewise any essay with a mark of 67 percent or under was considered 'bad'.

In addition to this first classification scheme by grade, for the purposes of a separate experiment, essays were grouped as 'good' examples if they had a mark of 70 percent or above and as 'bad' examples if they had a mark of 65 percent or less. This method removed the essays that had a middle-band grade of between 66 and 69 percent. The

idea was derived from Finn (2002:71), whose pre-classification of restaurant reviews as negative for marks above 15 or positive for marks above 23, precludes the middle-band marks, which are more difficult to classify. In this project, it was **hypothesised** that the perceptible difference (in terms of the linguistic features used here) of essays that are of 'average' grade, that is less prototypical, is not sufficiently discriminant to determine prototypical ('good') or non-prototypical ('bad'). This is illustrated in Figure 3.1 below. It therefore seems sensible to rather use only essays that are most likely to be definitely 'good' or 'bad', by using only the upper- and lower-band grades.

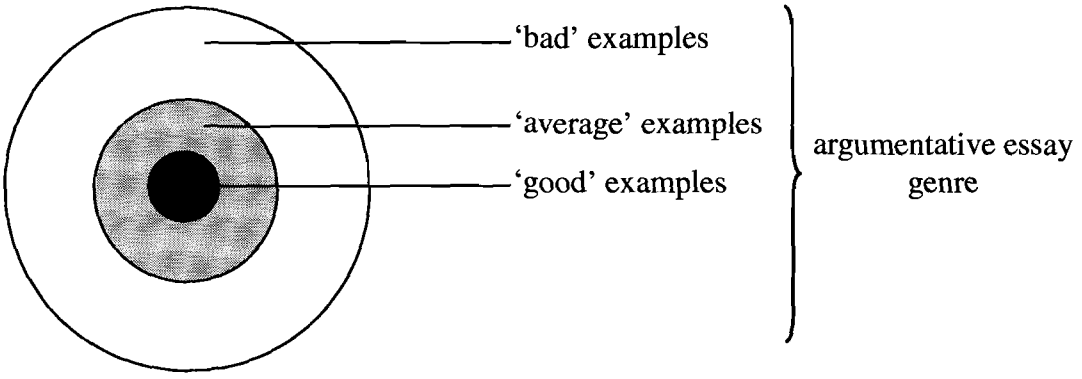

Figure 3.1: Illustrating the genre classification task, 'average' examples

The classifier was trained separately, on both these sets, the results of which will be reported in Chapter 4. It should be noted that, for the real application of this classifier, essays with middle-band grades will need to be classified. The removal of essays of 'average' grade is experimented with in this project, in order to determine whether the features used in this project are indicative of essays on either side of the continuum. This experiment does not imply that the classifier will be unable to classify essays falling in the middle-band grade.

Secondly, because of multiple (two or three) occurrences of terms in the word lists used for feature extraction, some data cleaning was necessary[69]. Multiple occurrences were therefore removed from the data. If this was not done these words would be given more weight in classification than other singly occurring words. This is because the frequencies of features lend them weight in classification, rather than multiple counts of their frequencies. The multiple occurrences arise from more general word lists such as

[69] The list of terms with multiple occurrences in the word lists can be found in Appendix 3.

the top fifty words in the BNC and the top function keywords. They do not arise because the words serve different functions. For example, *whilst*, occurs in the subordinating conjunctions list and in the top function keywords list.

Thirdly, before raw frequencies of terms could be given to the classifier, normalisation was required. *Normalisation* in this sense entails bringing each feature vector to unit-length according to the notions of length in a particular vector space. For this project, the data were normalised to text length, which corresponds to the L_1-norm.[70] Text-length normalisation provides a way for texts of different lengths to be compared and is common practice in corpus linguistics and genre classification studies (see for example, Dewdney *et al.*, 2001; also Santini, 2005c). It was **hypothesised** that for this project text-length normalised frequencies would yield better results than raw frequencies as frequencies normalised to text length are less skewed. Text-length normalisation is given by (Biber, Conrad & Reppen, 2000:265):

$$rwf/tw \times n, \tag{3.3}$$

where rwf is the raw frequency of a particular word/term in a document, tw is the total number of words in a document, and n is the basis for norming. This basis should be selected based on the average number of words in all the documents being compared. For the normalisation of the raw frequency counts for this project $n = 1000$. For example, if the raw frequency count for variable VBD is 53, the normalised frequency is given by: $53/1483 \times 1000 = 35,72$ per 1000 words. Once features were extracted they were each normalised in this way by using an equation in the spreadsheet they were extracted to. This process was not automatic, however, as this equation was re-entered for each feature (but not for each text).

Obviously, no normalisation was required for word length, sentence length, or paragraph length as they are not raw frequency counts and represent averages per document.

The raw frequency counts were also transformed, using logarithmic transformation and inverse document frequency transformation (*idf*, as explained in Section 2.3.2), which

[70] This norm is referred to as the block-city norm and corresponds to the linguistic notion of text length. See Leopold and Kinderman (2002) for more on this, and also on using the L_2-norm (Euclidean norm) instead of the L_1-norm.

are both typical transformations used in text classification (Forsbom, 2005:7; see Lee & Myaeng, 2002 for a genre classification study that uses *idf* with good results; also Leopold & Kinderman, 2002 for an SVM study in text classification that uses *idf*). Logarithmic transformation is a type of normalisation used to transform data that do not follow a normal distribution, which is an important distribution assumption, essential to many statistical tests and the interpretation of results.[71] *Normalisation* in this sense means that frequency distributions are transformed to follow a normal distribution. Logarithmic transformation is one of three options provided by Rietveld and Van Hout (2005:130–131), in order to overcome situations where the variables do not follow normal distribution. It is computed, using STATISTICA Text Mining and Document Retrieval (Statsoft, 2006) and is given by:[72]

$$f(wf) = 1 + \log(wf), \text{ for all } wf > 0, \tag{3.4}$$

where wf is the frequency of a particular word/term in a document. Using the previous example, the logarithmic transformation frequency for variable VBD is given by:

$f(wf) = 1 + \log(53)$.

Inverse document frequency is also computed, using STATISTICA Text Mining and Document Retrieval (Statsoft, 2006) and is given by:[73]

$$idf_{ij} = \begin{cases} 0 & \text{if } wf_{ij} = 0 \\ \left(1 + \log\left(wf_{ij}\right)\right) \log \dfrac{\ell}{df_i} & \text{if } wf_{ij} \geq 1 \end{cases}, \tag{3.5}$$

where wf is the frequency of a particular word/term in a document, ℓ is the number of documents, df_i is the number of documents in collection ℓ that contain the word/term w_i, and j refers to the document.

All the raw frequencies of variables were normalised and transformed in this manner.[74] This resulted in four sets of data:

1. a set of raw data frequencies;
2. a set of text-length normalised frequencies;

[71] See Oakes (1998:3–5) for more on normal distribution.
[72] This transformation uses log base e (Statsoft, 2004).
[73] Like the logarithmic transformation, the *idf* uses log base 10.
[74] All the raw and normalised frequencies for the actual extracted features can be found on the accompanying CD.

3. a set of log-transformed frequencies; and
4. a set of *idf* frequencies.

All four data sets were used to train four separate classifiers. The results of testing these classifiers will be discussed in Chapter 4, and the complete results in terms of both training and testing are available on the accompanying CD.

Finally, various feature selection methods were used to select those variables that are better predictors of the dependent variable.[75] Feature selection was only conducted on the text-length normalised data as this data set proved to perform the best (this will be discussed in Chapter 4). Feature selection appears to be not as relevant for SVMs as for many other machine learning and statistical classification tests (such as decision trees); nevertheless, some studies have reported that with many irrelevant features SVMs may perform poorly (see for example, Weston, Mukherjee, Chapelle, Ponntil, Poggio & Vapnik, 2001; also Liu & Zheng, 2006).

Feature selection was thus done for this project, in order to compare the performance of a classifier built on the full feature set, that is, with no feature selection, to a classifier built on a reduced feature set, that is, with feature selection. Variables were selected, using two kinds of tests:

1. parametric tests, which assume normal distribution; and
2. non-parametric tests, which do not make this assumption.

These two types of tests were required to compare predictive results, as some of the variables used, follow a normal distribution and others do not.[76]

The tests assuming normality were the *t*-test for independent samples and STATISTICA Feature Selection and Variable Screening module (Statsoft, 2006). This test requires specification of the number of cuts (k) into which the values of the predictor variables must be divided. The value selected for the number of cuts detects monotone to complex non-monotone relationships between the predictor variables and the dependent variable. The default value for k is 10, with smaller values detecting simpler relationships. For this research project, k was set equal to 2, 3, 4, 5, 6, 7, 8, 9, and 10.

[75] Note that the term *feature selection* has two meanings, see Section 3.7.3 for more on this.
[76] For each variable's descriptive statistics and for a sense of variable distributions see the accompanying CD.

The tests not making use of the normality assumption were the Mann-Whitney U test (see Oakes, 1998:17 for more on this test)[77] and the Kolmogorov-Smirnov Two-Sample Test. The non-parametric tests are the second of the three options given by Rietveld and Van Hout (2005:132) for cases in which the normality assumption is violated. Many of these tests are available; the two tests used here were selected as they can be used for comparisons between two independent groups (see Siegel & Castellan, 1988 for a discussion of non-parametric tests). The parametric tests and non-parametric tests were only used on normal and non-normal variables respectively. Furthermore, in order for any particular feature to be accepted, both tests for each test type were required to agree on the inclusion of this feature.

The best predictor variables, according to the feature selection tests discussed above and all the results using the feature selection techniques as described above, are documented in Appendix 5. The results of the classifier built on the set of the best predictor variables will be presented in Chapter 4.

Following McClave and Sincich (2000:229–232) normality was determined, using four descriptive methods.[78] The first two are visual assessment methods and involve constructing a histogram fitted with a normal curve, as well as a normal probability plot for each variable. The histogram shows that the data are normal if they follow the curve. Similarly, the normal probability plot shows that the data are normal if they follow the straight line closely. The third method for assessing normality was to determine the interquartile range (IQR) and the standard deviation (σ), in order to calculate the IQR/σ ratio. This ratio must be approximately equal to 1.3 to show that the data are approximately normal (McClave & Sincich, 2000: 230). The fourth method for determining normality was to calculate the intervals $\bar{x} \pm \sigma$, $\bar{x} \pm 2\sigma$ and $\bar{x} \pm 3\sigma$, where $\bar{x}$ is the sample mean, and to determine the percentage of cases falling in each interval. The data are normal if the percentage of cases falling in interval 1, interval 2 and interval 3 are approximately equal to 68 percent, 95 percent and 100 percent

[77] This non-parametric test is also known as the Wilcoxon Rank Sums test.

[78] There are various tests available on standard statistical packages, for example, the Shapiro-Wilk's W test (Shapiro, Wilk, & Chen, 1968; see also Royston, 1982) and the Kolmogorov-Smirnov test (Lilliefors, 1967). These were not used here, as it was deemed more appropriate to examine the properties of the normal distribution of the features, rather than complex statistical values that require interpretation in light of the properties of the normal distribution.

respectively (McClave & Sincich, 2000: 230). A particular variable was considered to approximately follow normal distribution if normal distribution could be shown, using all four of the above methods. The detailed normality assessments for all of the variables can be found on the accompanying CD.

The next section details the machine learning technique used to train the classifier.

3.7 Support vector machines

As was mentioned in Chapter 1, this research project made use of STATISTICA SVM (Statsoft, 2006). This implementation of SVMs supports both ν-SV and C-SV classification and linear, polynomial, RBF, and sigmoid kernels. It can handle missing data either by mean substitution or by the deletion of cases with missing data. The parameters to be selected are the ν or C values, and the kernel parameters: degree for polynomial kernels; γ for polynomial, RBF, and sigmoid kernels; and bias parameter for polynomial and sigmoid kernels.

The package provides a facility for handling imbalanced data by implementing a penalty for unbalanced classes. It also offers v-fold cross-validation (this will be described in Chapter 4). This package implements caching, scaling and shrinking. Caching and shrinking both reduce computational time. Scaling scales variables' values to fall in a numerical range; in STATISTICA SVM this range is $[0,1]$. This technique is recommended by Hsu, Chang, and Lin (2003:4), in order to improve on performance. For a review of these implementation issues and a detailed explanation, see Kroon (2003:50–54).

This package was selected because it is well documented and supported, and easy to use. Furthermore, it was selected for convenience, as it forms part of a larger statistical package. Thus, all feature extraction, data transformation, normality testing, and feature selection could be done, using the same package.

This section describes SVMs, which is the machine learning technique that this research project employs, in order to train the classifier. First, the learning problem is sketched and the concept of SVMs introduced, embedded in learning theory. Thereafter, SVMs

are described with regard to their basic concept, with some reference to their mathematical representation, drawing on optimisation theory. Support vector machines for both the linearly separable and non-linearly separable cases are detailed with illustrations. The more technical mathematical derivation of SVMs is detailed in Appendix 4.

The basic concept of SVMs was initially developed by Vapnik and Lerner (1963; also Vapnik & Chervonenkis, 1964) and later formalised by Boser, Guyon and Vapnik (1992). Support vector machines have numerous applications in both pattern recognition and regression. This project is concerned with using SVMs to classify texts, which is a pattern recognition task. Other applications are not directly relevant and for this reason are not discussed here. See Burges (1998:121) for a summary of the main studies in the various tasks of pattern recognition and regression, and Kroon (2003:54–59) for a good review of SVM applications.

Support vector machines have been chosen to train the classifier in this research project because the generalisation performance of SVMs has been found to be at least equal to or "significantly better than that of competing methods" (Burges, 1998:121; see also Schölkopf & Smola, 2002:22; Kroon, 2003:55; and Stecking & Schebesch, 2003). The four genre classification studies reviewed in Chapter 2 that made use of SVMs, achieved good performance and in these studies, where SVMs were compared with other machine learning and statistical methods, SVMs were found to achieve the highest accuracy (Dewdney *et al.*, 2001; Argamon & Dodick, 2004a; Meyer zu Eissen & Stein, 2004; and Santini, 2005b). Furthermore, SVMs are found to be robust to noisy data and can generalise well in high dimensional feature space, which is essential to genre classification as many features may prove to be irrelevant or redundant (Hechter, 2004:131).

3.7.1 The learning problem

In Section 2.1, some introduction to machine learning and the problem of this research project was already provided. This introduction to Vapnik's (2000:20) model of learning (see Figure 2.1) is now further developed. The *LM* has already been described as the learner that outputs y^* for any given x such that y^* approximates S's y

response. More generally, this problem of learning can be described as selecting from a set of functions $f(x,\alpha)$, $\alpha \in \Lambda$, where Λ is a set of parameters, the one that best approximates S's response (Vapnik, 2000:17). Selection of this function is based on training examples: training set S_t of ℓ independent and identically distributed observations.[79] This is an important assumption of SVMs and will be discussed in terms of the normal distribution and potential data problems, in Chapter 4.

In order to determine the best approximate to S's response, the loss $L(y, f(x,\alpha))$ between the S's y and $f(x,\alpha)$ of the *LM* needs to be measured. The expected value of the loss is given by the expected risk functional (Vapnik, 2000:18; also Schölkopf & Smola, 2002:66):

$$R(\alpha) = \int L(y, f(x,\alpha)) dF(x,y), \tag{3.6}$$

Thus, the function $f(x,\alpha_0)$ needs to be found. This function minimises (3.6) where the joint probability distribution function $F(x,y)$, as generated by G, is unknown, except for the information available from the training set.

This sketches the general problem of learning, but for the purposes of this project, this needs to be further specified for the problem of pattern recognition. In this case, the S's output y can only be –1 or +1 because the classification task is a binary one (a two-class problem). Similarly, the *LM*'s response $f(x,\alpha)$, $\alpha \in \Lambda$, is a set of functions taking only the values of 0 or 1 (these are called indicator functions). The loss function can now be given by:

$$L(y, f(x,\alpha)) = \begin{cases} 0 & \text{if } y = f(x,\alpha), \\ 1 & \text{if } y \neq f(x,\alpha) \end{cases}. \tag{3.7}$$

The function in (3.6) determines the probability of different answers (classifications) given by the S and the *LM* for the function in (3.7). Re-sketched then, the problem of learning for this research project is to find a function that minimises the probability of classification error, as given by Equation 3.6 (Vapnik, 2000:19).

[79] This is commonly abbreviated to i.i.d.

As the type of problem involved in learning here involves finding a solution/hypothesis function that minimises a function (the loss function), it looks to optimisation theory, which is concerned with the solutions to such problems. The next section expands further on how the problem of machine learning in this project is analysed within the framework of optimisation theory.

3.7.2 The optimal hyperplane classifier and the hard margin classifier

The simplest form of SVMs makes use of a special hyperplane (the generalisation of a straight line to a high dimensional space) classifier called the maximal margin classifier, which separates only linearly separable data. These data can be separated in linear space, that is, by a line. In order to describe this classifier, an explanation of hyperplane classifiers is required. Consider the case where S_t is $(\mathbf{x}_1, y_1),...,(\mathbf{x}_\ell, y_\ell)$, $\mathbf{x} \in \mathbb{R}^n$, and $y \in \{+1,-1\}$, where the values of y indicate binary classification. Thus, the input $\mathbf{x}$ is assigned to a class $y=-1$ if $f(\mathbf{x}) \leq 0$, or to $y=+1$ if $f(\mathbf{x}) \geq 0$. This function, $f(\mathbf{x})$, is a linear function of $\mathbf{x} \in X$ and is given by Cristianini and Shawe-Taylor (2000:9):

$$\begin{aligned} f(x) &= \langle \mathbf{w} \cdot \mathbf{x} \rangle + b \\ &= \sum_{i=1}^{n} w_i x_i + b \end{aligned} \tag{3.8}$$

where $\mathbf{w}$ and b are the parameters controlling $f(\mathbf{x})$; $\mathbf{w}$ is the weight vector perpendicular to the hyperplane and b is the threshold (bias) that moves the hyperplane parallel to its former position.[80] The decision rule, which governs the estimate of the target function $f(\mathbf{x})$, determines that an instance will be classified as -1 if $\mathbf{x} < 0$ or $+1$ if $\mathbf{x} \geq 0$. This situation can be more easily understood if interpreted geometrically as the training space is divided into two (classes) by the hyperplane, defined by Equation 3.8, as illustrated in Figure 3.2.

[80] The vector is a line with direction in space and it is said to be normal to the hyperplane.

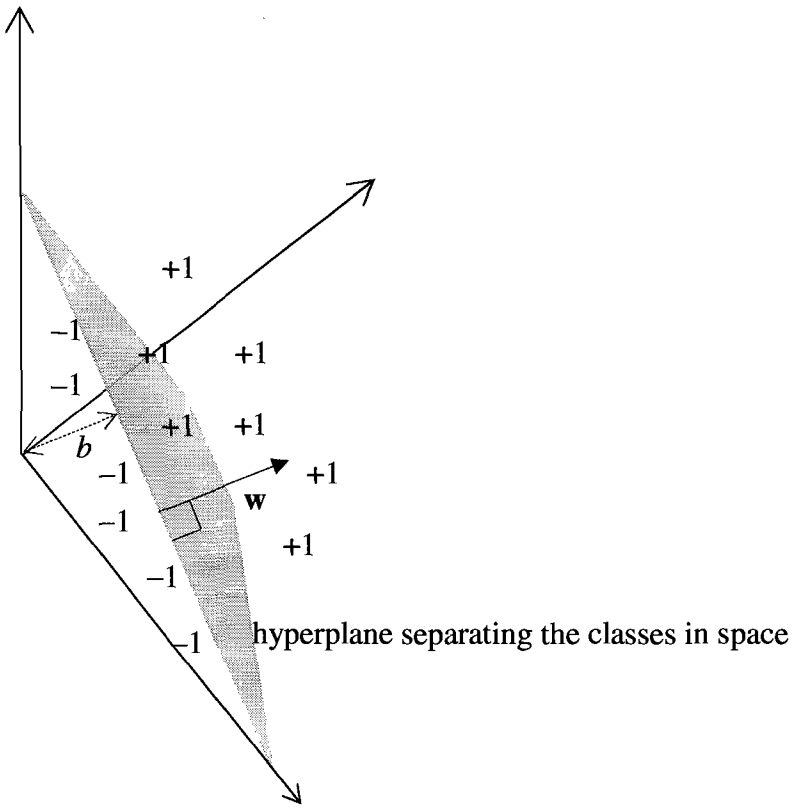

Figure 3.2: A separating hyperplane (adapted from Cristianini & Shawe-Taylor, 2000:10)

This method of learning by a separating hyperplane is called an ill-posed[81] problem because it has many different solutions, not all of which are equally helpful. In order to make this problem well-posed, one solution is to optimise a cost function, which ensures a unique solution (on the condition that a solution exists). This would mean selecting, from a hypothesis space of various hyperplanes, the hyperplane that is a maximum distance from the data points, thus maximising the margin (Cristianini & Shawe-Taylor, 2000:19). This is the maximal margin (optimal) hyperplane that separates the data points (feature vectors) without error and such that the distance between the plane and the nearest vector is maximised. The unique solution is thus characterised by maximum distance and minimum error. The functional margin of a training instance $(\mathbf{x}_i, y_i)$ is defined with respect to the hyperplane $(\mathbf{w}, b)$ as (Cristianini & Shawe-Taylor, 2000:11):

$$\gamma_i = y_i\left(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b\right). \tag{3.9}$$

[81] See Vapnik (2000:233–240) for more on ill-posed problems.

This definition encompasses an inherent degree of freedom in that $\mathbf{w}$ and b can be rescaled from $(\mathbf{w},b)$ to $(\lambda\mathbf{w},\lambda b)$, $\lambda\in\mathbb{R}^+$, without changing the function, while making the functional margin bigger. The solution to this scaling situation implies constraining the value of $\mathbf{w}$. The geometric margin is defined as the functional margin of a normalised[82] weight vector and it remains unchanged if $\mathbf{w}$ and b are scaled as it is scaled by $\|\mathbf{w}\|_2$, where the length of $\mathbf{w}$ is defined in terms of the Euclidean notion of length (Cristianini & Shawe-Taylor, 2000:94–95; also Schölkopf & Smola, 2002:192). This norm is termed the Euclidean or L_2-norm. In order to determine the maximal margin hyperplane, the geometric margin must be maximised. This can be done by minimising the weight vector's (Euclidean) norm and by making the functional margin equal to 1. The latter ensures a constraint on the value of $\mathbf{w}$ and it implies that $\langle\mathbf{w}\cdot\mathbf{x}^+\rangle+b=+1$, and $\langle\mathbf{w}\cdot\mathbf{x}^-\rangle+b=-1$. Using substitution, the geometric margin can be given by:

$$\begin{aligned}\gamma&=\frac{1}{2}\left(\left\langle\frac{\mathbf{w}}{\|\mathbf{w}\|_2}\cdot\mathbf{x}^+\right\rangle-\left\langle\frac{\mathbf{w}}{\|\mathbf{w}\|_2}\cdot\mathbf{x}^-\right\rangle\right)\\&=\frac{1}{2\|\mathbf{w}\|_2}\left(\langle\mathbf{w}\cdot\mathbf{x}^+\rangle-\langle\mathbf{w}\cdot\mathbf{x}^-\rangle\right)\\&=\frac{1}{\|\mathbf{w}\|_2}.\end{aligned}\tag{3.10}$$

The above discussion leads to the actual optimisation problem. The maximal margin hyperplane, which will successfully separate the training set by realising the geometric margin given in Equation 3.10, must minimise the objective function:

$$\Phi(\mathbf{w})=\langle\mathbf{w}\cdot\mathbf{w}\rangle=\frac{1}{2}\|\mathbf{w}\|^2\tag{3.11}$$

subject to the following inequality constraint:

$$y_i\left(\langle\mathbf{w}\cdot\mathbf{x}_i\rangle+b\right)\geq 1,\ i=1,\ldots,\ell\tag{3.12}$$

(Vapnik, 2000:132). This problem is known as the primal optimisation problem.

[82] See Section 3.6 for an explanation of normalisation.

The primal optimisation problem can be solved by using the Lagrangian function. This function is defined as the objective function, in this case, that given in (3.11), and a linear combination of the constraints, in this case those in (3.12) (Cristianini & Shawe-Taylor, 2000:83). The coefficients of this linear combination are the Lagrange multipliers. The primal Lagrangian must be minimised with respect to $\mathbf{w}$ and b (these are termed the primal variables) and be maximised with respect to the dual variables.

This must be done, in order to find the dual representation of the Lagrangian, which can be easier to solve than the primal, and has the same optimal solution as the primal (Cristianini & Shawe-Taylor, 2000:85–86). The dual is relevant for SVMs, as it has the number of variables equal to the size of the learning set rather than to the number of attributes, consequently reducing dimensionality (Abe, 2005:18). In order to find the dual Lagrangian, a saddlepoint must be found and the primal variables eliminated. A saddlepoint indicates the optimal solution of both the primal and dual problem values (Cristianini & Shawe-Taylor, 2000:86). The saddlepoint will ensure the minimisation of the primal variables. The Lagrangian variables can thus be used to solve the quadratic optimisation problem, by solving the value of weight vector $\mathbf{w}$, the geometric margin of γ and the threshold b.

The constraints on the dual variables of the Lagrangian indicate what are termed the *support vectors*. An equality constraint is active if the solution weight vector that satisfies the inequality constraint is equal to zero and is inactive if it is not equal to zero (Cristianini & Shawe-Taylor, 2000:80). *Support vectors* refer to training instances for which the constraint is active. Vapnik (1982) shows that if the data are separated, using an optimal hyperplane the probability of classification error on a test set is bounded by the ratio of the expected number of support vectors and training vectors. The support vectors thus lie on the margin and determine the maximal margin hyperplane. As only the support vectors are used to find the optimal solution, they are the only relevant training instances. Thus, the same optimal hyperplane would be found if only the training instances that are support vectors were used. Furthermore, because these training instances are often very few in number (Cristianini & Shawe-Taylor, 2000:101), it may be the case that for certain optimisation problems, the number of training variables can be significantly reduced, thus resulting in a sparse solution. This

means that such classifiers can have high generalisability "even in an infinite dimensional space" (Cortes & Vapnik, 1995).

The decision rule for new data points (with each point a natural number) is thus that a data point will be classified as -1 if $\mathbf{x} < 0$. Similarly, if $\mathbf{x} \geq 0$, a data point will be classified as $+1$ (Abe, 2005:20). This is illustrated in Figure 3.3 below.

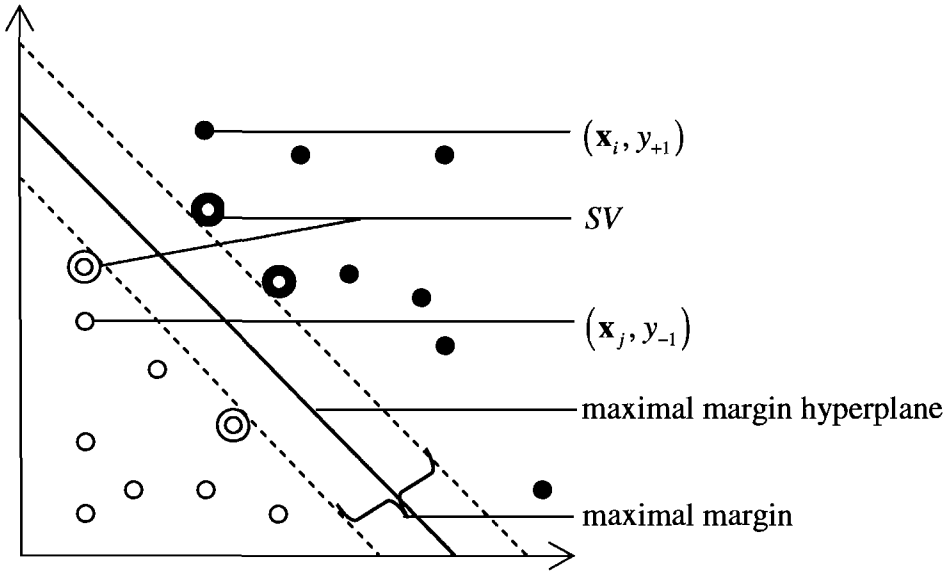

Figure 3.3: A maximal margin hyperplane that perfectly separates S_t, with its support vectors (SVs)

The optimal hyperplane classifier that has been explained in this section, is referred to as the hard margin classifier (Abe, 2005:19; also Hechter, 2004:45). It is the simplest form of the SVM and serves to demonstrate the theory of SVMs. However, for real problems, data are often not linearly separable or it may be that the data are linearly separable overall except for a few data points. The former problem can be resolved by kernels, which are introduced in the following section, and the latter problem can be solved by an extension of the hard margin classifier called the soft margin classifier, which is discussed in Section 3.7.4. The discussion on soft margin classifiers does not follow the discussion on hard margin classifiers directly as kernels are used in the soft margin equations, and some prior explanation of them is thus required.

3.7.3 Kernels[83]

As mentioned above, the optimisation problem for non-linearly separable data cannot be solved, using the hard margin classifier discussed in the preceding section. Kernels provide a way of transforming non-linearly separable data into a high dimensional feature space, in order to separate the data points linearly. This project only made use of linear kernel SVM classification but will report on preliminary test results in Chapter 4, using the Radial Basis Function (RBF) kernel, in order to motivate the choice of linear kernel SVM classification for this research project. Therefore, the discussion that follows, is brief, presenting essentially the basic concept behind kernels; see Cristianini and Shawe-Taylor (2000:26–49), Schölkopf and Smola (2002:200–204), and Hechter (2004:53–59) for more detailed discussion of kernels and their mathematical derivation.

In order to understand how kernels work it is necessary to first explain input space and feature space. In machine learning, pre-processing of data involves selecting a particular representation of the data, in order to suit the particular learning task. Changing the representation of that input data into another representation essentially maps the input variables in the input space X into features in the feature space F. This situation can be represented as follows (Cristianini & Shawe-Taylor, 2000:27):

$$\mathbf{x} = (x_1, ..., x_\ell) \mapsto \phi(\mathbf{x}) = (\phi(\mathbf{x}), ..., \phi_L(\mathbf{x})). \tag{3.13}$$

The variables, which have been described in Section 3.3, are the *input variables*, and their representations in F are termed *features* (Cristianini & Shawe-Taylor, 2000:27).[84] In this context, there are two uses of the term *feature selection*. In the first usage, the term refers to selecting the representation of the input variables; in the second usage, the term refers to detecting and eliminating irrelevant attributes (Cristianini & Shawe-Taylor, 2000:29).[85] Feature selection of the former usage often seeks to reduce the number of features, in order to reduce the dimensionality of the problem. This is generally considered useful as generalisation and computational performance degrade as the feature set increases; this is commonly referred to as 'the curse of dimensionality'

[83] For a detailed discussion of kernels see Schölkopf and Smola (2002: 25–54).
[84] These variables have been termed *features* according to common practice. However, in the context of machine learning this term is not strictly correct. Henceforth, these variables will be termed *attributes* or *input variables*.
[85] It is this latter term that is meant in Section 3.6.

(Cristianini & Shawe-Taylor, 2000:28).[86] The problem can be overcome in SVMs by using kernels (Cristianini & Shawe-Taylor, 2000:28).

The basic premise of kernels is that non-linearly separable data cannot be separated by a linear function in X but can be separated in F.[87] This situation is illustrated in Figure 3.4.

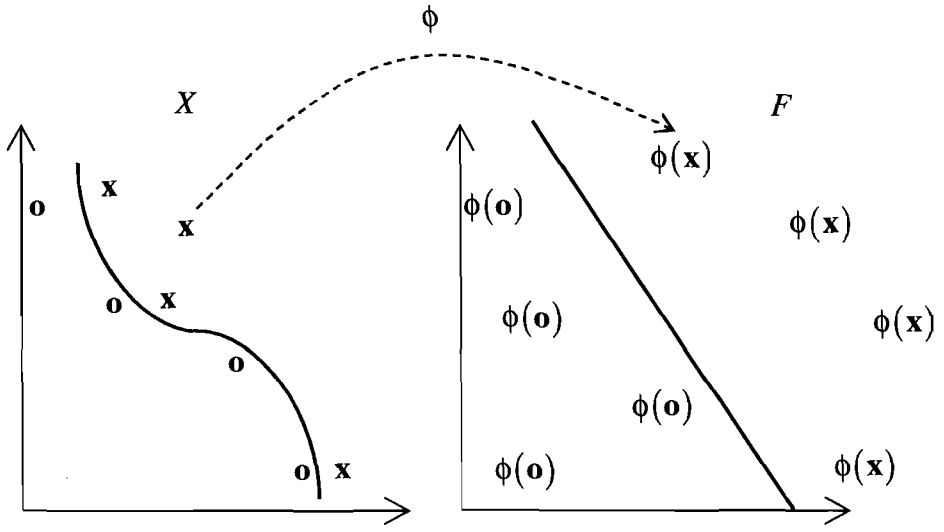

Figure 3.4. Feature mapping $\phi: X \rightarrow F$ (Cristianini & Shawe-Taylor, 2000:28).

This means that non-linearly separable data need to be mapped to a feature space where a linear machine can be used, using a set of non-linear features. The building of non-linear learning machines is thus conducted in two steps. The first step is to transform the data into F, and the second step is to classify the transformed data in this F, using a linear machine (Cristianini & Shawe-Taylor, 2000:30).

The dual representation introduced for linear machines in Section 3.7.2, allows each hypothesis to be expressed as a linear combination of each training point. This means that the decision rule can be evaluated, using the inner products (also dot products) of the test and training points (Cristianini & Shawe-Taylor, 2000:30).

[86] See Bellman (1961).
[87] This is mathematically motivated by Cover's theorem (Cover, 1965).

A kernel function allows for the direct calculation of the inner product in F, as a function of the original input data points. In this way, it merges the two steps of building non-linear machines. It is defined as (Cristianini & Shawe-Taylor, 2000:27):

$$K(\mathbf{x},\mathbf{z}) = \langle \phi(\mathbf{x}) \cdot \phi(\mathbf{z}) \rangle, \text{ for all } \mathbf{x},\mathbf{z} \in X. \tag{3.14}$$

This means that the kernel function will allow for the calculation of the decision function of optimal separating hyperplanes, without needing to know the feature map, that is, knowing K means not needing to know ϕ. This is referred to as the 'kernel trick' (Schölkopf & Smola, 2002:201). It is based on work by Aizerman, Braverman and Rozoner (1964) but was first applied to non-linear SVMs by Boser *et al.* (1992).

As previously mentioned, this project reports on results obtained, using the RBF kernel. These results will be discussed in Chapter 4 and are available for review on the accompanying CD. The next section discusses the extension of hard margin classifiers to soft margin classifiers.

3.7.4 The soft margin classifier

The chief problem with the maximal margin classifier is that its hypothesis solution can be severely affected by any noise, for example an outlier data point. Cortes and Vapnik (1995) introduce a solution to this problem by the introduction of the slack variables. These variables allow the constraints of the margin (3.12) to be violated on condition that there is some increase in the value of the objective function (3.11) at the optimal solution. This allows the training set to be separated with some errors, but without allowing too many misclassification errors (underfitting the data), whilst ensuring that the maximal margin requirements are achieved. Redefinition of the margin in this way results in what is referred to as the soft margin approach (Schölkopf & Smola, 2002:204).

The simplest way of including the slack variables in (3.6) is called the C-SV classifier. The objective function is given by Schölkopf and Smola (2002:205):

$$\Phi(\mathbf{w},\xi) = \frac{1}{2}\langle \mathbf{w} \cdot \mathbf{w} \rangle + \frac{C}{\ell}\sum_{i=1}^{\ell} \xi_i, \; C > 0, \tag{3.15}$$

subject to:

$\xi_i \geq 0,\ i = 1,...,\ell$; and (3.16)

relaxed separation constraints:

$y_i\left(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b\right) \geq 1 - \xi_i,\ i = 1,...,\ell.$ (3.17)

As with the hard margin classifier, the primal optimisation problem for the C-SV classifier can be solved by using the Lagrangian function. Again, the primal Lagrangian must be minimised with respect to the primal variables and maximised with respect to the dual variables. In this way, the dual representation of the Lagrangian is found. The dual is found in the same way as with the hard margin classifier by finding a saddlepoint and eliminating the primal variables. The Lagrangian variables are then used to solve the quadratic optimisation problem, by solving the value of weight vector **w**, the geometric margin of γ and the threshold b. The decision function is similar to that of the hard margin classifier.

The cost parameter, C of the C-SV classifier determines the trade-off between minimising the training error and maximising the margin. The optimal value of C cannot be calculated *a priori*. Instead, it must be selected with prior knowledge of the amount of noise in S_t (Hechter, 2004:51). In order to establish this value of C it is common practice to test performance on a separate test set or v-fold cross-validation, using the training set with a wide range of values for C (Cristianini & Shawe-Taylor, 2000:104; also Abe, 2005:72–74).[88] As this value cannot be calculated *a priori*, a modification was suggested, in the form of the ν-parameter that replaces the C-parameter (Schölkopf, Smola, Williamson & Bartlett, 2000).

The ν-parameter, taking a value between 0 and 1, controls the number of margin errors and *SVs*. The realisation of a soft margin classifier that uses this parameter is called the ν- *SV* classifier. For this classifier the objective function in (3.11) is given by Schölkopf and Smola (2002:206):

[88] V-fold cross-validation and the values used for calculating the optimal C value for this project will be discussed in Chapter 4.

$$\Phi(\mathbf{w},\boldsymbol{\xi},\rho)=\frac{1}{2}\langle\mathbf{w}\cdot\mathbf{w}\rangle-\nu\rho+\frac{1}{\ell}\sum_{i=1}^{\ell}\xi_i, \tag{3.18}$$

subject to the constraints:

$$y_i\left(\langle\mathbf{w}\cdot\mathbf{x}_i\rangle+b\right)\geq\rho-\xi_i,\ i=1,...,\ell,\ \text{and} \tag{3.19}$$

$$\xi_i\geq 0,\ \rho\geq 0. \tag{3.20}$$

As with the hard margin classifier and the C-SV classifier, the primal optimisation problem for the ν-SV classifier can be solved by using the Lagrangian function. Again, the primal Lagrangian must be minimised with respect to the primal variables and be maximised with respect to the dual variables. In this way, the dual representation of the Lagrangian is found. The dual is found in the same way as with the hard margin classifier and the C-SV classifier, by finding a saddlepoint and eliminating the primal variables. The Lagrangian variables are then used to solve the quadratic optimisation problem, by solving the value of weight vector $\mathbf{w}$, the geometric margin of γ and the threshold b. The decision function is similar to that of the hard margin classifier and the C-SV classifier.

This research project used C-SV classification but will report preliminary results on ν-SV classification, in Chapter 4. For this reason, the ν-SV classifier has been briefly described with reference to the C-SV classifier. For a more detailed discussion on the ν-SV classifier see Chen, Lin and Schölkopf (2005), and for information on the differences between these two soft margin classifiers, see Chang and Lin (2001).

Besides the two variants of SVMs discussed above, there are other SVM variants that are not relevant to this study and, therefore, are not discussed here. One widely used variant for text classification, the transductive SVM, which is based on transductive inference (Vapnik, 2000:293), was first introduced and discussed in Joachims (1999b). Good discussions of further variants of SVMs can be found in Abe (2005:129–154) and Kroon (2003:33–41).

3.8 Summary

This chapter provided detailed background of the data and learning methodology used to develop the genre classifier. The data for this study were derived from a corpus of essays, which represent a section of the BAWE corpus. These essays have already been graded by human evaluators. The task now is to select the right types of features that will allow the classifier to learn to assign a similar evaluation to human evaluators.

This chapter discussed all the features deemed potentially relevant as good predictors of prototypical or non-prototypical argumentative essays. The features were selected from those used in previous automatic genre classification studies and research in the field of academic discourse in general. Thereafter the processes of preparing the texts before feature extraction were detailed, and the annotation of features that could not be easily extracted discussed. Then information regarding the standardisation of the dependent variable, the removal of multiple occurrences of terms, data transformation, and feature selection were provided. Lastly, SVMs, the machine learning method that was used to train the genre classifier, was presented for both the linearly separable and the linearly inseparable cases.

Three hypotheses arose from the data and were introduced in Chapter 3. The first **hypothesis** was that text-length normalised frequencies would be likely to yield better results than the raw frequencies as text-length normalised frequencies are less skewed. The second **hypothesis** was that a classifier trained on the CLAWS7 feature set would outperform the classifier trained on the Penn Treebank feature set because the CLAWS7 tagset extracts finer linguistic detail than the Penn Treebank tagset. The third **hypothesis** that has been postulated is that essays of an average grade are linguistically harder to distinguish and that the removal of such essays would improve the accuracy of the classifier. These hypotheses will be further examined in Chapter 4.

Chapter 4 will detail how the data were used in training the SVM classifier, what training parameters were used, and the manner in which they were selected. It will also discuss the results of the classifier, detail well-known evaluation measures, and provide the most suitable accuracy measure for this project.

CHAPTER 4

Results: Training, evaluation and interpretation

SVMs are a rare example of a methodology where geometric intuition, elegant mathematics, theoretical guarantees, and practical algorithms meet
(Bennett & Campbell, 2000:9)

4.1 Introduction

In Chapter 3, the features and motivation for their selection were discussed. The machine learning technique used to train a *LM* to classify prototypical ('good') and non-prototypical ('bad') argumentative essays was also explained in the preceding chapter. The results of training such classifiers are required to be interpreted in terms of the objectives of this research project. As stated in Chapter 1, the classifier is trained to distinguish between instances of prototypical and non-prototypical examples of the argumentative essay genre, and this is done with a practical aim in mind. The classifier is intended to reduce the amount of marking for a senior marker while ensuring that non-prototypical essays are given the necessary attention. Thus, if the classifier succeeds in classifying essays into prototypical and non-prototypical classes, prototypical essays can be given to a junior marker and non-prototypical essays to a senior marker. The assumption here is that prototypical essays require less specialised attention, while the opposite is the case for non-prototypical essays. Moreover, the workload of a senior marker can be reduced by such a classifier, allowing more attention to be given to non-prototypical essays.

In light of these objectives, it can be seen that classifying a 'good' example as 'bad' is undesirable but not detrimental, whereas misclassification of a 'bad' example will mean that a less experienced marker will be assigned the essay. This could prove to be problematic, as 'bad' examples require specialised attention. From this, it is clear that the latter misclassification is the more serious of the two. Therefore, the results presented here are discussed in light of these misclassification errors.

This chapter commences with Section 4.2, in which the selection of the learning parameters used for training the classifier, v-fold cross-validation and theoretical concerns of this technique are detailed. In this section, three potential data problems that

are relevant to this research project, are introduced and, where relevant, their implications and solutions are presented. These problems relate to imbalanced data sets, differing misclassification costs and the normal distribution assumption. Thereafter, in Section 4.3, the most common evaluation measures and indicators are described, and the most relevant measure for this project is presented. Next, in Section 4.4, the results of the classifiers trained and tested are reported and discussed, and the results of the best classifier are determined.[89] Furthermore, comparison of the classifiers built on each data set, using different kernels, features and cases is made. The hypotheses introduced in Chapter 3 are addressed in terms of the results of the classifiers built on the data concerned. In addition to this, the features selected, using the feature selection methods reported in Chapter 3, are discussed. Finally, in Section 4.5, the results presented in Section 4.4 are analysed, and potential explanations for the classifier's performance are examined.

4.2 Training the classifier, using support vector machines

This section discusses the process of training the SVM classifier in terms of the selection of the learning parameters used for training the classifier. The section also addresses potential data problems such as imbalanced data sets, differing misclassification costs and the normal distribution assumption of the data set.

4.2.1 Parameter selection

The parameters that influence the position and orientation of the maximal hyperplane are known as learning parameters (Eitrich & Lang, 2006:427). Selection of these parameters is essential for obtaining an accurate SVM (Hechter, 2004:108). This selection is based on performance measures of various values of the parameters. The values are usually based on a grid search of an interval of values for each parameter (Eitrich & Lang, 2006:428).[90] Each value within this interval is determined according to the increment selected, that is, the amount by which the previous value increases in the following test. Each of the values within the value interval can be tested, using v-fold

[89] The results reported are those obtained on the test sets, unless otherwise stipulated. Also note that the words *train* and *build* are used synonymously.
[90] Exhaustive search of all potential parameter values.

cross-validation or a separate training and testing set. The former is the most commonly used method (Bennett & Campbell, 2000; also Schölkopf & Smola, 2002:217:10).

For v-fold cross-validation, the training set is first divided into v sets, then the LM is trained on $v-1$ sets and validated on the one set excluded from training. This is performed over v training runs, moving one set up for each run so that once the process has been completed each set has served as a validation set. The performance over each run is averaged. This determines the accuracy of the particular parameter value being tested. The parameter value with the highest cross-validation accuracy is then selected as the most appropriate value, and this value is used to train the whole training set. Similarly, in cases where separate training and testing sets are used, the parameter with the highest accuracy is used to train the whole training set.

Unfortunately, there are some theoretical precautions to v-fold cross-validation. Firstly, this approach can lead to overfitting as the parameters are optimised on the same set as the one used for training. Secondly, the optimal parameter settings for the full training set ℓ and the $\frac{v-1}{v}\ell$ set are often not the same; and thirdly, it is possible that there is a phase-transition on the learning curve between ℓ and $\frac{v-1}{v}\ell$ (Schölkopf & Smola, 2002:217). This could result in a large generalisation error size, as a function of the set size between both sets (Schölkopf & Smola, 2002:217).

Schölkopf and Smola (2002:217) suggest that the process of a grid search, using v-fold cross-validation could be avoided by making an educated guess at the optimal parameter values. Among others, they present the following, as methods of making such an educated guess (Schölkopf & Smola, 2002:218):

1. Select parameter settings that have been known to perform accurately for similar problems; and
2. Incorporate prior expectation of the error rate. For ν-SV classifiers, the knowledge of the typical error rate for a particular problem can be incorporated by selecting a value for ν, which is in the test error range. For C-SV classifiers, this knowledge can be incorporated by selecting a large value for C and reducing it until the number of margin errors is in a suitable range (below the error rate).

Unfortunately, both these suggestions proved problematic for this research project. The first suggestion was followed but revealed very little as there is no other research (to the best of my knowledge) that attempts to use genre classification to distinguish between prototypical and non-prototypical student essays, or indeed any kind of academic prose, in the manner that this project seeks to do. The most relevant automatic genre classification studies, which were reviewed in Chapter 2 that use SVMs do not provide details of the steps they followed in selecting their models and parameters. Dewdney *et al.* (2001:5) mention that they make use of SVM^{light}'s default settings, but they do not specify any more detail; and Argamon and Dodick (2004a:5) also state that they use the default settings of the SMO algorithm (Platt, 1999) implemented in the Weka package (Witten & Frank, 1999). The second suggestion could not be implemented, as no direct access to the actual equations in the algorithm used was available.

Despite their cautions regarding the use of *v*-fold cross-validation to determine the optimal parameter settings, Schölkopf and Smola (2002:217) state that excellent results can still be obtained, using the same parameter settings as determined during a grid search, using *v*-fold cross-validation on the whole training set. In following Hsu *et al.* (2003:5–7), this project makes use of a grid search where the values for C are 2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}, 2^{7}, 2^{9}, 2^{11}, 2^{13}, and 2^{15}; the values for ν are 0.08, 0.16, 0.24, 0.32, 0.4, 0.48, 0.56, 0.64, 0.72, and 0.8; and the values for γ are 2^{-15}, 2^{-13}, 2^{-11}, 2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3}, and 2^{5}.

These values were tested, using a 75 percent training set and a 25 percent testing set, for the text-length normalised data and 5-fold cross-validation for the raw data, log-transformed data and the *idf* data.[91] The most commonly used values for *v*-fold cross-validation in the research projects reviewed in Chapter 2 are 5, 10, and 20. This project made use of 5-fold cross-validation because the case set is relatively small (it is comprised of 346 texts). The C value interval that was used for the raw data, log-transformed data and the *idf* data was searched, using an increment of 0.09375, that is, the value of the parameter was incremented by 0.09375 for each 5-fold test. This value

[91] Please note that a seed of 2000 was used for all tests, using a 75 percent training set and a 25 percent testing set. The default seed in Statistica (Statsoft, 2006) is 1000, but better performance was achieved, using a seed of 2000.

was selected because it represents the smallest difference between the value intervals suggested by Hsu *et al.* (2003:5–7), as discussed above. Once the best value for C had been determined this value was tested, using a 75 percent training set and a 25 percent testing set.[92] It is important to note that the best performance on the 75 percent training set and 25 percent testing set was measured by the recall of 'bad' examples. This performance measure was selected in accordance with the objective of the classifier, and is discussed in detail in Section 4.3. Note further that no ν-SV classification was performed on the raw data, log-transformed data, or the *idf* data as this type of classification proved to be less accurate than C-SV classification, both in preliminary experiments and also for the text-length normalised data (the results of the ν-SV classification are discussed in more detail in Section 4.4). For the same reason, only experiments, using the linear kernel were conducted for the raw data, log-transformed data, and the *idf* data. For the text-length normalised data the values of the grid search were tested individually. As STATISTICA SVM (Statsoft, 2006) does not allow v-fold cross validation unless a value interval is specified, the individual parameter values had to be tested, using separate training and testing sets.

Hsu *et al.* (2003:7) further recommend conducting a finer grid search once values that perform best, using the values above, had been determined. Once the individual values that performed best had been determined for all the data sets, a finer grid search was conducted. The values used for this finer search differed for each set; the exact values used can be found on the accompanying CD. These values were once again tested using a 75 percent training set and a 25 percent testing set.

4.2.2 Potential data problems: imbalanced data sets, misclassification costs and the normal assumption

This section introduces three potential data problems that are relevant to this research project. The potential problems raised are those of imbalanced data sets, differing misclassification costs, and the normal distribution assumption.

[92] As previously mentioned, the studies reviewed in this dissertation mainly made use of v-fold cross-validation. Therefore, there was no sense of standard practice on which to base the split used here. However, because of the small data set, a testing set smaller than 25 percent was considered too small (see also Biber, 1993b on representative sample size).

In addition to the problems put forth by Schölkopf and Smola (2002:217), as discussed in Section 4.2.1, Eitrich and Lang (2006:425) point out that the learning parameters are difficult to determine, using a grid search for an imbalanced data set and/or if the cost for a false negative classification is very high while a false positive is acceptable. Both these conditions hold for this project as the 'bad'/'good' essay ratio is 137/209, and the cost for classifying an essay labelled 'bad' as 'good' is high, whereas classifying an essay labelled 'good' as 'bad' is not desirable but acceptable. The terms *false negative*, *false positive*, *true positive* and *true negative* are derived from hypothesis testing, and thus work from the basis of a null hypothesis. As such, these terms are evaluated from the point of view of the null hypothesis or from one class. For this project, evaluation is on the basis of the non-prototypical ('bad') essays. This is illustrated in Figure 4.1 below:

TRUE POSITIVES 'bad' examples classified 'bad'	*FALSE POSITIVES* 'good' examples classified 'bad'
FALSE NEGATIVES 'bad' examples classified 'good'	*TRUE NEGATIVES* 'good' examples classified 'good'

Figure 4.1: Matrix showing classification and misclassification terms

The first problem, that of imbalanced data sets, put forth by Eitrich and Lang (2006:425) above, has a solution in STATISTICA SVM (Statsoft, 2006). This SVM implementation provides a penalty option as a solution for imbalanced data sets (Statsoft, 2006). In the case of this project, 'good' examples were penalised by the ratio of 'good'/'bad' examples, for example, 209/137, which gives a penalty of 1.53 for *v*-

fold cross-validation tests. This means that the *LM* will not classify a particular essay as 'good' simply because there are more 'good' examples.[93]

The subject of the second problem, put forth by Eitrich and Lang (2006:425), is differing misclassification costs. These costs indicate the relative severity of different misclassifications. The standard SVM learner generally assumes these costs are equal, while learning situations in which they are not can be considered non-standard (Lin, Lee & Wahba, 2002). In order to take the differing misclassification costs into consideration for training, the decision surface of the hyperplane needs to be selected to minimise the expected cost of future misclassifications (Hand, 1997:7). The SVM package used for this research project implements SVMs for the standard situation and thus does not allow for the resolution of this problem.[94]

The third potential data problem is the assumption of normality. It should be noted that establishing normality, as discussed in Section 3.6, was important for determining tests to be used for feature selection but that normal distribution is not important for the SVM training itself. In Section 3.7, it was stated that the entire data set used for SVMs are assumed to be independent and identically distributed observations. This has two important implications for SVM training: first, both training and testing data are assumed to be drawn from the same distribution, and second, this distribution is not assumed to be a normal one (Noble, 2006:1566). The first implication can be a problem for SVMs as the training and testing data set may not share the same distribution (see Hand, 1997:9–10). Furthermore, the distribution of the entire data set, which is intended to be a sample of the target population, may not actually share the same distribution as the target population. This is addressed in more detail by Lin *et al.* (2002:192). For this project, the latter problem is not considered to be relevant, as the classification task is not a population problem. It is concerned with the features distinguishing prototypical and non-prototypical examples of the argumentative essay and the distribution of these features, rather than the distribution of prototypical and non-prototypical examples of the argumentative essay in general. For this project, it was **hypothesised** that normalisation of the data in the form of the logarithmic transformation would not affect

[93] See Karakoulas and Shawe-Taylor (1999) for more on imbalanced datasets.
[94] See Lin *et al.* (2002) for a study that addresses SVM learning solutions for the non-standard situation.

the classifier's results, if compared with the classifier trained and tested on the raw data set.

The next section introduces various commonly used evaluation measures. These measures are discussed in terms of their relevance to this project, and the measure indicating good performance for this research project is introduced.

4.3 Evaluation indicators and metrics

The objective of evaluating performance is to determine how well a classifier will perform in classifying new instances (Hand, 1997:9), that is, how well it will generalise. This is considered a foremost concern for this project. However, performance on the training set is also relevant as it provides an indication of the good design of the features as well as the fit of the model. For example, the performance on the training set, if compared with the performance on the testing set, can show whether the model is overfitting the data. A variety of evaluation metrics and indicators are used in the automatic genre classification studies reviewed in Chapter 2, the most common of which are error rate, accuracy, the baseline, improved performance, precision, recall, and the F-measure. This section discusses all of these measures and detail which measures are most relevant for this research project.

The first evaluation metric that is addressed is error rate or misclassification rate, which gives the proportion of objects that were misclassified. In classification studies, it is desirable to reduce the error rate as much as possible, and this measure is therefore the most popular evaluation measure (Hand, 1997:98). It is problematic because it treats all types of misclassification equally. For example, in the case of this project, error rate will treat cases of misclassification of 'bad' examples and misclassification of 'good' examples as equally undesirable. This is problematic because misclassification of 'bad' examples will result in some essays requiring specific attention by a senior marker to be neglected and instead passed on to a junior marker. Therefore, misclassification of 'bad' examples is to be regarded as more serious than misclassification of 'good' examples. As discussed in Section 4.2.2, differing misclassification costs need to be incorporated into the selection of the decision surface. The decision surface, which simply minimises the error rate, arises from the assumption that the misclassification costs are equal

(Hand, 1997:7). This measure is therefore not considered to be helpful for this research project and is not reported on.[95]

The second evaluation metric, that of accuracy, provides a measure of the overall performance of the classifier, that is, of the proportion of correctly classified data and is given by Weiss and Provost (2003:322) as:

$$accuracy = \frac{(tp + tn)}{(tp + fp + tn + fn)}, \tag{4.1}$$

where *tp* are true positives ('bad' examples correctly classified as 'bad'), *fp* are false positives ('good' examples incorrectly classified as 'bad'), *tn* are true negatives ('good' examples correctly classified as 'good') and *fn* are false negatives ('bad' examples incorrectly classified as 'good'). For this research project, accuracy is reported for the training set and the testing set. However, this measure is not sensitive to false negatives in particular, which is the misclassification type that most concerns this project.

Accuracy is often evaluated in terms of a baseline, which is a useful indicator to use in determining whether performance is actually good or just appears to be good. It is determined by assigning the class label of the biggest class to each instance in the test data (Manning & Schütze, 1999:234; see also Kessler *et al.*, 1997). In this research project, the testing baseline for the experiments, using a 75 percent training set and a 25 percent testing set is 63.2 percent. This is derived from the test set, which consists of fifty-five 'good' and thirty-two 'bad' texts. Should the classifier simply label all the texts 'good' this will yield an accuracy of 63.2 percent. This figure is considered the baseline for this project. The training baseline for the experiments, using a 75 percent training set and a 25 percent testing set is 59.5 percent. This is derived from the training set, which consists of 154 'good' and 105 'bad' texts. Should the classifier simply label all the texts 'good' this will yield an accuracy of 59.5 percent. Any accuracy above the training and testing baselines can be considered to indicate an improved performance. However, the baselines are simply reported here for means of comparison with other studies. They are not inherently useful for this project, as they provide no indication of false negatives, that is, 'bad' examples incorrectly classified as 'good'.

[95] All the confusion matrices of the tests conducted can be found on the accompanying CD. The error rate can be calculated from these matrices if required.

In addition to baselines, performance that improves on the accuracies obtained by previous projects, on a particular task, is well established in Natural Language Processing as a means of evaluation (Manning & Schütze, 1999:267). This too, however, will not adequately suffice as a means of evaluation for this research project as there is no previous work of the exact same classification task to which to make adequate comparison.

As mentioned above, accuracy measures are not sensitive to false positives and false negatives. Precision, recall, and the F-measure are metrics that address this issue. Precision is a measure of the proportion of items correctly classified of those that were classified as a particular class and is given by Manning and Schütze (1999:268) as:

$$precision = \frac{tp}{tp + fp}. \tag{4.2}$$

Recall is a measure of the proportion of items classified as a particular class and is given by Manning and Schütze (1999:269) as:

$$recall = \frac{tp}{tp + fn}. \tag{4.3}$$

Precision and recall can be combined into a measure known as the F-measure, a variant of the E-measure that was introduced by Van Rijsbergen (1975:174–175), where $F = 1 - E$. It is given by Manning and Schütze (1999:269) as:

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}, \tag{4.4}$$

where P is precision, R is recall and α is a factor determining the weighting of precision and recall. Depending on the value selected for α, the F-measure can be made more sensitive to either precision or recall.

This project reports on overall recall, but it is rather the recall of 'bad' examples (particularly of the testing set) that is most relevant in evaluating the results of the SVM classifier. This project is the most concerned with false negatives because classifying a 'bad' text as 'good' would mean that the incorrectly classified text would not be given sufficient attention by the senior marker, which would defeat the purpose of the classifier. Therefore, good performance for this project is to be determined by high

recall of 'bad' texts. Naturally, recall of the 'good' examples is also important, as it would not be helpful for the classifier to simply label every text 'bad'. In summary, errors in labelling 'good' examples are more acceptable (yet still undesirable) than errors in labelling 'bad' examples.

The F-measure combines overall precision and recall and, as stated previously, can be adjusted to favour one above the other. However, as this project is concerned with the recall of 'bad' examples specifically and not just overall recall, as described above, the F-measure is not helpful. Therefore, this evaluation measure is not reported on (for an example of a study that makes use of this measure, see Shepard, Waters & Kennedy, 2004).[96]

In addition to the measures already discussed, this research project reports on the average recall for both 'bad' and 'good' examples. It should be noted that *average* does not mean *overall*, rather the *average* is calculated by dividing the sum of correctly classified texts (in percent) by 2. For this average, the baseline is 50 percent. Similar to the accuracy baselines, the average baseline is derived from the test set, which consists of fifty-five 'good' and thirty-two 'bad' texts. Should the classifier simply label all the texts 'good' this will yield a recall of 100 percent for the 'good' texts and a recall of 0 percent for the 'bad' texts. This value divided by 2 will yield a baseline of 50 percent.

The next section presents the results of the SV classifiers built on each data set, using different kernels, features, and cases.

4.4 Results and discussion

This section presents and discusses the results of the SV classifiers. The results for all four data sets are presented in Table 4.1 and are the best results obtained on the recall of 'bad' examples (henceforth RB). Detailed results for all tests using each parameter value, as discussed in Section 4.2.2, are available on the accompanying CD.[97] The discussion of the results given in Table 4.1 follows below.

[96] As with the error rate, the F-measure can be calculated from the confusion matrices on the accompanying CD if desired. The best value for the weight is usually 0.5. If this value is set to less than 1 recall is favoured; conversely, if set to more than 1 precision is favoured (Jurafsky & Martin, 2000:578).
[97] These results are given in percent.

Table 4.1: Results on different data sets for SVM training, using the best parameter values

Data set	Feature set	Type of SV classification	Type of kernel	Cases[98]	RB	RG	AR	Training accuracy	Testing accuracy
RAW	Complete set	C-SV	Linear	All (346)	62.50000	74.54545	68.52273	100	70.115
LOG	Complete set	C-SV	Linear	All (346)	31.25000	89.09091	60.17045	86.873	67.816
IDF	Complete set	C-SV	Linear	All (346)	34.37500	60.00000	47.18750	100	50.575
TLN	Complete set	C-SV	Linear	All (346)	46.87500	72.72727	59.80114	100	63.218
TLN	Feature selection	C-SV	Linear	All (346)	62.50000	74.54545	68.52273	62.934	70.115
TLN	Penn Treebank POS tags	C-SV	Linear	All (346)	40.62500	76.36364	58.49432	70.656	63.218
TLN	CLAWS7 POS tags	C-SV	Linear	All (346)	50.00000	63.63636	56.81818	76.448	58.621
TLN	Top fifty BNC	C-SV	Linear	All (346)	50.00000	61.81818	55.90909	58.301	57.471
TLN	Complete set	C-SV	RBF	All (346)	46.87500	80.00000	63.43750	100	67.816
TLN	Complete set	nu-SV	Linear	All (346)	46.87500	70.90909	58.89205	100	62.069
TLN	Complete set	nu-SV	RBF	All (346)	50.00000	43.63636	46.81818	56.757	45.977
TLN	Complete set	C-SV	Linear	MGR (263)	48.27587	52.94118	50.60852	98.446	50.794

[98] Please note that the number of cases used follow in brackets.

The results of the SV classifier, using each data set are discussed with reference to the text-length normalised data set. This is the set upon which further training and testing were conducted as it proved to perform best and is the most linguistically intuitive. Results for the raw data set are reported here, in order to determine whether it is necessary to normalise the data to the text length or whether satisfactory results can be obtained, using the data as is. This is a worthy pursuit as the text-length normalised data require substantially more manual effort than the raw data set. In addition to this, the results for the raw data set are reported for comparison with the results obtained on the log-transformed data set, in order to determine whether this transformation affects classification performance.

The results of the C-SV classifier built on the text-length normalised data set, using the complete feature set, a linear kernel and with the parameter $C = 0.3125$, show a RB of 46.9 percent and a recall of 'good' examples (henceforth RG) of 72.7 percent. This yields an average recall (henceforth AR) of 59.8 percent, which is a 9.8 percent improvement on the average baseline, as discussed in Section 4.3.

The results of this classifier can be compared with the results of the classifiers built on the log-transformed data and the *idf* data. The results of the C-SV classifier built on the log-transformed data set, using the complete feature set, a linear kernel and with the parameter $C = 0.046625$, have a RB of 31.3 percent and a RG of 89.1 percent. This yields an AR of 60.2 percent, which is a 10.2 percent improvement on the average baseline. However, as this project is concerned with RB rather then RG, it is this value that shows that the classifier built on the log-transformed data set underperformed the classifier built on the text-length normalised data set.

The results of the classifier built on the text-length normalised data set can also be compared with the classifier built on the *idf* data set. The results of the C-SV classifier built on the *idf* data set, using the complete feature set, a linear kernel and with the parameter $C = 1.953125$, have a RB of 34.4 percent and a RG of 60.0 percent. This yields an AR of 47.2 percent, which is 2.8 percent below the average baseline. The RB of this classifier shows a 3.1 percent improvement over the RB of the classifier built on the log-transformed data set. As with the classifier built on the log-transformed data set,

the classifier built on the text-length normalised data set outperformed the classifier built on the *idf* data set. This is shown by a 12.5 percent gain on the RB value on the text-length normalised data set. From this, it is clear that results on the text-length normalised data are better than on the log-transformed data set and the *idf* data set.

As previously noted in Section 4.2, the distribution assumptions made for SVM learning do not assume that the data is normally distributed. Thus, it would be expected that whether the data is normal or not would not affect performance. As discussed in Section 3.6, logarithmic transformation is intended to transform non-normal data so that they approximate a normal distribution. The AR values of the classifier built on the raw data set (68.5 percent) and the classifier built on the log-transformed data set (60.2 percent) differ by 8.3 percent. This difference is fairly substantial, especially if the separate RB values and RG values of the two classifiers are examined. It can be seen that the classifier built on the raw data set yields a RG of 14.6 percent more and a RB of 31.2 percent more than the classifier built on the log-transformed data set. It thus appears that the log-transformation had a detrimental effect on the classifier, which implies that the normality of the distribution has a negative effect on SVM learning. This means that this experiment cannot provide conclusive evidence for SVM learning not being affected by the normality of the data distribution.

As the classifier built on the text-length normalised data set outperformed both the classifiers built on the log-transformed data set and the *idf* data set, as discussed above, all further tests were conducted, using the text-length normalised data set. Initially, tests were conducted by training on subsets of variables. These tests were conducted separately on selected variables, the Penn Treebank tags, the CLAWS7 tags, and the top fifty words of the BNC. Next, tests were conducted, using only the cases without middle-band marks. Thereafter, tests were conducted, training on the complete feature set but using the RBF kernel instead of the linear kernel. Finally, tests were conducted, using the complete feature set but using ν-SV classification (with the linear kernel and the RBF kernel) instead of C-SV classification. First, the results on the raw data set and then the results of the tests as described above are presented.

As previously mentioned, the results for the raw data set are reported here for comparison with the results obtained on the text-length normalised data set, in order to determine whether the gain in recall on the 'bad' examples necessitates the extra human effort required for the latter data set. The results of the C-SV classifier built on the raw data set, using the complete feature set, a linear kernel and with the parameter $C = 1.953125$, have a RB of 62.5 percent and a RG of 74.5 percent. This yields an AR of 68.5 percent, which is an 18.5 percent improvement on the average baseline. These results are favourable and show a 15.6 percent increase in RB over the text-length normalised data set. Furthermore, although both classifiers obtained 100 percent in training accuracy, the classifier trained on the raw data set overfitted the data by a smaller margin than the classifier trained on the text-length normalised data set. It is possible that if the texts are all of a similar length, text-length normalisation would not produce any perceptibly different classification results. However, the length of the essays used in this project range from approximately 800 to 4000 words. Therefore, the results would seem to indicate that text-length normalisation may be an unnecessary step for good classification.

The good results on the raw data set equal the best results obtained on the C-SV classifier built on the text-length normalised data set, using feature selection, a linear kernel and with the parameter $C = 32$, in terms of RB and RG on the testing data set (the feature selection process is discussed in Section 3.6). The major difference between these two classifiers is their training results. As mentioned above, training on the raw data set with the complete feature set resulted in a 100 percent training accuracy. This contrasts greatly with the results of training on the text-length normalised data set with feature selection. The latter classifier resulted in a training accuracy of only 62.9 percent, which may be 3.4 percent above the baseline but which is 37.1 percent below the best training accuracy achieved, indicating substantially poorer overall performance. This provides further evidence for using the raw data set as is.

The low training accuracy also has implications for feature selection itself. The training result shows that although feature selection may have helped improve testing accuracy, the removal of the rest of the feature set was detrimental to training. This means that the classifier trained on the reduced feature set underfitted the training data. It may be that

the features used in this project are all valuable to some extent in discriminating between the essays, and so the removal of the majority of the features resulted in a 37.1 percent loss in discriminatory ability.

In Section 3.6, it was stated that feature selection does not appear to be as essential to SVM learning as to other machine learning techniques, such as decision trees, because SVM learning can learn well in high dimensions. That this is the case is not disputed here, but it does appear that there is some evidence in favour of more careful feature selection. This is so, not because too many features can produce poor results, but rather because too many irrelevant features can produce poorer results, and it is difficult to determine *a priori*, with certainty, which features are irrelevant. A good case is provided for more exploratory techniques, such as cluster analysis, before presenting the features to the *LM*. The higher accuracy results, using feature selection, thus provide some evidence towards the findings of some studies that have reported poor performance of SVMs during learning, with many irrelevant features (such as Weston *et al.* 2001).

Thirty-eight of the 812 features were selected in the manner described in Section 3.6. These are:

1. seven Penn Treebank tags: *JJ*, *NN*, *PDT*, *POS*, *TO*, *VBD*, and *VBG*;
2. eleven CLAWS7 tags: *GE*, *PPHO2*, *PPHS2*, *RP*, *VBDR*, *VBDZ*, *VDD*, *VDI*, *VHD*, *VHN*, and *VVZ*;
3. two quotation tags: quotations non-integrated with text and sentence counts for non-integrated quotations;
4. readability scores: The ARI readability score;
5. number of references;
6. eight from the BNC top fifty: *had*, *they*, *this*, *to*, *up*, *was*, *were*, and *when*;
7. key function words of the top 1000 key words: *this* and *to*;
8. simple prepositions: *to*;
9. two-word complex prepositions: *as to* and *on to*;
10. private factual verbs: *recall* and *find*;
11. subordinating conjunctions: *when*;
12. factual stance adverbs: *actually* and *never*;
13. two-word factual stance adverbs: *in fact*; and

14. factual stance nouns: *result.*

It should be noted that seven of the above features have multiple membership and are thus listed more than once in the above list. These are:

1. *'* and *'s*, CLAWS7 tags and Penn Treebank tags;
2. *had*, CLAWS7 tags and the BNC top fifty;
3. *was*, CLAWS7 tags and the BNC top fifty;
4. *were*, CLAWS7 tags and the BNC top fifty;
5. *when*, subordinating conjunctions and the BNC top fifty;
6. *this*, key function words of the top 1000 key words and the BNC top fifty; and
7. *to*, Penn Treebank tags, key function words of the top 1000 key words, the BNC top fifty, and simple prepositions.

It is interesting to note that the features that were selected are, in the main, function words or indicative of function words (POS tags). Words typically associated with academic writing (as discussed in Chapter 3) such as hedges, downtoners, and various kinds of linking adverbials were not selected as features. Contrary to what is expected, other words linked to academic writing, such as subordinating conjunctions and reporting verbs are composed of only three features: *recall*, *find* and *when*. Of course, this does not indicate that other academic words are not useful indicators of prototypical or non-prototypical argumentative essays. They might simply be less (by varying margins) useful indicators than the features selected. It is also possible that these words would have been selected if other kinds of feature selection techniques were used.

In addition to the above remarks, other significant features were also selected. These features are quotations non-integrated with the text, sentence counts for non-integrated quotations, the ARI readability score, and the average number of references. The first two of these features are unique to this research project, and it therefore appears to be promising that they were selected as good discriminators of prototypical and non-prototypical argumentative essays. This means that they can be considered useful for future projects (ideas in this regard will be developed further in Chapter 5). The ARI readability score indicates that readability measures can provide useful information for

classifying prototypical and non-prototypical argumentative essays. This also encourages further exploration of readability measures (see Chapter 5).

The selection of the average number of references as a good classifying feature is also encouraging. This is because it is expected that prototypical argumentative essays will put forth opposing or collaborative arguments and provide evidence for their argumentation, which would entail references to other research. Therefore, it makes linguistic sense that this feature is considered highly discriminatory.

From the feature selection, it can also be deduced that the Penn Treebank tags, the CLAWS7 tags, and the top fifty words of the BNC are very useful for discriminating between prototypical and non-prototypical essays on this data set. The discriminatory ability of POS tags and the top fifty words of the BNC for automatic genre classification (as discussed in Chapter 2) is thus confirmed to some extent by the feature selection. Additional evidence for the usefulness of these two feature sets is presented in the form of results from three C-SV classifiers built on the text-length normalised data set, using the Penn Treebank tags, the CLAWS7 tags, and the top fifty words of the BNC respectively.

The first classifier built on the Penn Treebank tags, using a linear kernel, and with the parameter $C = 8.09375$, yields a RB of 40.6 percent and a RG of 76.4 percent. This results in an AR of 58.5 percent, which is an 8.5 percent improvement on the average baseline. The second classifier built on the CLAWS7 tags, using a linear kernel and with the parameter $C = 32$, yields a RB of 50.0 percent and a RG of 63.6 percent. This results in an AR of 56.8 percent, which is a 6.8 percent improvement on the average baseline. The third classifier built on the top fifty words of the BNC, using a linear kernel and with the parameter $C = 128$, yields a RB of 50.0 percent and a RG of 61.8 percent. This results in an AR of 55.9 percent, which is a 5.9 percent improvement on the average baseline.

These last two classifiers, built on the CLAWS7 tags and the top fifty words of the BNC, show an improvement on RB in comparison with the results obtained on the text-length normalised data set, using the full feature set. Used on their own, however, these

two feature sets do not have as high a discriminatory ability on the prototypical texts. It can thus be deduced that for this data set the CLAWS7 and the top fifty words of the BNC feature sets are useful for classifying non-prototypical texts, and that the Penn Treebank feature set is more useful for classifying prototypical texts. This is important for this project, as too many false positives would also be unhelpful for the user of such a classifier. This is so because a large number of false positives will result in many essays misclassified as 'bad'. This in turn will mean that more essays will be required to be marked by a senior marker. As the purpose of this classifier is to reduce the amount of unnecessary work for a senior marker, misclassifications that increase the amount of work for a senior marker will defeat the purpose of the classifier.

The tests discussed above, which examined the discriminatory power of the Penn Treebank and CLAWS7 tags also provided partial evidence for the **hypothesis** put forth in Section 3.5.1. This hypothesis suggested that the CLAWS7 tagset, which extracts finer linguistic detail than the Penn Treebank tagset, might extract potentially useful discriminatory features. It suggested further that a classifier trained on this tagset would report more accurate results than a classifier trained on the Penn Tree tagset. If the results of the classifier trained on the CLAWS7 tagset are compared with the results of the classifier trained on the Penn Tree tagset, it can be seen that the discrimination ability for the non-prototypical examples is substantially improved by using the CLAWS7 tagset. As discussed above the classifier trained on the CLAWS7 tagset is better at classifying non-prototypical examples than the classifier trained on the Penn Treebank tagset, but worse at classifying prototypical texts. This implies that the hypothesis can be accepted, if better performance is defined by RB only. However, linguistically, it would be expected that a classifier trained on the CLAWS7 tagset would perform better than one trained on the Penn Treebank tagset in terms of both RB and RG. This is the intended interpretation of the hypothesis, which means that as the classifier trained on the CLAWS7 tagset did not perform better than the classifier trained on the Penn Treebank tagset for both classes, it can only be accepted partially.

In addition to tests conducted, using different feature sets, tests were conducted, using the complete feature set but with a reduced selection of cases. These cases contain only those essays with upper- and lower-band marks, which were selected according to the

procedure detailed in Section 3.6, resulting in 263 cases for training and testing. The results of the classifier trained and tested, using this reduced set of cases, using a linear kernel and with the parameter $C = 0.5$, yields a RB of 48.3 percent and a RG of 55.9 percent. This yields an AR of 52.1 percent, which is 2.1 percent improvement on the average baseline. This can be compared with the results obtained for the classifier trained and tested on the text-length normalised data set, using the full case set (346 instances). Recall from the first results presented in this section, that this classifier has a RB of 46.9 percent and RG of 72.7 percent, yielding an AR of 59.8 percent. This shows that on RB the classifier trained and tested on the upper- and lower-band essays outperformed the full case set classifier on the RB values by 1.4 percent.

This result is difficult to interpret, as on the one hand it may mean that the **hypothesis** presented in Section 3.6, which proposed that the linguistic differences of 'average' grade essays are not sufficiently discriminant to determine prototypical or non-prototypical essays, is to be accepted. One the other hand, the RG on the classifier trained and tested on the upper- and lower-band essays is 16.8 percent worse than the RG on the classifier trained and tested on the full case set. This implies that the hypothesis is to be rejected. Furthermore, the results of the classifier trained and tested on the upper- and lower-band essays are being compared with a classifier trained and tested on more instances. Naturally, the larger the case set the more likely it is that the classifier will classify training and testing instances more accurately.

This idea was examined by training and testing ten classifiers, using the text-length normalised data set, on the complete feature set and a linear kernel, but with successive classifiers trained and tested on 10 percent more of the case set each time. This meant that the results for classification on 10 percent, 20 percent, 30 percent, 40 percent, 50 percent, 60 percent, 70 percent, 80 percent, 90 percent, and 100 percent of the case set were available for comparison.[99] As previously mentioned, the classifier trained and tested on the upper- and lower-band grade instances was trained and tested on 263 instances. This corresponds to 76.0 percent of the case set and means that these results

[99] These sets were all sampled, using stratified random sampling, implemented in Statistica (Statsoft, 2006). This technique samples both the 'good' and 'bad' examples randomly, while retaining the same class proportions as those of the complete set. This technique is never less representative and often more representative than simple random sampling (Biber, 1993b:244).

can be reasonably compared with the results of the classifier trained and tested on 80 percent of the case set. The results of this classifier, using the parameter $C = 0.5$, yields a RB of 44.8 percent and a RG of 64.4 percent. This yields an AR of 54.6 percent, which is 4.6 percent above the average baseline. The performance shows a 3.5 percent increase in RB for the classifier trained and tested on the upper- and lower-band grade cases. However, this classifier also performed more poorly than the classifier trained and tested on 80 percent of the case set in terms of RG. Again, this provides mixed evidence as to whether the performance of the classifier trained and tested on the upper- and lower-band grade cases was necessarily due to case set size.

It is possible that the results on both these classifiers are not due to overall case size but rather to the amount of texts for each class. In comparison, the 80 percent case set contains 125 training and 45 testing examples of 'good' texts, and 94 training and 29 testing examples of 'bad' texts, whereas the data set, using the upper- and lower-band grade cases has 96 training and 34 testing examples of 'good' texts, and 97 training and 29 testing examples of 'bad' texts. This means that the second classifier had less instances upon which to learn 'good' examples, but more instances upon which to learn 'bad' examples. This may partly explain the results on RG and RB.

Evidence against the **hypothesis** presented in Section 3.6, that the linguistic differences of 'average' grade essays are not sufficiently discriminant to determine prototypical or non-prototypical essays, is given if the particular types of mistakes made on the training and testing on the full case set are considered. Fifteen out of the thirty-two mistakes made on the testing set are not expected according to the hypothesis, that is, the misclassified essays fall decidedly in the top band of prototypical argumentative essays and in the lower band of the non-prototypical argumentative essays. Similarly, forty out of the ninety-six mistakes made on the training set are not expected. That 46.9 percent for the testing set and 41.7 percent for the training set are not expected mistakes is additional evidence against the hypothesis that essays with middle-band marks are less easily distinguishable. This would seem to imply that either the features used do not adequately extract the linguistic differences between the essays, or that more training data are required. This is addressed further in Section 4.5.

After testing, using different cases, tests were conducted using a RBF kernel rather than the linear kernel. The RBF kernel was used because it is one of the most commonly used kernels for SVM classification (Hechter, 2004:56; also Statsoft, 2004) and is recommended as a first choice by Hsu *et al.* (2003:4). The results of this C-SV classifier built on the text-length normalised data set, using the complete feature set, with the parameter $C = 31.8125$ and with the parameter $\gamma = 0.013139006$, have a RB of 46.9 percent and a RG of 80.0 percent. This yields an AR of 63.4 percent, which is a 13.4 percent improvement on the average baseline. These results are better overall than the results obtained on the classifier trained on the text-length normalised data set, using the complete feature set but with a linear kernel because the former classifier achieved a RG of 7.3 percent more than the latter classifier.

Finally, tests were conducted, using ν-SV classification rather than C-SV classification and using the linear and RBF kernels. The results of the ν-SV classifier trained on the text-length normalised data set, using the complete feature set, a linear kernel and with the parameter $\nu = 0.08$, have a RB of 46.9 percent and a RG of 70.9 percent. This yields an AR of 58.9 percent, which is 8.9 percent above the average baseline. In comparison with the C-SV classifier trained on the text-length normalised data set, using the complete feature set and a linear kernel, the RB is equal but the RG is less by 1.8 percent.

The results of the ν-SV classifier trained on the text-length normalised data set, using the complete feature set, a RBF kernel and with the parameter $\nu = 0.08$, have a RB of 50.0 percent and a RG of 43.6 percent. This yields an AR of 46.8 percent, which is 3.2 percent below the average baseline. This classifier performed better in RB than either the C-SV classifier trained on the text-length normalised data set, using the complete feature set and a RBF kernel, or the C-SV classifier trained on the text-length normalised data set, using the complete feature set and a linear kernel. However, the RG of the ν-SV classifier is very poor: it is 36.4 percent below the RG of the C-SV classifier trained on the text-length normalised data set, using the complete feature set and a RBF kernel. Furthermore, it is also 29.1 percent below the RG of the C-SV classifier trained on the text-length normalised data set, using the complete feature set

and a linear kernel. As previously discussed, such a low RG will defeat the purpose of the classifier.

The next section analyses why the results of the best classifier did not reach 100 percent recall of the 'bad' and 'good' examples.[100]

4.5 Analysis of the classifier's performance

From the results and discussion presented in Section 4.4 above, it is apparent that the best results, those of the classifier trained on the text-length normalised data set, using feature selection, a linear kernel and $C = 32$, are quite good in terms of both RB, as well as the average and testing baselines. Furthermore, if judged in the light of the difficulty of the classification problem, which examines a very subtle genre type, these results are very good indeed.

However, it is obviously still desirable for the classifier to have 100 percent recall of 'bad' and 'good' examples. This section suggests seven potential explanations for why the classifier did not perform more accurately and proposes several ways to improve on the classifier's performance. The following reasons are suggested:

1. The argumentative essays are not truly representative of non-prototypical examples.
2. The training set size could have been too small.
3. Cross-domain classification could have reduced accuracy.
4. Personal language differences could have confused classification.
5. The features used could not have adequately extracted the linguistic differences between the classes.
6. The feature selection might not have been adequate.
7. More knowledge of the discourse 'struggle' might need to have been considered.

The first reason suggested for the classifier's performance is that the argumentative essays are not truly representative of non-prototypical examples. This reason is listed first because it is the most likely explanation for the classifier not being able to identify

[100] The code for this classifier can be found on the accompanying CD.

non-prototypical examples as easily as prototypical examples. Unfortunately, the essays this project uses are actually considered instances of good student academic writing (and thus argumentative essays) by the BAWE corpus compilers. The lowest grade for the essays in the corpus is 60 percent, which in comparison with a grade of 80 percent can be considered poor but in actuality is average rather than poor. This by no means implies that the whole corpus is average, as the majority of the essays are in the region of 68 percent and above. However, it is likely that the classifier would be able to better distinguish between the two classes, using the current features if the classes were more readily distinguishable, that is, if essays of fail marks made up the non-prototypical case set.

This problem could not be rectified for this study, as very few L1 English argumentative student essays are easily available for study. Furthermore, where such corpora are available, they do not always have the necessary features for training. For example, the LOCNESS (Granger, 1994) has had quotations removed from the texts, an important feature used in this research project.

The second possible reason for the classifier not having achieved 100 percent RB and RG is a concern that was already raised in Section 2.3.2, that of adequate data set size. It is possible that the data set that was used to train the classifier could have been too small to allow optimal performance. In other words, the classifier may require more learning examples, in order to generalise well. It is often the case that accuracy is a function of the size of the training set. For example, Stamatatos *et al.* (2000b:491) show that the classification accuracy of their problem does improve with more training data. The best way of establishing whether accuracy is a function of training size is to train the classifier on an increasing number of texts. However, this approach requires the same testing set, which was not possible with the SVM implementation used for this project. This is because the training and testing examples are selected randomly by the algorithm.

The third reason, that cross-domain classification could have reduced accuracy, is first preceded by the background to this problem. In Section 4.3, an evaluation indicator was introduced in the form of performance that improves on the accuracies obtained by

previous projects, on a particular task. It was noted that this means of evaluation could not be used directly for this research project as there is no previous work of the exact nature to which to make adequate comparison. However, it is clear that previous studies have provided some guidance in terms of features to be used, data transformation, and machine learning techniques. These studies were selected on the basis of their overall high performance. One study to which more direct comparison can be made in terms of evaluation of results is Finn's (2002) study.

This is for two reasons, first because it is a two-class classification problem, as with this study, unlike the majority of all the studies reviewed in Chapter 2, which are concerned with multi-class classification. And second, because it is a more subtle classification task, more like the task of this project, than for example the letters and poems of Kelih *et al.*'s (2005) study, which are two genres that differ so widely as to render classification errors unlikely. This latter type of genre classification task is fairly representative of many of the other types of genres examined in the studies reviewed in Chapter 2.

In addition to suitability because of classes and type of classification, Finn's (2000) study also focuses specifically on cross-domain performance, something that is not an explicit focus of any other study, as discussed in Section 2.3.2. As previously remarked, this focus on cross-domain accuracies is relevant to this research project where all the data are drawn from and classified across domains. It is thus expected that because the texts are drawn from across very different domains, accuracy will be negatively affected (see Finn, 2002:86). The idea that cross-domain classification adversely affects classification is thus put forth as an explanation of why the best SVM classifier did not achieve 100 percent recall of 'bad' and 'good' examples.

Further evidence can be provided in this regard by Karlgren and Cutting's (1994:1072) poor results on the *learned/humanities* subset. The texts in this subset were instances of academic prose written within disciplines in the humanities faculty. As this project made use of texts of this nature, drawn not only from various disciplines in the humanities faculty but also from across all faculties except Engineering (see Section

3.2), this provides further evidence for cross-domain accuracy reducing classification accuracy.

Wolters and Kirsten (1999) provide strong evidence that different word types (function and content words, versus function words only) are not equally discriminatory for texts from the humanities and from science and technology. This too shows that accuracy, particularly in academic discourse, is negatively affected by cross-domain classification.

The fourth explanation is that personal language differences could have confused classification. Van Halteren, Baayen, Tweedie, Haverkort and Neijt (2005) argue for the existence of a human stylome, that is, idiosyncratic stylistic differences in writing that varies from individual to individual. They prove that a human stylome exists in the framework of authorship attribution studies, and find that "the differences between the 'personal' language versions of even nonspecialist writers are greater than expected so far" (2005:73).

As the essays used in this research project are written by many different authors, it is also quite possible that the individual stylistic variations, as established by Van Halteren *et al.* (2005), could play a role in classification. In other words, differences that are not simply indicative of prototypical or non-prototypical examples but also of personal language styles were also extracted. Naturally, this would result in lower classification accuracy in terms of this classification task.

The fifth potential explanation for why the classifier did not perform more accurately is that the features used could not have adequately extracted the linguistic differences between the classes. This is the most obvious and simple explanation, because identifying good discriminant features is one of the main concerns of automatic genre classification. Since the choice of highly discriminant features is central to automatic genre classification, solutions to selecting better features are important for future work. Therefore, suggestions in the form of different kinds of features that could be added, as well as different combinations of current features will be made in Chapter 5, as recommendations for future work.

The sixth explanation is related not to overall feature choice, but rather to the selection of a subset of features from the main feature set. It is suggested that the feature selection that was made, might not have been adequate (see Yang & Pedersen, 1997 for more on other feature selection techniques).

The seventh explanation is that more knowledge of the discourse 'struggle' might need to have been considered. Recall that in Chapter 1 the stance of this research project towards student academic writing entails viewing academic writing within disciplines as sites of discourse and identity struggle. Nystrand (1987:204) claims that any good text analysis needs to distinguish between "the structure of argument and the structure of communication". He explains these two notions in terms of the metaphor of a 'contract' writers have with readers. This 'contract' requires writers to establish shared knowledge, contextualise new information, and mark text boundaries, in order to indicate conceptual and narrative shifts. This project attempted to extract relevant information relating to the linguistic features of argumentative essays. It was done with the assumption that such features of writing are indicative of successful essays. It is quite possible, however, that in addition to this success in structuring argument, there may be evidence of a communicative role with the reader-marker, which has not been extracted and could provide more accurate classification.

Furthermore, Swales (1990:54) cautions that knowledge of the underlying logic of gatekeeping (discourse norms) is also very important and that both surface features and underlying discourse norms are crucial to communicative success (and perhaps entry into the discourse community). As this project examined only surface features, it is possible that further knowledge and incorporation of the discourse norms and context would have allowed for more accurate classification of the essays. Attempting to trace and extract the structure of communication, as well as incorporating discourse contexts will require further research; this may prove troublesome, for Nystrand (1987:205) claims that a writer must strike "an effective balance between what needs to be said and what may remain unsaid".

The next section provides a summary of the main results of the classifier, as well as the main conclusions reached regarding the classifier's performance.

4.6 Summary

This chapter began by establishing the importance of selecting learning parameters, in order to obtain high generalisability. To this end, an overview of the grid search values used to determine the best learning parameters for this project was presented. As grid searches are often conducted in the context of *v*-fold cross-validation, this validation technique was elucidated. Furthermore, some theoretical concerns about the validity of this technique were delineated and alternatives to this technique were put forth, following Schölkopf and Smola (2002:217). Counters to why these suggestions could not be followed for this research project were also discussed.

In the following section, three potential data concerns were addressed. The first was the problem of imbalanced data sets. A solution to this in the form of a penalty for the 'good' examples was provided, in order to prevent the *LM* assigning the label of the larger class ('good') to any particular text simply because there are more 'good' examples.

The second problem that was addressed was differing misclassification costs. This refers to the relative severity of different misclassifications. It was placed in light of the difference in errors for this project, where the cost for classifying a 'bad' example as 'good' is more serious then classifying a 'good' example as 'bad'. The problem could not be resolved for this research project because the SVM package used offers SVM classification in the standard situation.

The third problem that was raised was the normal distribution assumption of the data set. The problem is related to the SVM learning method, which assumes that both training and testing data are drawn from the same distribution, but which does not assume that this distribution is normal. The assumption established that the data are not required to be normally distributed. From this, it was **hypothesised** that data normalised by way of the logarithmic transformation would affect the classifier's results.

Next, the following evaluation measures were presented: error rate, accuracy, the accuracy baselines, improved performance, precision, recall, the average recall baseline,

and the F-measure. In this section, it was shown that the recall of 'bad' examples is the most relevant measure for this research project.

The results were then presented in terms of this measure in the following section. The results of the classifiers trained on various data and feature sets, using C-and ν-SV classification, and the linear and RBF kernels were compared. The best performance of 62.5 percent RB was reported on a 75 percent training set and 25 percent testing set. This C-SV classifier was trained on the text-length normalised data set, using feature selection and a linear kernel.

In the comparisons drawn between the results of the different classifiers, various hypotheses were addressed. These hypotheses were raised in Chapters 3 and 4. The first **hypothesis** was that text-length normalised frequencies would yield better results as frequencies would be less skewed. However, it was found that the text-length normalised frequencies, although linguistically and statistically motivated, appeared to yield no better results than the raw frequencies. The raw data set results were also easier to compute, as no feature selection was required, unlike that of the text-length normalised data set.

The second **hypothesis** was directly assumed from the distribution assumption addressed earlier. It is expected that, because the data set is not assumed to be normal, normalisation of the data should not affect the classifier's results. If evaluated in terms of AR the results of the classifier trained on the text-length normalised data set and the classifier trained on the log-transformed data set are approximately the same. However, RB values of the two classifiers differed, which meant that the hypothesis could not be accepted or rejected.

The third **hypothesis** was that the classifier trained on the CLAWS7 feature set would outperform the classifier trained on the Penn Treebank feature set because the CLAWS7 tagset extracts finer linguistic detail, than the Penn Treebank tagset. This hypothesis was only partially accepted because, on the one hand, the recall of 'bad' examples was much improved by using the CLAWS7 tagset, which is the most relevant measure of

evaluation for this project. On the other hand, the recall of the 'good' examples was higher when the Penn Treebank tagset was used.

The fourth **hypothesis** put forth was that the middle-grade essays are linguistically harder to distinguish and that the removal of these essays would improve the accuracy of the classifier. This hypothesis could not be rejected or accepted because, although the classifier trained and tested on the set of middle-grade essays outperformed the classifier trained and tested on all the essays on RB, it was outperformed by the classifier trained on all the essays on RG. That sample size could have an affect on learning, was tested. This was disproved by tests conducted on a similar size data set, with instances of upper-, middle- and lower-band essays. Rather, it was found that the number of instances of each class that the classifier was trained and tested on played a role in classification performance. Furthermore, the types of errors made on the classifier trained and tested on all the instances were examined. This showed that errors made on 46.9 percent of the testing set and 41.7 percent of the training set were not expected, that is they were errors made on essays that were clear cases of upper- and lower-band grades.

In addition to examining these hypotheses, feature selection was also explored. In Chapter 3, it was stated that feature selection does not appear to be as essential to SVM learning as to other machine learning techniques. This was not invalidated here, but rather, because of the favourable results obtained, using feature selection, evidence in favour of feature selection was presented.

In the final section of this chapter, the best classifier's performance was analysed and seven potential explanations were provided to account for the less than 100 percent recall. In summary, these reasons were that the argumentative essays are not truly representative of non-prototypical examples, the training set size could have been too small, cross-domain classification could have reduced accuracy, personal language differences could have confused classification, the features used could have not adequately extracted the linguistic differences between the classes, the feature selection could not have been adequate, and more knowledge of the discourse 'struggle' might need to have been considered.

Chapter 5 will provide a conclusion to this dissertation by presenting a review of the findings of this research project. Thereafter, implications of this study for genre classification and academic writing will be discussed, and in light of these, recommendations for future research will be made.

CHAPTER 5

Conclusion and recommendations

A picture says more than a thousand words, and for me the issue is how to listen to the picture the text paints, without being distracted by its words
(Karlgren, 2000:131)

5.1 Introduction

Chapter 5 provides a conclusion to this dissertation. This chapter commences with a detailed summary of the preceding chapters in Section 5.2. In this section, the research questions posed in Chapter 1 are discussed, in order to determine whether this study has answered them adequately. Thereafter, in Section 5.3, the results and findings presented in Chapter 4 are discussed, with particular reference to the hypotheses presented in Chapters 1, 3, and 4. The acceptance or rejection of these hypotheses, as discussed in Chapter 4, is also reported in this section. Section 5.4 concludes this chapter with recommendations for further studies on genre classification and academic writing, which are made on the basis of the findings of this study.

5.2 Summary of chapters

In order to determine whether an automatic genre classification approach could be used for the classification task of this study, a detailed review of the state-of-the-art of this approach was provided in Chapter 2. The notions of machine learning and supervised learning, necessary to understanding the perspective of the classification task that this research project takes, were introduced along with the basic notation used for the machine learning process. These concepts were placed in the framework of automatic genre classification and the genre task of this project. Thereafter, the use of genre in this project was elucidated and the genre examined in this project explained.

Once the preliminaries to understanding the nature of the classification task of this research project were established, studies in automatic genre classification were reviewed. This review was in three parts. In the first section, Biber's (1988) language variation study, which forms the background to studies in automatic genre classification, and two seminal works in the field of automatic genre classification were reviewed. In the second section, relevant contemporary automatic genre classification

studies were reviewed. These studies were deemed relevant to this research project regarding application, corpus, features, or methodology. The last three aspects are essential to any text classification task; therefore, these three aspects were detailed for each study. This was done for two reasons: first, to determine contemporary practice in the field of automatic genre classification, which partly established whether an automatic genre classification approach would suit the classification task of this research project. Second, to determine the potential value of these studies for informing the features, methodology and interpretation for this research project through critical assessment of the validity of pre-defined genre classes, results, evaluation measures, and the features of each study. This partly answered the second research question, by determining potential linguistic features that could be easily computed and used for this research project.

The third section of the review reported on automatic genre classification studies that use SVMs for the purposes of genre classification. Studies using this technique were reviewed in a separate section because this machine learning technique had demonstrated the best results in comparison with other techniques in automatic genre classification tasks. In this manner, the expected performance of SVMs, as a machine learning technique, for this research project was determined and the fourth research question was thus partly addressed.

In Chapter 3, the features and machine learning technique used in this research project were reported. It provided the background to the corpus and the data derived from the corpus. The features presented in this chapter were selected from studies reviewed in Chapter 2 and research in the field of academic discourse. They were deemed potentially relevant as good predictors of prototypical or non-prototypical argumentative essay. This provided the foundation for Chapter 4, and thus began to address the first research question.

The processes of text preparation, annotation of the features, removal of multiple instances of features, data transformation, and pre-classification exploration of the feature set were also detailed in Chapter 3. The texts in the corpus were prepared before features extraction by removal of formatting, essay questions, essay titles, bibliography,

appendices, headings, footnotes, graphs, illustrations, tables, some punctuation, and equations. The character set of the texts was converted, in order to be compatible with the language models used in the SVMTagger (Giménez & Màrquez, 2004). In addition to this preparation, apostrophes and quotation marks were also standardised, in order to remove any ambiguity of single closing quotation marks and apostrophes, and to separate actual citations from play titles and book titles. Lastly, the texts were tokenised.

The corpus was POS tagged, using SVMTagger (Giménez & Màrquez, 2004) and the CLAWS4 tagger (Garside, 1987:30). These two taggers were used to extract POS tags from the Penn Treebank and the CLAWS7 POS sets. This was done to determine whether coarse or fine POS distinctions would be more helpful as genre revealing features. Next, sentences, paragraphs, quotations, references, punctuation marks, nominalisations, two-word complex prepositions, and three-word complex prepositions were marked up, using XML tags. This was done, in order to extract features that could not be directly extracted from the text. Features were extracted, using STATISTICA Text Mining and Document Retrieval (Statsoft, 2006).

Thereafter, the value of the dependent variable was standardised to percentage grading, and the data were cleaned by removing multiple occurrences of terms and thus double counts of single features. Next, the data were normalised to text length (as measured in words) and transformed using logarithmic transformation and inverse document frequency transformation. This resulted in four data sets: a set of raw data frequencies, a set of text-length normalised frequencies, a set of log-transformed frequencies, and a set of frequencies.

Chapter 3 also addressed the process of feature selection, which removed potentially irrelevant features by selecting those features that contribute most to the classification of the texts. This feature selection partly addressed the first research question. The chapter concluded by presenting SVMs for the linearly separable and the non-linearly separable case. In this section, it was shown that in theory, this machine learning technique provides good generalisability and is manually undemanding, thus partially addressing the fourth research question.

The research questions that were only partially addressed in Chapters 2 and 3, as discussed above, were fully attended to in Chapter 4. This chapter established the importance of selecting learning parameters, in order to obtain high generalisability. As grid searches are usually used to select the optimal parameters (Eitrich & Lang, 2006:428), this project too conducted a grid search of different values, in order to determine the learning parameters that provided the best performance. To this end, an overview of the grid search values used to determine the best learning parameters for this project was presented. The performance of each of the values of the grid search can be tested using v-fold cross-validation or a separate training and testing set. The former method is the most usual method of determining the best learning parameters; therefore, this process was elucidated (Bennett & Campbell, 2000; also Schölkopf & Smola, 2002:217:10). Moreover, theoretical concerns of the validity of v-fold cross-validation were raised and two alternatives to this technique were suggested, following Schölkopf and Smola (2002:217). These suggestions, however, could not be followed for this research project.

This chapter also raised and addressed three potential data problems that concern this research project. The first problem, that of imbalanced data sets, is relevant because the 'bad'/'good' essay ratio is 137/209, which could result in the classifier classifying any particular instance as 'good' simply because there are more examples of 'good'. This problem was solved, using a classification penalty on the 'good' examples. The second problem was that although both kinds of misclassifications are not equally severe, the SVM classifier treats the costs of both these misclassifications as equal. This problem could not be solved for the SVM implementation used in this research project. The third data problem was that SVMs assume the training and testing data to be drawn from the same distribution but that this distribution is not assumed to be normal. This led to a hypothesis, which is discussed in Section 5.3, regarding the normal distribution of the data.

The following evaluation measures were presented next: error rate, accuracy, accuracy baselines, improved performance, precision, recall, the average recall baseline, and the F-measure. In light of the purpose of the classification task of this project the most suitable accuracy measure was discussed. For this project, correct classification of the

non-prototypical examples is the most important, therefore, the recall of 'bad' examples was the measure used to evaluate the classifier's performance.

Furthermore, Chapter 4 reported on and drew comparisons between the results of the various classifiers trained and tested on different data and feature sets, using C-and ν-SV classification, and linear and RBF kernels. In this chapter, it was shown that C-SV classification and the linear kernel showed better results than ν-SV classification and the RBF kernel. For this reason, all further tests were conducted, using C-SV classification and the linear kernel. In this chapter, the various hypotheses that were raised in Chapters 3 and 4 were discussed. Finally, the best classifier's performance was analysed and seven possible reasons were suggested to explain the results of this classifier. The results of the classifiers, the hypotheses, and explanations for the classifier's performances are presented in Section 5.3, below. In this section, the research questions posed in Chapter 1 are fully answered.

5.3 Summary of results and findings

As mentioned above, Chapter 4 detailed the results of all the classifiers trained, using ν- and C-SV classification, linear and RBF kernels as well as various data, feature and case sets. Tests on the different data, feature, and case sets were motivated by hypotheses postulated in Chapters 3 and 4.

The first **hypothesis** was that normalising the features to text length would remove skewness from the frequencies and would thus yield better results than the raw frequency set. Contrary to this, the classifier trained on the text-length normalised data set yielded the same results as the raw data set, if feature selection was used on the former set. Without feature selection, the text-length normalised data set yielded poorer results than the raw data set. It is possible that text-length normalisation would not produce any perceptibly different classification results from the raw frequencies if the texts were of similar length. This was not the case for this project as the essays' lengths ranged from approximately 800 to 4000 words. It would thus appear that, although normalisation for text length is linguistically and statistically motivated, raw frequencies could be used without normalisation.

The second **hypothesis** was drawn directly from the data distribution assumption of SVMs, as mentioned in Section 5.2 above. As discussed in Section 3.6, the log-transformed data set was transformed, using logarithmic transformation, in order to transform non-normal data to approximate a normal distribution. It was expected that, because the data set was not assumed to follow a normal distribution, normalisation of the data would have no effect on the classifier's results (on 'good' and 'bad' examples). The results on the text-length normalised data set and the log-transformed data set showed approximately the same AR values, which seems to imply that normalisation had not affected performance. Nevertheless, the RB values and RG values of the classifiers differed. As a result, the second hypothesis could not be accepted or rejected because this experiment could not provide conclusive evidence of the normality of the distribution not affecting the results of SVM learning.

The third **hypothesis** was that a classifier trained on a tagset, which extracts finer linguistic detail (the CLAWS7 tagset) than the Penn Treebank tagset, would perform more accurately than one built on the Penn Treebank tagset. Comparison of the RB values of the classifier trained on the CLAWS7 tagset and the classifier trained on the Penn Tree tagset showed that the classifier trained on the CLAWS7 tagset outperformed the classifier trained on the Penn Tree tagset. Conversely, comparison of the RG values of the classifier trained on the Penn Treebank tagset and the classifier trained on the CLAWS7 tagset showed that the classifier trained on the Penn Treebank tagset outperformed the classifier trained on the CLAWS7 tagset. Hence, it could be deduced that the CLAWS7 tagset was more discriminant of non-prototypical argumentative essays and the Penn Treebank tagset more discriminant of prototypical argumentative essays. As with the second hypothesis, the third hypothesis was intended to be interpreted in terms of better classification of both classes. Therefore, the third hypothesis could only be partially accepted.

The fourth **hypothesis** put forth was that middle-grade essays are linguistically harder to distinguish, and therefore, the difference in linguistic features used in this project were not sufficiently discriminant to determine whether texts were instances of prototypical or non-prototypical argumentative essays. Therefore, it was postulated that the removal of middle-grade essays would improve the classifier's performance. The results of the

RB values of the classifier demonstrated that the classifier trained and tested on the reduced cases was not outperformed in RB by the classifier trained and tested on the full case set. It would thus appear that the hypothesis should be accepted. However, the experiment could not be considered conclusive without further evidence because the results on RG showed that the classifier trained and tested on the full case set outperformed the classifier trained and tested on the reduced cases, and because the former classifier was being compared with a classifier trained and tested on more observations.

Training sample size, first raised in Chapter 2, was a concern because a training sample should be linguistically representative of the target population and the classifier should have equal experience of both classes. As a result, the larger the data set the more likely accurate classification would be. The effect of sample size was examined by training a classifier, using the text-length normalised data set, on the complete feature set and a linear kernel, but with 80 percent of the case set. This classifier could be reasonably compared to the classifier trained and tested on the upper- and lower-band grade instances as it was trained and tested on 76 percent of the case set. From the RB values it could be seen that the classifier trained and tested on 76 percent of the case set showed a better performance than the classifier trained and tested on 80 percent of the case set. Conversely, from the RG values, it could be seen that the classifier trained and tested on 76 percent of the case set showed a poorer performance than the classifier trained and tested on 80 percent of the case set. This did not suffice to provide evidence that the performance of the middle-grade case set classifier was not due to data set size.

The size of the two classes that each classifier had to train and test on were examined. The 80 percent case set contained 125 training and 45 testing instances of 'good' examples, and 94 training and 29 testing instances of 'bad' examples. The middle-grade case set had 96 training and 34 testing instances of 'good' examples, and 97 training and 29 testing instances of 'bad' examples. Thus, it was suggested that the results of the two classifiers trained, using the different case sets could be explained in terms of the number of examples of each class that they were trained and tested on.

Evidence against the fourth hypothesis was given by the particular types of errors made on the training and testing on the full case set. According to this hypothesis, fifteen out of the thirty-two errors on the testing set were not expected, that is, these misclassified essays fell unequivocally into the top band of prototypical essays and into the lower band of the non-prototypical essays. Equally, forty out of the ninety-six errors made on the training set were not expected. As 46.9 percent of errors on the test set and 41.7 percent of errors on the training set were not expected, this provided evidence for rejecting the fourth hypothesis.

In addition to these hypotheses, feature selection was also explored. In Chapter 3, it was stated that as SVMs can learn well in high dimensions, feature selection does not appear to be essential to this type of learning. This was not shown to be false here, but rather, evidence in favour of careful feature selection was presented because of the good results obtained, using feature selection. It appeared to be the case that for SVM learning it was not too many features, but rather too many irrelevant features that could produce poorer results. As previously mentioned, it was difficult to determine *a priori* which features were irrelevant. A good case, however, was made in this research project for exploratory techniques, such as cluster analysis and decision trees (as used by Wolters & Kirsten, 1999), before attempting final classification tests.

The good results on the raw data set equalled the best results obtained on the C-SV classifier trained on the text-length normalised data set, using feature selection, a linear kernel and with the parameter $C = 1.953125$, in terms of RB and RG on the testing data set. The major difference between these two classifiers was their training results. As mentioned above, training on the raw data set with the complete feature set resulted in a 100 percent training accuracy. This contrasted greatly with the results of training on the text-length normalised data set with feature selection. The latter classifier resulted in a training accuracy of only 62.9 percent, which is 3.4 percent above the baseline but 37.1 percent below the best training accuracy achieved, indicating substantially poorer overall performance. This provided further evidence for using the raw data set with no feature selection.

In Chapter 4, it was reported that although the testing accuracy was much improved by using feature selection, training accuracy was reduced (this was judged in comparison with the results on the raw data set, and could also be seen in the results of the classifier trained on the full feature set, using the text-length normalised data set). The training result showed that the removal of the rest of the feature set was detrimental to training, as the classifier trained on the reduced feature set was shown to underfit the training data. This provided evidence in favour of the feature set as a whole, that was selected for this classification task, as it appeared that these features were all valuable to some extent in discriminating between the essays. Thus, this research project provided evidence in support of Dewdney *et al.* (2001:7) and Finn (2002:76) who find that both BOW and linguistic features can be used well together. Moreover, this project confirmed that the BOW approach, if used with more careful word selection could be useful and did not need be viewed as opposing a linguistic feature set (this was not presented as such but is evident in Argamon & Dodick, 2004a; and Santini, 2005b).

As feature selection and particular feature selection techniques are not guaranteed to select only those features that are relevant to the classification task the selected features must be interpreted with caution and inferences made tentatively. It was pointed out in Chapter 3 that the feature selection was conducted to test classification performance, not to make any inferences on any particular feature selection, with final implications for this study. The thirty-eight features that were selected were thus reviewed and their significance discussed but these did not indicate the absolute usefulness of the remaining features as it had already been shown that the majority of the features had a useful discriminating function. These features were mainly function words or indicative of function words (POS tags). Three features were kinds of subordinating conjunctions and reporting verbs that are words linked to academic writing. Other features were non-integrated quotations, sentence counts for non-integrated quotations, the ARI readability score, and the number of references. The first two features are unique to this research project, thus, it appears promising that these features were selected as good indicators of prototypical and non-prototypical essays. This could then indicate that they can be considered useful for future automatic genre classification projects. The selection of the average number of references is pleasing because argumentative essays are required to

provide evidence for their argumentation, which would entail references to other research. Therefore, this feature is expected to be characteristic of argumentative essays.

Contrary to what was expected, words characteristic of academic writing, such as hedges, downtoners, and conjuncts were not selected during feature selection. As indicated above, this could not be used to infer that these words are not useful features for this classification task. These features may rather be less (by varying margins) discriminatory indicators than the features selected. Also, it is also possible that these features would have been selected if other feature selection techniques had been used.

Thus, from the above discussion it can been seen that the results of the various classifiers as well as the testing of the hypotheses, postulated in Chapters 3 and 4, have fully addressed the first, second and fourth research questions.

As reported in Chapter 4, the best classifier's performance was 62.5 percent RB, 74.5 percent RG and 68.5 percent AR on a 75 percent training set and 25 percent testing set. This C-SV classifier ($C = 32$) was trained on the text-length normalised data set, using feature selection and a linear kernel. In practice, this means that out of 87 essays classified, 34 essays would be given to the senior marker and 53 essays to the junior marker. Of the essays given to the senior marker, 14 would be incorrectly classified, while 12 essays given to the junior marker would be incorrectly classified.

From the discussion in Chapter 4, as well as in light of the difficulty of the classification task it is apparent that these results are very good. They are judged so in terms of both RB, and the average and testing baselines. Nevertheless, it is still desirable to achieve 100 percent recall on 'bad' and 'good' examples.

As mentioned in Section 5.2, Chapter 4 analysed the best classifier's performance and provided seven potential explanations to account for the less than 100 percent recall performance of the classifier. The first reason was that argumentative essays were not truly representative of non-prototypical examples. This reason was postulated as the most likely explanation for the classifier not having been able to identify non-prototypical essays as easily as prototypical essays. The data used in this project are all

relatively prototypical to very prototypical argumentative essays, and considered indicative of proficient student writing by the BAWE corpus compilers. The lowest grade for the essays in the corpus is 60 percent. The classifier ought to still generalise well if provided true instances of prototypical and non-prototypical argumentative essays because it is very good at identifying prototypical argumentative essays. On the data set used in this project it would be more likely that the classifier would be able to distinguish better between the two classes if the 'bad' class was truly non-prototypical, that is, if essays of fail marks made up the non-prototypical case set.

This problem could not be solved for this research project, as very few L1 English argumentative student essays are available for study. Moreover, where such corpora exist, they do not always have the necessary features for training; for example the LOCNESS (Granger, 1994) has had quotations removed from the texts, from which five important features were derived for this research project.

The second reason was that the size of the data set may not have been adequate. It is possible that the classifier might have required more learning examples, in order to generalise well and thus have optimal performance. The best way of establishing whether accuracy is a function of training size is to train the classifier on an increasing number of texts, and then test the classifier on the same testing set each time. This was not possible with the SVM package used for this project because the training and testing examples were selected simultaneously (and randomly) by the algorithm. This meant that the testing set could not be independently selected.

The third reason was that cross-domain classification could have reduced accuracy. Evidence for this was provided in Finn (2002:86), Wolters and Kirsten (1999), and Karlgren and Cutting (1994:1072). Thus, it was expected that for this research project, in which all the data were drawn from and classified across domains, accuracy would be negatively affected.

The fourth explanation is that personal language differences could have confused classification. Van Halteren, Baayen, Tweedie, Haverkort and Neijt (2005) prove that a human stylome exists in the framework of authorship attribution studies, and find that

personal language differences play a great role in classification of these studies, even in the case of non-specialist writers (2005:14). Since the essays used in this research project were written by different authors, it is likely that individual stylistic variations, as established by Van Halteren *et al.* (2005) could affect classification. Thus, differences that are not only indicative of prototypical or non-prototypical essays but also of personal language styles were also extracted, resulting in a poorer performance in classifying prototypical and non-prototypical essays.

The fifth explanation was the most obvious and simple explanation that the features used could not have adequately extracted the linguistic differences between the classes. As the identification of good discriminant features is one of the main concerns of automatic genre classification, solutions to choosing better features are important for future work. Therefore, suggestions in this regard are made in the next section.

The sixth explanation was that the feature selection techniques used may not have been adequate to extract the most relevant features from the feature set. This topic is addressed in the form of potential solutions to better feature selection in Section 5.4, below.

The seventh explanation was that more knowledge of the discourse 'struggle' could have needed to be considered. This explanation is derived from the approach this project took towards student academic writing, as stated in Chapter 1. Nystrand (1987:204) claims that any good text analysis needs to distinguish between "the structure of argument and the structure of communication". This project aimed to extract relevant linguistic features of argumentative essays. It is possible that, in addition to these features, there may be evidence of a communicative role with the reader-marker, which had not been extracted, and which could provide more accurate classification.

This section has thus shown that results of the best classifier as presented in Chapter 4, answered the fourth research question fully. This study can thus be shown to have achieved all of its aims, by developing a classifier, using an automatic genre classification approach, which will categorise prototypical and non-prototypical argumentative essays of student writers, into 'good' or 'bad' classes. In order to do this

the most discriminating features between prototypical argumentative essays and non-prototypical argumentative essays were determined in an initial literature review and experiments on classifier results, using different feature sets. Furthermore, only linguistic features that can be easily computed were used. The classifier that was developed was trained on SVM learning, a technique that has been shown to provide good generalisability in automatic genre classification studies (Dewdney *et al.*, 2001; Argamon & Dodick, 2004a; Meyer zu Eissen & Stein, 2004; and Santini, 2005b) and in other classification tasks (Burges, 1998:121, Schölkopf & Smola, 2002:22; Kroon, 2003:55; and Stecking & Schebesch, 2003). This machine learning technique requires human effort only in determining the optimal learning parameters, as discussed in Chapter 4. The next section discusses recommendations for further research in light of the above discussion.

5.4 Recommendations for further research

The recommendations in this section are for furthering a study of the nature of this project. Therefore, the recommendations could be applied to automatic genre classification studies seeking to classify academic texts and to similar NLP applications where the object is to explore the linguistic features of academic writing. These recommendations are derived from this research project as well as other studies. They are grouped into three areas: features, evaluation, and learning technique, and are listed separately below.

5.4.1 Features

In Section 5.3, it has been suggested that it is possible that the linguistic features of this project did not adequately extract the differences between the two classes under examination. A variety of features could potentially improve accuracy for a further study of a similar nature to this research project. However, only a few features are suggested here: additional quotation information, word length, readability scores, feature ratios, POS trigrams, parsing, and stability.

The first four features are surface features, which are the most used type of features because they are easy to compute and have been demonstrated to achieve good accuracy

in the field of automatic genre classification (this study, and also see the studies reviewed in Chapter 2). The first feature is additional quotation information. This study considered quotations and the verbs associated with them (see reporting verbs in Section 3.3.6.4) as potentially revealing features for distinguishing prototypical and non-prototypical examples of argumentative essays. Quotations could be further explored by examining their location in sentences, and the types of reporting verbs could be grouped according to the type of reporting: direct quotation, paraphrase, summary, and generalisation (Hyland, 2002:116).

The second feature is word length in syllables. Kelih *et al.* (2005), reviewed in Chapter 2, provide compelling evidence for word length being a powerful discriminator of genres, if defined as syllables per word.

The third feature is readability measures. In Chapter 4, it is shown that the ARI readability score was a good discriminant feature for classifying prototypical and non-prototypical essays. This encourages further exploration of readability measures, such as those in Boese (2005).

The fourth feature is ratios of features, which can be of many types. Two examples are the ratio of determiners to nouns and the ratio of prepositions to nouns. These kinds of ratios are used in Ross and Hunter's (1994) stylistic description tool.

The final three features go beyond more typical surface features. The first suggested feature set is POS trigrams that Santini (2004a) used, as indicators of deeper syntactic features. As was seen in this study, reviewed in Chapter 2, good accuracies resulted from POS trigrams. The second feature set is the functional cues and syntactic patterns used by Santini (2005b), reviewed in Chapter 2, which are extracted with a parser. These two features sets are expected to reveal more information regarding a text's syntax than, for example, simple POS tags.

The third feature is the measure on linguistic features introduced by Koppel, Argamon, and Shimoni (2003), which they term *stability*. Features are considered stable if they have semantic equivalents, which can replace them without changing the meaning of the

text. They show that features with low stability can be useful for stylistic text classification. This feature is expected to yield potentially useful information with regard to a text's semantic structure. Up to the present, automatic genre classification has not made use of semantically-orientated linguistic features. Semantic analysis has already been proven successful in essay grading (Lemaire & Dessu, 2001).

In addition to experimenting with the features detailed above, features could be collapsed into smaller categories. This could be done on the basis of analysis by hand and then grouping according to function. Alternatively, features could be combined, using, for example, Principal Components Analysis (Hand, 1997:149–151).

5.4.2 Evaluation

A variety of standard evaluation measures were detailed in Chapter 4. According to Hand (1997:100–114), the following four evaluation measures provide a more comprehensive method of evaluation (these measures require access to the probabilities of belonging to the classes being examined):

1. Inaccuracy: judges how ineffective a classification rule is in classifying an object correctly. This is based on a measure of the difference between the true class and the estimated probability of belonging to this class.

2. Imprecision: the difference between estimated probabilities and true probabilities of belonging to a particular class.

3. Inseparability: this determines whether the classification situation is indeed completely separable, by establishing whether the true probabilities of belonging to each class at $\mathbf{x}$, averaged over $\mathbf{x}$, are similar. Similarity will indicate that the problem is inseparable.

4. Resemblance: measures the variation between the true probabilities of belonging to one class above the decision surface and to another class below the decision surface.

Thus, a good classifier is one with high accuracy, high precision, and low resemblance. This means low inaccuracy coefficients, imprecision coefficients and resemblance coefficients (Hand, 1997:100).

5.4.3 Learning technique

Recommendations by way of the learning technique used, refer to either using a different implementation of SVMs or combining classifiers.[101] Thus, different classifiers can be trained, using the same learning methodology but, for example, training the classifiers, using different feature sets. The final classifications can be done, using bagging, where the chosen classification is the one most often selected by all the classifiers, or boosting, where all the classifiers contribute to the final vote by using weighted voting (Hand, 1997:159–162). An example of such an approach is the ensemble learning approach of the follow-up study to Finn (2002), which was reviewed in Chapter 2. In this study, Finn (2002:76) suggests combining feature sets in classification. This suggestion is taken up in Finn and Kushmerick (2003) in which they compare accuracies of Finn (2002) with a classifier built, using the combined feature sets. They make use of an ensemble learner, which is the result of three classifiers trained on the same data, each using a different feature set. They find that their ensemble learner often performs significantly better in comparison with the performance of individual feature sets.

5.5 Summary

This chapter provided a conclusion to this dissertation. It summarised the preceding chapters, presented the results and findings of Chapter 4, and reviewed the hypotheses postulated in Chapters 3 and 4. This was done to show how the research questions posed in Chapter 1 were adequately addressed by the research project. Thus, this chapter demonstrated that the research aims of this study have been achieved by developing an automatic genre classifier that categorises prototypical and non-prototypical instances of the argumentative essay genre, written by students.

[101] For a comprehensive list of SVM implementations see http://www.support-vector-machines.org/SVMsoft.html.

This was achieved by first reviewing automatic genre classification studies, in order to determine the potentially most discriminating features of prototypical and non-prototypical argumentative essays. Thereafter, the features used for this research project were selected from these studies, as well as two well-known English grammar books: Biber *et al.* (1999) and Quirk *et al.* (1985). In Chapters 3 and 4, various hypotheses regarding best feature sets, case sets and data transformations were posed. These were examined and then accepted or rejected. This acceptance or rejection, summarised in this chapter, holds relevant implications regarding the most discriminant linguistic features for prototypical and non-prototypical argumentative essays. Finally, the fourth research aim of this study was achieved by developing the classifier, using SVM learning.

Finally, this chapter concluded by putting forth suggestions for future research regarding features, evaluation, and learning technique.

REFERENCE LIST

ABE, S. 2005. Support vector machines for pattern classification. London: Springer. 343 p.

AIRES, R.; ALUÍSIO, S.; & SANTOS, D. 2005. User-aware page classification in a search engine. (*In* the Proceedings of Stylistic Analysis of Text for Information Access, SIGIR Workshop, Salvador, Brazil, August 19). www.linguateca.pt/documentos/AiresetalSIGIR2005.pdf Date of access: 6 Jan. 2008.

AIZERMAN, A.; BRAVERMAN, E.M; & ROZONER, L.I. 1964. Theoretical foundations of the potential function method in pattern recognition learning. Automation and remote control, 25:821–837.

ALUÍSIO, S.M.; PINHEIRO, G.M.; FINGER, M.; NUNES, M.G.V.; & TAGNIN, S.E. 2003. The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. (*In* Archer, D.; Rayson, P.; Wilson, A.; & McEnery, T., *eds.* The Proceedings of Corpus Linguistics, Lancaster, U.K. UCREL Technical Papers, Vol. 16, p. 14–21.) ftp://ftp.ime.usp.br/pub/mfinger/2003/Lacio.pdf.gz Date of access: 6 Jan. 2008.

ARGAMON, S. & DODICK, J. 2004a. Conjunction and modal assessment in genre classification: a corpus-based study of historical and experimental science writing. (*In* Qu, Y.; Shanahan, J.; & Wiebe, J., *eds.* The Proceedings of the AAAI Spring Symposium on Attitude and Affect in Text: Theories and Applications, Palo Alto, California, U.S.A., March. Menlo Park, CA: AAAI Press. p. 1–8.)

ARGAMON, S. & DODICK, J. 2004b. Linking rhetoric and methodology in formal scientific writing. (*In* Forbus, K.; Gentner, D.; & Regier, T., *eds.* The Proceedings of the 26th Annual Conference of the Cognitive Science Society, Chicago, Illinois, U.S.A., August. p. 61–65.) http://www.cogsci.rpi.edu/CSJarchive/Proceedings/2004/CogSci04.pdf Date of access: 6 Jan. 2008.

ARGAMON, S.; KOPPEL, M.; & AVNERI, G. 1998. Routing documents according to style. (*In* Schwartz, D.; Divitini, M. & Brasethvik, T., *eds.* The Proceedings of IIIS-98: 1st International Workshop on Innovative Internet Information Systems, Pisa, Italy, June 8–9.) http://citeseer.ist.psu.edu/cache/papers/cs/2891/http:zSzzSzwww.idt.ntnu.nozSz~monicazSziis-98zSzpaperszSzargamon.pdf/argamon98routing.pdf Date of access: 6 Jan. 2008.

ASTON, G. & BURNARD, L. 1998. The BNC handbook: exploring the British National Corpus with SARA. Edinburgh: Edinburgh University Press. 256 p.

BELLMAN, R.E. 1961. Adaptive control processes: a guided tour. Princeton, NJ: Princeton University Press. 255 p.

BENNETT, K.P. & CAMPBELL, C. 2000. Support vector machines: hype or hallelujah? SIGKDD Explorations, 2(2):1–13, Dec. http://www.acm.org/sigs/sigkdd/explorations/issue2-2/bennett.pdf Date of access: 6 Jan. 2008.

BEUTEL, B. 1998. Malaga User Manual. http://www.linguistik.unierlangen.de/Malaga.de.html Date of access: 6 Jan. 2008.

BIBER, D. 1986. Spoken and written textual dimensions in English: resolving the contradictory findings. Language, 62(2):384–414, Jun.

BIBER, D. 1988. Variation across speech and writing. London: Cambridge University Press. 299 p.

BIBER, D. 1989. A typology of English texts. Linguistics: an interdisciplinary journal of the language sciences, 27(1):3–34.

BIBER, D. 1993a. Using register-diversified corpora for general language studies. Computational linguistics, 19(2):219–242, Jun.

BIBER, D. 1993b. Representativeness in corpus design. Literary and linguistic computing, 8(4):243–257.

BIBER, D.; CONRAD, S.; & REPPEN, R. 2000. Corpus linguistics: investigating language structure and use. Cambridge: Cambridge University Press. 300 p.

BIBER, D.; CONRAD, S.; REPPEN, R.; BYRD, P.; HELT, M.; CLARK, V.; CORTES, V.; CSOMAY, E.; & URZUA, A. 2004. Representing language use in the university: analysis of the TOEFL 2000 Spoken and Written Academic Language corpus. Report: ETS TOEFL Monograph series MS-25, January, RM-04-03. Princeton, NJ: Educational Testing Service. 303 p. http://www.ets.org/Media/Research/pdf/RM-04-03.pdf Date of access: 6 Jan. 2008.

BIBER, D.; JOHANSSON, S.; LEECH, G.; CONRAD, S.; & FINEGAN, E. 1999. The Longman grammar of spoken and written English. Pearson: Essex. 1204 p.

BISANT, D. 2005. An application of neural networks to sequence analysis and genre identification. International journal of pattern recognition and artificial intelligence, 19(2):199–215, Mar.

BLANKENSHIP, J. 1962. A linguistic analysis of oral and written style. Quarterly journal of speech, 48(4):419–422, Dec.

BOESE, E.S. 2005. Stereotyping the web: genre classification of web documents. Fort Collins: Colorado State University. (Thesis — M.Sc.) 68 p. http://www.cs.colostate.edu/~boese/Research/masters.pdf Date of access: 6 Jan. 2008.

BOESE, E.S. & HOWE, A.E. 2005. Effects of web document evolution on genre classification. (*In* Herzog, O.; Schek, H.; Fuhr, N.; Chowdhury, A.; & Teiken, W., *eds*. The Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM), Bremen, Germany, October 31 – November 5. New York, NY: ACM Press. p. 632–639.) http://www.cs.colostate.edu/~boese/Research/Publications/CIKM05-boese.pdf Date of access: 6 Jan. 2008.

BOSER, B.E.; GUYON, I.M.; & VAPNIK, V.N. 1992. A training algorithm for optimal margin classifiers. (*In* D. Haussler, *ed*. The Proceedings of the 5th Annual Workshop on Computational Learning Theory, Pittsburgh, Pennsylvania, U.S.A. New York, NY: ACM Press. p. 144–152.) http://coblitz.codeen.org:3125/citeseer.ist.psu.edu/cache/papers/cs/11260/http:zSzzSzwww.clopinet.comzSzisabellezSzPaperszSzcolt92.pdf/boser92training.pdf Date of access: 6 Jan. 2008.

BRANTS, T. 2000. TnT: a statistical part-of-speech tagger. (*In* the Proceedings of the 6th Conference on Applied Natural Language Processing, Seattle, Washington, U.S.A., April 29 – May 3. San Francisco, CA: Morgan Kaufmann. p. 224–231). http://www.aclweb.org/anthology-new/A/A00/A00-1031.pdf Date of access: 6 Jan. 2008.

BRASLAVSKI, P. & TSELISCHEV, A. 2005. Experiment on style-dependent document ranking. (*In* the Proceedings of the 7th Russian Conference on Digital Libraries (RCDL'05), Yaroslavl, Russia, October 4–6). http://www.rcdl2005.uniyar.ac.ru/ru/RCDL2005/papers/sek7_1_paper.pdf Date of access: 6 Jan. 2008.

BRILL, E. 1992. A simple rule-based tagger. (In the Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP'92), Trento, Italy, March 31 – April 3. Morristown, NJ: Association for Computational Linguistics. p. 152–155). http://www.aclweb.org/anthology-new/A/A92/A92-1021.pdf Date of access: 6 Jan. 2008.

BUCKINGHAM SHUM, S.; UREN, V.; LI, G.; DOMINGUE, J.; & MOTTA, E. 2002. Visualizing internetworked argumentation. (*In* Kirschner, P.A.; Buckingham Shum, S.J. & Carr, C.S., *eds*. Visualizing argumentation: software tools for collaborative and educational sense-making. London: Springer. 216 p.)

BURGES, C.J.C. 1998. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2):121–167, Jun.

BURSTEIN, J.; MARCU, D.; ANDREYEV, S.; & CHODOROW, M. 2001. Towards automatic classification of discourse elements in essays. (*In* the Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01), Toulouse, France, July 9–11. San Francisco, CA:Morgan Kaufmann. p. 98–105.) http://www.isi.edu/~marcu/papers/thesis-statements-acl01.pdf Date of access: 6 Jan. 2008.

CALANDO, A. 2004. Utrac-0.3.0. Alliance MCA. http://utrac.sourceforge.net/download/utrac-0.3.0.tar.gz Date of access: 6 Jan. 2008.

CHAFE, W. 1982. Integration and involvement in speaking, writing and oral literature. (*In* Tannen, D., *ed.* Spoken and written language: exploring orality and literacy. Norwood, N.J.: Ablex. p. 35–54.)

CHAFE, W & DANIELEWICZ, J. 1987. Properties of spoken and written language. (*In* Horowitz, R & Samuels, S.J., *eds.* Comprehending oral and written language. New York, NY: Academic Press. p. 83–113.)

CHANG, C.-C. & LIN, C.-J. 2001. Training ν-support vector classifiers: theory and algorithms. Neural computation, 13(9):2119–2147, Sept.

CHEN, P.-H.; LIN, C.-J.; & SCHÖLKOPF, B. 2005. A tutorial on ν-support vector machines. Applied stochastic models in business and industry 21(2), 111–136, March/April. http://www.kyb.mpg.de/publications/pdfs/pdf3353.pdf Date of access: 6 Jan. 2008.

COVER, T. M. 1965. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE transactions on electronic computers, 14(3):326–334, Jun.

CORTES, C. & VAPNIK, V.N. 1995. Support vector networks. Machine learning, 20(3):273–297, Sept.

COUTURE, B., *ed.* 1986. Functional approaches to writing: research perspectives. Norwood, NJ: Ablex. 271 p.

CRISTIANINI, N. & SHAWE-TAYLOR, J. 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press. 189 p.

CROWSTON, K. & KWASNIK, B. 2004. A framework for creating a facetted classification for genres: addressing issues of multidimensionality. (*In* the Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS'04), Track 4, Waikoloa, Island of Hawaii, January 5–8. Los Alamitos, CA: IEEE Computer Society. 9 p.) http://csdl2.computer.org/comp/proceedings/hicss/2004/2056/04/205640100a.pdf Date of access: 6 Jan. 2008.

DAELEMANS, W.; VAN DEN BOSCH, A.; & WEIJTERS, T. 1997. IGTtree: using trees for compression and classification in lazy learning algorithms. Artificial intelligence review, 11(1–5):407–423, Feb.

DAVIDSON, C. & TOMIC, A. 1999 (*In* Jones, C.; Turner, J.; & Street, B., *eds.* Students writing in the university: cultural and epistemological issues. Philadelphia: John Benjamins. p. 161–169.)

DEWDNEY, N.; VANESS-DYKEMA, C.; & MACMILLAN, R. 2001. The form is the substance: classification of genres in text. (*In* The Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management, Toulouse, France, July 6–7. Morristown, NJ: Association for Computational Linguistics. p.1–8.) http://www.aclweb.org/anthology-new/W/W01/W01-1007.pdf Date of access: 6 Jan. 2008.

EITRICH, T. & LANG, B. 2006. Efficient optimization of support vector machine learning parameters for unbalanced datasets. Journal of computational and applied mathematics, 196(2):425–436, Nov.

ENGLISH, F. 1999. What do students really say in their essays? Towards a descriptive framework for analysing student writing. (*In* Jones, C.; Turner, J.; & Street, B., *eds.* Students writing in the university: cultural and epistemological issues. Philadelphia: John Benjamins. p. 17–36.)

EMDE, W. & WETTSCHERECK, D. 1996. Relational instance based learning. (*In* Saitta, L., *ed.* The Proceedings of the 13th International Conference on Machine Learning (ICML'96), Bari, Italy, July 3–6. San Francisco, CA: Morgan Kaufmann p. 122–130).

ERIKSON, F. & SCHULTZ, J. 1982. The counsellor as gatekeeper: social interaction in interviews. New York, NY: Academic Press. 263 p.

EXPERT ADVISORY GROUP ON LANGUAGE ENGINEERING STANDARDS (EAGLES). 1996. Preliminary recommendations on text typology. EAGLES Document EAG-TCWG-TTYP/P, June. http://www.ilc.cnr.it/EAGLES/texttyp/texttyp.html Date of access: 6 Jan. 2008.

FINN, A. 2002. Machine learning for genre classification. Dublin: University College. (Thesis — M.Sc.) 133p. http://www.aidanf.net/publications/afthesis.ps Date of access: 6 Jan. 2008.

FINN, A. & KUSHMERICK, N. 2003. Learning to classify documents according to genre. (Paper presented at the 18th International Joint Conference on Artificial Intelligence (IJCAI-03) Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico, August 9–15.) (Unpublished.) http://www.aidanf.net/publications/finn03learninggenre.pdf Date of access: 6 Jan. 2008.

FINN, A., KUSHMERICK, N., & SMYTH, B. 2002. Genre classification and domain transfer for information filtering. (*In* Crestani, F.; Girolami, M.; & Van Rijsbergen, C.J., *eds.* Advances in information retrieval. The Proceedings of the 24th BCS-IRSG European Colloquium on IR Research Glasgow, U.K., March 25–27. Berlin: Springer. p. 353–362.)

FLESCH, R. 1948. A new readability yardstick. Journal of applied psychology, 32(3): 221–233, Jun.

FLOWERDEW, L. 2002. Corpus-based analyses in EAP. (*In* Flowerdew, J., *ed.* Academic discourse. Harlow: Pearson. p. 95–114.)

FORSBOM, E. 2005. Feature extraction for genre classification. http://stp.lingfil.uu.se/~evafo/gslt/statmet/statmet05forsbom.pdf Date of access: 6 Jan. 2008.

FRANCIS, W.N. & KUČERA, H. 1982. Frequency analysis of English usage: lexicon and grammar. Boston, MA: Houghton Mifflin. 561 p.

FRIEDL, J. 1997. Mastering regular expressions: powerful techniques for Perl and other tools. Sebastopol, CA: O'Reilly. 342 p.

GARSIDE, R. 1987. The CLAWS word-tagging system. (*In* Garside, R.; Leech, G.; & Sampson, G., *eds.* The computational analysis of English: a corpus-based approach. London: Longman. p. 30–41.)

GARSIDE, R. & SMITH, N. 1997. A hybrid grammatical tagger: CLAWS4. (*In* Garside, R.; Leech, G.; & McEnery, A., *eds.* Corpus annotation: linguistic information from computer text corpora. London: Longman. p. 102–121.)

GIMÉNEZ, J. (jgimenez@lsi.upc.edu) 12 Sept. 2006. Re: SVM Tool. E-mail to: Raaff, S. (20226209@puknet.puk.ac.za).

GIMÉNEZ, J. (jgimenez@lsi.upc.edu) 23 Jan. 2007. Character encoding help. E-mail to: Raaff, S. (20226209@student.nwu.ac.za).

GIMÉNEZ, J. & MÀRQUEZ, L. 2004. SVMTool: a general POS tagger generator based on support vector machines. (*In* Lino, M.T.; Xavier, M.F.; Ferreira, F.; Costa, R.; & Silva, R., *eds.* The Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal, May 26–28. Paris: ELDA. p. 43–46). http://www.lsi.upc.edu/~nlp/SVMTool/lrec2004-gm.pdf Date of access: 6 Jan. 2008.

GIMÉNEZ, J. & MÀRQUEZ, L. 2006. SVMTool: Technical Manual v1.3, August. http://www.lsi.upc.edu/~nlp/SVMTool/SVMTool.v1.3.pdf Date of access: 6 Jan. 2008.

GRANGER, S. 1994. The learner corpus: a revolution in applied linguistics. English today, 39(3): 25–29.

GRANGER, S. & RAYSON, P. 1998. Automatic profiling of learner texts. (*In* Granger, S., *ed.* Learner English on computer. London: Longman. p. 119–131.)

HALLIDAY, M.A.K. 1978. Language as social semiotic: the social interpretation of language and meaning. London: Arnold. 256 p.

HALLIDAY, M.A.K. 1987. Spoken and written modes of meaning. (*In* Horowitz, R & Samuels, S.J., *eds.* Comprehending oral and written language. New York, NY: Academic press. p. 55–82.)

HALLIDAY, M.A.K. & MATTHIESSEN, C.M.I.M. 2004. An introduction to functional grammar. 3rd ed. London: Arnold. 689 p.

HAND, D. J. 1997. Construction and assessment of classification rules. Chichester: Wiley. 214 p.

HASTIE, T; TIBSHIRANI, R; & FRIEDMAN, J. 2001. The elements of statistical learning: data mining, inference and prediction. New York, NY: Springer. 533 p.

HECHTER, T. 2004. A comparison of support vector machines and traditional techniques for statistical regression and classification. Stellenbosch: Universiteit Stellenbosch. (Dissertation — M.Com.) 159 p.

HENNING, J.G. 2006. Linking adverbials in first, second and foreign language English student writing corpora. Potchefstroom: Noordwes-Universiteit. (Dissertation — M.A.) 161 p.

HOLMES, J. (jasper.holmes@gmail.com) 21 Feb. 2007a. Re: Help with BAWE missing data. E-mail to: Raaff, S. (20226209@student.nwu.ac.za).

HOLMES, J. (jasper.holmes@gmail.com) 21 Feb. 2007b. Re: Help with BAWE missing data. E-mail to: Raaff, S. (20226209@student.nwu.ac.za).

HSU, C.-W.; CHANG, C.-C.; & LIN, C.-J. 2003. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University (Taiwan). http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf Date of access: 6 Jan. 2008.

HYLAND, K. 1998. Hedging in scientific research articles. Amsterdam: John Benjamins. 307 p.

HYLAND, K. 2002a. What do they mean? Questions in academic writing. Text, 22(4): 529–557. Available: Academic Search Premier. Date of access: 6 Jan. 2008.

HYLAND, K. 2002b. Activity and evaluation: reporting practices in academic writing. (*In* Flowerdew, J., *ed.* Academic discourse. Harlow: Pearson. p. 115–130).

HYLAND, K. 2004a. Patterns of engagement: dialogic features and L2 undergraduate writing. (*In* Ravelli, L.J. & Ellis, R.A., *eds.* Analysing academic writing: contextualised frameworks. London: Continuum. p. 5–23.)

HYLAND, K. 2004b. Genre and second language writing. Ann Arbor, MI: University of Michigan Press. 244 p.

HYLAND, K. & MILTON, J. 1997. Qualifications and certainty in L1 and L2 students' writing. Journal of second language writing, 6(2):183–205, May. Available: ScienceDirect. Date of access: 6 Jan. 2008.

IVANIČ, R.; CLARK, R.; & RIMMERSHAW, R. 2000. What am I supposed to make of this? The messages conveyed to students by tutors' written comments. (*In* Lea, M.R & Stierer, B, *eds*. Student writing in higher education. Buckingham: SRHE and Open University Press. p. 47–65.)

JACKSON, P. & MOULINIER, I. 2002. Natural language processing for online applications: text retrieval, extraction and categorization. Amsterdam: John Benjamins. 225 p.

JOACHIMS, T. 1999a. Making large-scale SVM learning practical. (*In* Schölkopf, B.; Burges, C.; & Smola, A., *eds*. Advances in kernel methods — support vector learning. Cambridge, MA: MIT. p. 169–184.)

JOACHIMS, T. 1999b. Transductive inference for text classification using support vector machines. (*In* Bratko, I. & Dzeroski, S., *eds*. The Proceedings of the 16th International Conference on Machine Learning (ICML'99), Bled, Slovenia, June 27–30. San Francisco, CA: Morgan Kaufmann. p. 200–209.) http://coblitz.codeen.org:3125/citeseer.ist.psu.edu/cache/papers/cs/6683/http:zSzzSzwww-ai.cs.uni-dortmund.dezSzDOKUMENTEzSzJoachims_99c.pdf/joachims99transductive.pdf Date of access: 6 Jan. 2008.

JOHANNESSON, E. & WALLSTRÖM, C. 1999. Automatic analysis and visualization of stylistic genres. (*In* Käkölä,T.K., *ed*. The Proceedings of the 22nd Information Systems Research Seminar in Scandinavia (IRIS 22): Enterprise Architectures for Virtual Organisations, Keuruu, Finland, August 7–10. Jyväskylä: Department of Computer Science and Information Systems, University of Jyväskylä. p. 149–162.) http://www.cs.jyu.fi/~timokk/iris22/IRIS22_Volume2.pdf Date of access: 6 Jan. 2008.

JOHANSSON, S., LEECH, G., & GOODLUCK, H. 1978. Manual of information to accompany the Lancaster-Oslo-Bergen corpus of British English, for use with digital computers. Oslo: Department English, Oslo University. 147 p.

JURAFSKY, D. & MARTIN, J.H. 2000. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, NJ: Prentice Hall. 934 p.

KARAKOULAS, G. & SHAWE-TAYLOR, J. 1999. Optimizing classifiers for imbalanced training sets. (*In* Kearns, M. S.; Solla, S. A.; & Cohn, D. A., *eds*. Advances in neural information processing systems, Vol. 11. Cambridge, MA: MIT. p. 253–259.)

KARJALAINEN, A.; PÄIVÄRINTA, T.; TYRVÄINEN, P.; & RAJALA, J. 2000. Genre-based metadata for enterprise document management. (*In* the Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS'33), Vol. 3, Maui, Hawaii, January 4–7. Los Alamitos, CA: IEEE Computer Society. 10 p.) http://csdl2.computer.org/comp/proceedings/hicss/2000/0493/03/04933013.pdf Date of access: 6 Jan. 2008.

KARLGREN, J. 1999. Stylistic experiments in information retrieval. (*In* Strzalkowski, T., *ed.* Natural language information retrieval. Dordrecht: Kluwer. p. 147–166.)

KARLGREN, J. 2000. Stylistic experiments for information retrieval. Stockholm: Universitet Stockholms. (Dissertation — D.Phil.) 147 p. http://www.sics.se/~jussi/Artiklar/2000_PhD/thesis.ps.gz Date of access: 6 Jan. 2008.

KARLGREN, J. & CUTTING, D. 1994. Recognizing text genres with simple metrics using discriminant analysis. (*In* the Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994), Vol. 2, Kyoto, Japan, August 5–9. Morristown, NJ: Association for Computational Linguistics. p. 1071–1075.) http://eprints.sics.se/56/01/cmplglixcol.pdf Date of access: 6 Jan. 2008.

KELIH, E.; ANTIĆ, G.; GRZYBEK, P.; & STADLOBER, E. 2005. Classification of author and/or genre? The impact of word length. (*In* Weihs, C. & Gaul, W., *eds.* Classification: the ubiquitous challenge. Berlin: Springer. p. 498–505.)

KESSLER, B.; NUNBERG, G.; & SCHÜTZE, H. 1997. Automatic detection of text genre. (*In* the Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'97), Madrid, Spain, July 7–12. Morristown, NJ: Association for Computational Linguistics. p. 32–38.) http://www.aclweb.org/anthology-new/P/P97/P97-1005.pdf Date of access: 6 Jan. 2008.

KOHONEN, T.; KANGAS, J.; LAAKSONEN, J.; & TORKKOLA, K. 1995. LVQ PAK: the learning vector quantization package v. 3.1. Technical Report A30, Helsinki University of Technology (Norway). http://citeseer.ist.psu.edu/cache/papers/cs/5716/http:zSzzSzwww.ele.etsmtl.cazSzacademiquezSzele778zSzarticleszSzlvq_tr96.pdf/kohonen96lvq.pdf Date of access: 6 Jan. 2008.

KOPPEL, M.; AKIVA, N.; & DAGAN, I. 2003. A corpus-independent feature set for style-based text categorization. (Paper resented at the 18th International Joint Conference on Artificial Intelligence (IJCAI-03) Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico, August 9–15.) (Unpublished.) http://citeseer.ist.psu.edu/cache/papers/cs2/86/http:zSzzSzwww.cs.biu.ac.ilzSz~koppelzSzpaperszSzstability-workshop.pdf/koppel03corpusindependent.pdf Date of access: 6 Jan. 2008.

KOPPEL, M.; ARGAMON, S.; & SHIMONI, A. 2003. Automatically categorizing written texts by author gender. Literary and linguistic computing, 17(4):401–412.

KROON, R. S. 2003. Support vector machines, generalization bounds, and transduction. Stellenbosch: Universiteit van Stellenbosch. (Dissertation – M.Com.) 171 p.

KUHN, H.W. & TUCKER, A.W. 1951. Nonlinear programming. (*In* Neymann, J., *ed.* The Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, California, U.S.A., July 1 – August 12. Berkeley, CA: University of California Press. p. 481–492.)

KWASNIK, B.H.; CROWSTON, K.; NILAN, M.; & ROUSSINOV, D. 2000. Identifying document genre to improve web search effectiveness. Bulletin of the American Society for Information Science and Technology, 27(2): 23–26, Dec/Jan. http://www.asis.org/Bulletin/Dec-01/kwasnikartic.html Date of access: 6 Jan. 2008.

LAKOFF, G. 1973. Hedges: a study in meaning criteria and the logic of fuzzy concepts. Journal of philosophical logic, 2(4): 458–508, Oct.

LEA, M.R & STREET, B.V. 2000. Student writing and staff feedback in higher education: an academic literacies approach. (*In* Lea, M.R & Stierer, B, *eds.* Student writing in higher education. Buckingham: SRHE and Open University Press. p. 32–46.)

LEBART, L.; SALEM, A.; & BERRY, L. 1998. Exploring textual data. Dordrecht: Kluwer. 245 p.

LEE, D. Y. W. 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, Sept. http://llt.msu.edu/vol5num3/pdf/lee.pdf Date of access: 6 Jan. 2008.

LEE, Y.-B. & MYAENG, S.-H. 2002. Text genre classification with genre-revealing and subject-revealing features. (*In* the Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11–15, Tampere, Finland. New York, NY: ACM Press. p.145–150.) http://ir.icu.ac.kr/papers/text_genre.pdf Date of access: 6 Jan. 2008.

LEE, Y.-B. & MYAENG, S.-H. 2004. Automatic identification of text genres and their roles in subject-based categorization. (*In* the Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS'04), Track 4, Waikoloa, Island of Hawaii, January 5–8. Los Alamitos, CA: IEEE Computer Society. 10 p.) http://csdl2.computer.org/comp/proceedings/hicss/2004/2056/04/205640100b.pdf Date of access: 6 Jan. 2008.

LEMAIRE, B. & DESSU, P. 2001. A system to assess the semantic content of student essays. Journal of educational computing research, 24(3):305–320.

LEOPOLD, E. & KINDERMANN, J. 2002. How to represent texts in input space? Machine learning, 46(1–3):423–444, Jan.

LILLIEFORS, H.W. 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association, 62(318):399–402, Jun.

LIN, Y.; LEE, Y.; & WAHBA, G. 2002. Support vector machines for classification in nonstandard situations. Machine learning, 46(1–3):191–202, Jan.

LIU, Y. & ZHENG, Y.F. 2006. FS_SFS: a novel feature selection method for support vector machines. Pattern recognition, 39(7):1333–1345, Jul. Available: ScienceDirect. Date of access: 6 Jan. 2008.

LOUW, H. 2006. Standardising written feedback on L2 student writing. Potchefstroom: Noordwes-Universiteit. (Dissertation — M.A.) 218 p.

LUHN, H.P. 1957. A statistical approach to mechanical encoding and searching of literary information. IBM journal of research and development, 1(4):309–317.

MANNING, C.D. & SCHÜTZE, H. 1999. Foundations of statistical natural language processing. Cambridge, MA: MIT. 679 p.

MARCUS, M.P.; SANTORINI, B.; & MARCINKIEWICZ, M.A. 1993. Building a large annotated corpus of English: the Penn Treebank. Computational linguistics, 19(2): 313–330.

MARTIN, J.R. 1985. Process and text: two aspects of human semiosis. (*In* Benson, J.D & Greaves, W.S., *eds*. Systemic perspectives on discourse. Norwood, NJ: Ablex. p. 248–274.)

MCCLAVE, J.T. & SINCICH, T. 2000. Statistics. 8th ed. Upper Saddle River, NJ: Prentice-Hall. 484 p.

MEYER ZU EISSEN, S. & STEIN, B. 2004. Genre classification of web pages: user study and feasibility analysis. (*In* Biundo, S.; Frühwirth, T.; & Palm, G., *eds*. KI 2004: advances in artificial intelligence. Berlin: Springer. p. 256–269.)

MITCHELL, T.T. 1997. Machine learning. Boston, MA: McGraw-Hill. 414 p.

MOREALE, E. & VARGAS-VERA, M. 2003. Genre analysis and the automated extraction of arguments from student essays. (*In* Christie, J., *ed*. The Proceedings of the 7th International Computer Assisted Assessment Conference (CAA-2003), Loughborough, U.K., July 8–9. Loughborough: Loughborough University. p.271–286.) http://www.caaconference.com/pastConferences/2003/procedings/moreale.pdf Date of access: 6 Jan. 2008.

MOSTELLER, F. & WALLACE, D. L. 1964. Inference and disputed authorship: the Federalist papers. Reading, MA.: Addison-Wesley. 287 p.

NELSON, G. 2002. International Corpus of English markup manual for written texts. http://www.ucl.ac.uk/english-usage/ice/written.pdf Date of access: 6 Jan. 2008.

NESI, H.; SHARPLING, G.; & GANOBCSIK-WILLIAMS, L. 2004. Student papers across the curriculum: designing and developing a corpus of British student writing. Computers and composition, 21(4): 439–450, Dec. Available: Science Direct. Date of access: 6 Jan. 2008.

NOBLE, W.S. 2006. What is a support vector machine? Nature biotechnology, 24(12):1565–1567, Dec.

NYSTRAND, M. 1987. The role of context in written communication. (*In* Horowitz, R & Samuels, S.J., *eds.* Comprehending oral and written language. New York, NY: Academic Press. p. 197–214.)

OAKES, M. P. 1998. Statistics for corpus linguistics. Edinburgh: Edinburgh University Press. 287 p.

PLATT, J. 1999. Fast training of support vector machines using Sequential Minimal Optimization. (*In* Schölkopf, B.; Burges, C.; & Smola, A., *eds.* Advances in kernel methods — support vector learning. Cambridge, MA: MIT. p. 185–208.)

QUINLAN, J.R. 1993. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann. 302 p.

QUIRK, R. 1968. Essays on the English language: medieval and modern. London: Longman. 201 p.

QUIRK, R., GREENBAUM, S., LEECH, G., & SVARTVIK, J. 1985. A comprehensive grammar of the English language. London: Longman. 1779 p.

RAUBER, A. & MÜLLER-KÖGLER, A. 2001. Integrating automatic genre analysis into digital libraries. (*In* Fox, E.A. & Borgman, C.L., *eds.* The Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01), June 24–28, Roanoke, Virginia, U.S.A. New York, NY: ACM Press. p. 1–10.) http://www.ifs.tuwien.ac.at/ifs/research/pub_pdf/rau_jcdl01.pdf Date of access: 6 Jan. 2008.

REHM, G. 2002. Towards automatic web genre identification. (*In* the Proceedings of the 35th Hawaii International Conference on System Sciences (HICSS'02), Track 4, Waikoloa, Island of Hawaii, January 7–10. Los Alamitos, CA: IEEE Computer Society. 10 p.) http://csdl2.computer.org/comp/proceedings/hicss/2002/1435/04/14350101.pdf Date of access: 6 Jan. 2008.

RIETVELD, T. & VAN HOUT, R. 2005. Statistics in language research: analysis of variance. Berlin: Mouton de Gruyter. 265 p.

ROBERTS, P. 1960. The relation of Linguistics to the teaching of English. College English, 22(1):1–9, Oct.

ROSCH, E. 1975. Cognitive representations of semantic categories. Journal of experimental psychology, 104(3):192–233, Sept.

ROSS, D. & HUNTER, D. 1994. μ-EYEBALL: an interactive system for producing stylistic descriptions and comparisons. Computers and the humanities, 28(1):1–11.

ROSSO, M.A. 2005. Using genre to improve web search. Chapel Hill, NC: University of North Carolina. (Dissertation – Ph.D.) 275 p. http://ils.unc.edu/~rossm/Rosso_dissertation.pdf Date of access: 6 Jan. 2008.

ROUSSINOV, D.; CROWSTON, K.; NILAN, N.; KWASNIK, B.; CAI, J.; & LIU, X. 2001. Genre based navigation on the web. (*In* the Proceedings of the 34th Hawaii International Conference on System Sciences (HICSS'01), Track 4, Maui, Hawaii, January 3–6. Los Alamitos, CA: IEEE Computer Society. 10 p.) http://csdl2.computer.org/comp/proceedings/hicss/2001/0981/04/09814013.pdf Date of access: 6 Jan. 2008.

ROYSTON, J. P. 1982. An extension of Shapiro and Wilk's W test for normality to large samples. Applied statistics, 31(2):115–124.

SALKIND, N. 2004. Statistics for people who (think they) hate statistics. 2nd ed. Thousand Oaks, CA: Sage. 401 p.

SANTINI, M. 2004a. A shallow approach to syntactic feature extraction for genre classification. (*In* Lee, M., *ed.* The Proceedings of the 7th Annual Colloquium for the UK Special-Interest Group for Computational Linguistics (CLUK 7), 6–7 January, Birmingham, U.K.) ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-02.pdf Date of access: 6 Jan. 2008.

SANTINI, M. 2004b. State-of-the-art on automatic genre identification. Technical report, University of Brighton (U.K.). ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-03.pdf Date of access: 6 Jan. 2008.

SANTINI, M. 2005a. Genres in formation? An exploratory study of Web pages using cluster analysis. (*In* the Proceedings of the 8th Annual Colloquium for the UK Special-Interest Group for Computational Linguistics (CLUK 8), January 11, Manchester, U.K.) http://www.itri.brighton.ac.uk/~Marina.Santini/ITRI-05-01.pdf Date of access: 6 Jan. 2008.

SANTINI, M. 2005b. Linguistic facets for genre and text type identification: a description of linguistically-motivated features. Technical Report, University of Brighton (U.K.). http://www.nltg.brighton.ac.uk/home/Marina.Santini/linguistic_facets_tech_rep.pdf Date of access: 6 Jan. 2008.

SANTINI, M. 2005c. Clustering web pages to identify emerging textual patterns. (Poster presented at Rencontre des Etudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL), June 6–10, Dourdan, France.) (Unpublished.) http://www.nltg.brighton.ac.uk/home/Marina.Santini/_poster_recital_horizontal.pdf Date of access: 6 Jan. 2008.

SANTINI, M. 2006a. Common criteria for genre classification: annotation and granularity. (Paper delivered as part of the Workshop on Text-based Information Retrieval (TIR-06), Riva del Garda, Italy, August 29. p. 35–40) (Unpublished) http://www.uni-weimar.de/medien/webis/research/tir/tir-06/ Date of access: 6 Jan. 2008.

SANTINI, M. 2006b. Web pages, text types, and linguistic features: some issues. ICAME journal, 30: 67–86, Apr. http://icame.uib.no/ij30/ij30-page67-86.pdf Date of access: 6 Jan. 2008.

SAVILLE-TROIKE, M. 1982. The ethnography of communication: an introduction. Oxford: Blackwell. 209 p.

SCHÖLKOPF, B. & SMOLA, A.J. 2002. Learning with kernels: support vector machines, regularization, optimization, and beyond. Cambridge, MA: MIT. 624 p.

SCHÖLKOPF, B.; SMOLA, A.J.; WILLIAMSON, R. C.; & BARTLETT, P. L. 2000. New support vector algorithms. Neural computation, 12(5):1207–1245, May.

SCOTT, M. 2004. Oxford WordSmith Tools 4.0, Version 4.0.0.268. Oxford: Oxford University Press.

SHAPIRO, S. S.; WILK, M. B.; & CHEN, H. J. 1968. A comparative study of various tests for normality. Journal of the American Statistical Association, 63(324):1343–1372, Dec.

SHEPHERD, M.; WATERS, C.; & KENNEDY, A. 2004. Cybergenre: automatic identification of home pages on the web. The journal of web engineering, 3(3 & 4): 236–251. http://users.cs.dal.ca/~shepherd/pubs/JWE040722.pdf Date of access: 6 Jan. 2008.

SIEGEL, S. & CASTELLAN, N.J. JR. 1988. Nonparametric statistics for the behavioural sciences. 2nd ed. New York, NY: McGraw-Hill. 399 p.

SIGLEY, R. 1997. Text categories and where you can stick them: a crude formality index. International journal of corpus linguistics, 2(2):199–237.

SMITH, E.A. & KINCAID, P. 1970. Derivation and validation of the automated readability index for use with technical materials. Human factors, 12:457–464.

SMITH, N. 1997. Improving a tagger. (*In* Garside, R.; Leech, G.; & McEnery, A., *eds*. Corpus annotation: linguistic information from computer text corpora. London: Longman. p. 137–150.)

SPÄRCK JONES, K. 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1):11–21.

STAMATATOS, E.; FAKOTAKIS, N.; & KOKKINAKIS, G. 2000a. Text genre detection using common word frequencies. (*In* the Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000), Vol. 2, Saarbrücken, Germany, July 31 – August 4. San Francisco, CA: Morgan Kaufmann. p. 808–814.) http://delivery.acm.org/10.1145/1000000/992763/p808-stamatatos.pdf?key1=992763&key2=0100279911&coll=GUIDE&dl=GUIDE&CFID=11357499&CFTOKEN=78694650 Date of access: 6 Jan. 2008.

STAMATATOS, E.; FAKOTAKIS, N.; & KOKKINAKIS, G. 2000b. Automatic text categorisation in terms of genre and author. Computational linguistics, 26(4):471–495, Dec.

STATSOFT, Inc. 2006. STATISTICA (data analysis software system), Version 7.1.

STATSOFT, Inc. 2004. STATISTICA electronic manual. STATISTICA (data analysis software system), Version 7.

STECKING, R. & SCHEBESCH, K.B. 2003. Support vector machines for credit scoring: comparing to and combining with some traditional classification methods. (*In* Schader, M.; Gaul, W.; & Vichi, M., *eds*. Between data science and applied data analysis. Berlin: Springer. p. 604–612.)

SUC. 1997. SUC 1.0 Stockholm Umeå Corpus, Version 1.0. Department of Linguistics, Umeå University and Department of Linguistics, Stockholm University. http://www.ling.su.se/DaLi/suc/suc1.0_info.html Date of access: 6 Jan. 2008.

SVARTVIK, J. & QUIRK, R., *eds*. 1980. A corpus of English conversation. Lund: C.W.K. Gleerup. 893 p.

SWALES, J.M. 1990. Genre analysis: English in academic and research settings. Cambridge: Cambridge University Press. 260 p.

TAPANAINEN, P. & JÄRVINEN, T. 1997. A non-projective dependency parser. (*In* the Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97), Washington, DC, U.S.A., March 31 – April 3. San Francisco, CA: Morgan Kaufmann. p. 64–71.) http://portal.acm.org/ft_gateway.cfm?id=974568&type=pdf Date of access: 6 Jan. 2008.

TEUFEL, S. & MOENS, M. 1999. Discourse-level argumentation in scientific articles: human and automatic annotation. (*In* the Proceedings of the Association for Computational Linguistics (ACL'99) Workshop Towards Standards and Tools for Discourse Tagging, College Park, MD, U.S.A., June 22. Morristown, NJ: Association for Computational Linguistics.) http://www.aclweb.org/anthology-new/W/W99/W99-0311.pdf Date of access: 6 Jan. 2008.

TRUSHKINA, J. 2006. Automatic error detection in second language learners' writing. Language matters, 37(2):141–159.

TRUSS, L. 2003. Eats, shoots & leaves: the zero tolerance approach to punctuation. London: Profile. 209 p.

TULDAVA, J. 1995. Methods in quantitative linguistics. Trier: WVT Wissenschaftlicher. 187 p.

TURNER, J. 1999. Academic literacy and the discourse of transparency. (*In* Jones, C.; Turner, J.; & Street, B., *eds*. Students writing in the university: cultural and epistemological issues. Philadelphia: John Benjamins. p. 150–160.)

UCREL. 2006. CLAWS part-of-speech tagger. http://www.comp.lancs.ac.uk/ucrel/claws/ Date of access: 6 Jan. 2008.

VAN DE POEL, K. 2006. Scribende: academic writing for students. Acco: Leuven. 175 p.

VAN HALTEREN, H.; BAAYEN, R. H.; TWEEDIE, F.; HAVERKORT, M.; & NEIJT, A. 2005. New machine learning methods demonstrate the existence of a human stylome. Journal of quantitative linguistics, 12(1): 65–77, Apr.

VAN RIJSBERGEN, C.J. 1975. Information retrieval. London: Butterworths. 208 p.

VAPNIK, V.N. 1982. Estimation of dependencies based on empirical data. Berlin: Springer. 399 p.

VAPNIK, V.N. 2000. The nature of statistical learning theory. 2nd ed. New York: Springer. 314 p.

VAPNIK V.N & CHERVONENKIS, A. 1981. The necessary and sufficient conditions for uniform convergence of means to their expectations. Theory of probability and its applications, 26(3):532–553.

VAPNIK V.N & LERNER, A. 1963. Pattern recognition using generalized portrait method. Automation and remote control, 24:774–780.

WASTHOLM, P.; KUSMA, A.; & MEGYESI, B. B. 2005. Using linguistic data for genre classification. (*In* Funk P.; Rognvaldsson T.; & Xiong N., *eds*. The Proceedings of Swedish Artificial Intelligence Society and the Swedish Society for Learning Systems (SAIS-SSLS 2005), Mälardalen University, Västerås, Sweden, April 12–14.) http://stp.ling.uu.se/~bea/wastholm-megyesi-sais05.pdf Date of access: 6 Jan. 2008.

WEISS, G.M. & PROVOST, F. 2003. Learning when training data are costly: the effect of class distribution on tree induction. Journal of artificial intelligence research, 19:315–354, July/Dec.

WESTON, J.; MUKHERJEE, S.; CHAPELLE, O.; PONNTIL, M.; POGGIO, T.; & VAPNIK, V. 2001. Feature selection for SVMs. (*In* Solla, S.A.; Leen, T.K.; & Muller, K.-R. Advances in neural information processing systems, Vol. 13. Cambridge, MA: MIT. p. 668–674.)

WHITE, M.; CARDIE, C.; & NG, V. 2002. Detecting discrepancies in numeric estimates using multidocument hypertext summaries. (*In* the Proceedings of the 2nd International Conference on Human Language Technology Research, San Diego, California, U.S.A., March 24–27. San Francisco, CA: Morgan Kaufmann. p. 336–341.). http://www.hlt.utdallas.edu/~vince/papers/hlt02.pdf Date of access: 6 Jan. 2008.

WITTEN, I. & FRANK, E. 2000. Data mining: practical machine learning tools with Java implementations. San Francisco, CA: Morgan Kaufmann. 371 p

WITTGENSTEIN, L. 1953. Philosophical investigations. Oxford: Blackwell. 232 p

WOLTERS, M. & KIRSTEN, M. 1999. Exploring the use of linguistic features in domain and genre classification. (*In* the Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, Bergen, Norway, June 8–12. Morristown, NJ: Association for Computational Linguistics. p. 142–149.) http://delivery.acm.org/10.1145/980000/977055/p142-wolters.pdf?key1=977055&key2=2820279911&coll=GUIDE&dl=GUIDE&CFID=11357924&CFTOKEN=24954924 Date of access: 6 Jan. 2008.

YANG, Y. & PEDERSEN, J. O. 1997. A comparative study on feature selection in text categorization. (*In* Fisher, D.H., *ed.* The Proceedings of the 14th International Conference on Machine Learning (ICML'97), Nashville, Tennessee, U.S.A., July 8–12. San Francisco, CA: Morgan Kaufmann. p. 412–420.) http://citeseer.ist.psu.edu/cache/papers/cs/1982/http:zSzzSzwww.cs.cmu.eduzSz~yimingzSzpapers.yyzSzml97.pdf/yang97comparative.pdf Date of access: 6 Jan. 2008.

APPENDIX 1
Corpus information

A1.1. Subjects

Biochemistry	International Relations
Business	Law
Comparative American Studies	Literature
Computer Science	Management Science
Economics	Mathematics
English and Theatre	Medicine
European Industrial Relations	Philosophy
Film and Literature	Politics
French	Psychology
German	Sociology
History	Theatre and Performance Studies

Table A1.1.1: List of departments/courses from which essays used are drawn

A1.2. Essays used in this project

Essay number	**Grade**	**Department/subject**
0001a	71	Sociology
0001b	67	Sociology
0001c	68	Sociology
0001d	75	Sociology
0001e	67	Sociology
0002a	70	Sociology
0002b	71	Sociology
0004a	68	Sociology
0004b	75	Sociology
0004c	65	Sociology
0004d	67	Sociology
0005a	69	History

Essay number	Grade	Department/subject
0005b	72	History
0005c	74	History
0006a	85	Biological Sciences
0006c	65	Biological Sciences
0009d	85	Biological Sciences
0009e	65	Biological Sciences
0010a	66	History
0010b	65	History
0010c	73	History
0010d	70	History
0010e	72	Sociology
0011a	75	Psychology
0011b	75	Psychology
0011c	81	Psychology
0011d	75	Psychology
0011e	75	Psychology
0011f	85	Psychology
0011g	65	Psychology
0012a	70	History
0012b	72	History
0012c	70	Comparative American Studies
0012d	65	History
0013a	68	History
0013b	73	History
0013c	67	History
0013d	65	History
0014a	65	Psychology
0014b	65	Psychology
0014c	65	Psychology
0014d	85	Psychology
0014e	65	Psychology
0015a	68	History

Essay number	Grade	Department/subject
0015b	66	History
0016a	74	Psychology
0016c	72	Psychology
0017a	65	Psychology
0017b	65	Psychology
0019a	74	History
0019b	71	History
0019c	68	History
0019d	66	History
0019e	72	History
0019f	64	History
0019g	68	History
0019h	72	History
0019i	65	History
0019j	65	History
0020b	65	Psychology
0020c	75	Psychology
0020d	70	Psychology
0020e	68	Psychology
0020f	68	Psychology
0020g	68	Psychology
0020h	68	Psychology
0021c	75	Engineering
0022a	68	Psychology
0022b	65	Psychology
0022c	75	Psychology
0024a	67	Theatre and Performance Studies
0024b	71	Theatre and Performance Studies
0024c	68	Theatre and Performance Studies
0024d	67	Theatre and Performance Studies
0024e	71	Theatre and Performance Studies
0024f	67	Theatre and Performance Studies

Essay number	Grade	Department/subject
0024g	65	Theatre and Performance Studies
0024h	65	Theatre and Performance Studies
0026a	68	Philosophy
0026b	68	Philosophy
0029a	67	History
0029b	66	History
0029c	68	Comparative American Studies
0029f	65	History
0029h	65	Comparative American Studies
0029i	65	History
0029j	65	History
0029k	65	History
0029l	65	Comparative American Studies
0029m	65	Comparative American Studies
0029n	65	Comparative American Studies
0029o	65	History
0029p	65	Comparative American Studies
0030a	66	History
0030b	66	History
0031a	75	Psychology
0031b	85	Psychology
0031c	75	Psychology
0032a	85	Biological Sciences
0033a	84	Psychology
0033b	85	Psychology
0033c	90	Psychology
0033d	75	Psychology
0033e	75	Psychology
0034a	65	Politics
0034b	65	Politics
0034c	85	Politics
0034e	67	Politics

Essay number	Grade	Department/subject
0037a	72	Psychology
0037b	68	Psychology
0038a	69	Philosophy
0038b	65	Philosophy
0038c	69	Philosophy
0039a	67	History
0039b	66	History
0039d	65	History
0039e	64	History
0039f	64	History
0040a	70	History
0040b	67	History
0040d	67	History
0040e	68	History
0044a	70	History
0044b	73	History
0045a	79	Biological Sciences
0048a	85	Medicine
0053a	70	Economics
0053b	70	Economics
0053d	68	Economics
0058f	68	Economics
0063a	68	History and French
0063b	68	History and French
0063c	63	History
0063d	65	History and French
0064a	80	Law
0064b	68	Law
0064e	78	Law
0064f	70	Law
0069c	65	Law
0069d	65	Law

Essay number	Grade	Department/subject
0073a	74	Business
0073b	68	Business
0073c	73	Business
0073d	72	Business
0075a	76	Sociology
0075c	70	Politics
0075d	68	Politics
0075e	71	Politics
0075f	73	Politics
0075g	61	Sociology
0075i	63	Sociology
0075j	65	Politics
0075k	68	Politics
00751	66	Politics
0075m	72	Politics
0082a	68	Psychology
0082b	65	Psychology
0082c	62	Psychology
0082d	68	Psychology
0082e	80	Psychology
0082f	62	Psychology
0082g	68	Psychology
0082i	65	Psychology
0082j	75	Psychology
0084a	85	Psychology
0098a	65	History
0098b	68	History
0098c	68	Theatre
0117a	72	Economics
0118a	65	Economics
0119c	69	Law
0119d	70	Law

Essay number	Grade	Department/subject
0119e	65	Law
0119g	67	Law
0126b	68	English and Theatre
0126c	68	English and Theatre
0126d	73	English and Theatre
0126f	72	English and Theatre
0129a	65	Comparative American Studies
0129b	71	Comparative American Studies
0129c	63	History
0129d	70	Comparative American Studies
0129e	64	History
0129f	72	History
0129g	68	History
0129h	63	History
0129i	70	Comparative American Studies
0129j	68	Comparative American Studies
0130a	72	Comparative American Studies
0130b	69	Comparative American Studies
0130c	70	Comparative American Studies
0130d	70	Comparative American Studies
0130e	68	Comparative American Studies
0130f	72	Comparative American Studies
0130g	70	Comparative American Studies
0130h	68	Comparative American Studies
0130i	70	Comparative American Studies
0130j	73	Comparative American Studies
0131a	68	Psychology
0131b	65	Psychology
0135a	80	Politics
0135b	80	Politics
0135c	74	Politics
0135d	70	Politics

Essay number	Grade	Department/subject
0135e	64	Politics
0135f	85	Politics
0135g	65	Politics
0135h	70	Politics
0135i	71	Politics
0137a	66	History
0137b	67	History
0137c	60	Politics
0137e	69	Politics
0137f	71	Politics
0137h	65	Politics
0140a	65	Sociology
0140b	65	Sociology
0140c	65	Sociology
0140d	65	Sociology
0140f	65	Sociology
0143b	65	Law
0143c	65	Law
0143d	62	Law
0144a	65	History
0144b	71	History
0144c	67	History
0144d	66	History
0144e	68	History
0150a	65	Theatre
0150b	65	Theatre
0150c	65	Theatre
0169d	70	Business
0171a	60	Psychology
0171c	61	Psychology
0177b	85	Philosophy
0179a	85	Sociology

Essay number	Grade	Department/subject
0179b	65	Sociology
0179c	65	Sociology
0179d	65	Sociology
0179e	65	Sociology
0179f	65	Sociology
0179g	65	Sociology
0179h	65	Sociology
0179i	85	Sociology
0179j	65	Sociology
0181b	75	Physics
0190a	68	Psychology
0190b	68	Psychology
0190c	75	Psychology
0190d	68	Psychology
0191b	71	Law
0191c	65	Law
0192a	62	Psychology
0192b	62	Psychology
0192c	62	Psychology
0202a	70	Economics
0202b	65	Economics
0202c	74	Economics
0202d	63	Economics
0202e	70	Economics
0202f	64	Economics
0202g	65	Economics
0202h	62	Economics
0202i	67	Economics
0202m	70	Business
0202n	74	Business
0209a	70	Law
0209b	70	Law

Essay number	Grade	Department/subject
0209c	67	Law
0209e	70	Law
0209g	68	Law
0215b	70	Philosophy
0215c	67	Philosophy
0215d	68	Philosophy
0215e	62	Philosophy
0224a	64	Film and Literature
0224c	65	Film and Literature
0224d	72	Film and Literature
0224e	70	Film and Literature
0228c	65	Computer Science
0235a	72	Philosophy
0235b	78	Philosophy
0235c	63	Philosophy
0235d	74	Philosophy
0238a	62	Psychology
0238b	60	Psychology
0238c	62	Psychology
0238d	62	Psychology
0238e	60	Psychology
0240a	62	Psychology
0240b	75	Psychology
0240c	80	Psychology
0240d	65	Psychology
0240e	68	Psychology
0244d	62	Comparative American Studies
0244e	64	History
0244f	65	History
0244g	63	History
0244h	64	Comparative American Studies
0244i	63	Comparative American Studies

Essay number	Grade	Department/subject
0244j	64	Politics
0244k	70	Politics
0244l	63	Politics
0244m	60	Politics
0244n	61	Politics
0247a	65	Biological Sciences
0252a	67	Comparative American Studies
0252b	70	Comparative American Studies
0252c	67	History
0252d	67	Sociology
0252e	70	Sociology
0252f	64	Sociology
0252g	67	Sociology
0252h	69	Sociology
0252j	70	Sociology
0252k	72	Sociology
0252l	64	Sociology
0252m	64	Sociology
0252n	73	History
0252o	64	Comparative American Studies
0252p	66	Comparative American Studies
0252q	67	Sociology
0252r	72	History
0252s	70	Comparative American Studies
0252t	67	History
0255a	65	Politics
0255b	65	Politics
0255c	67	Politics
0255d	64	Politics
0255e	64	Politics
0255f	70	History
0255g	64	History

Essay number	Grade	Department/subject
0255h	70	History
0259a	70	Philosophy
0259b	70	Philosophy
0259c	80	Philosophy
0259d	73	Philosophy
0260a	65	Medicine
0260c	65	Medicine
0262a	75	Psychology
0262b	75	Psychology
0262c	68	Psychology
0262d	65	Psychology
0262e	67	Psychology
0262f	65	Psychology
0262g	68	Psychology

Table A1.2.1: List of essays from the BAWE and their grading in percentage

APPENDIX 2
Linguistic features

A2.1. Parts-of-speech

Tag	Description	Examples
CC	Coordinating conjunction	*and, but, either or, or*
CD	Cardinal number	includes roman numerals and fractions: *1, one, IV, two-thirds*
DT	Determiner	includes articles and determiners: *this, the, an, any, another*
EX	Existential *there*	*there* (was a party in progress)
FW	Foreign word[102]	*in vivo, a priori, et cetera,*
IN	Preposition or subordinating conjunction	*in, of, if, although*
JJ	Adjective	includes ordinal numbers: *first, green, mammary, acidic*
JJR	Adjective, comparative[103]	with the comparative ending *er* with a strictly comparative meaning: *greener, more, less, greater*
JJS	Adjective, superlative[104]	with superlative ending *est* and superlative meaning: *greenest, most, least, worst*
LS[105]	List item marker	*1., (a)*
MD	Modal	includes possibility, necessity and predictive modals: *can, ought, shall, may*
NN	Noun, singular or mass	*re-enactment, flour, dog, subclass*
NNP	Proper noun, singular	*John, UN, Brighton, Christmas*
NNPS	Proper noun, plural	*Labradors*
NNS	Noun, plural	*dogs, subclasses*
PDT	Pre-determiner	pre-qualifiers: *quite, rather, such*; pre-quantifiers: *all, half, many, nary*; and *both.*

[102] Foreign words would include *e.g.* and *i.e.* but this has not been done for this study.
[103] *More* and *less* are tagged JJR when alone or used as adjectives.
[104] *Most* and *least* are tagged JJS when alone or used as adjectives.
[105] It was decided to exclude LS as a variable for the final classifier development, as it is not reliably tagged.

Tag	Description	Examples
POS	Possessive ending	nouns ending in *'s*: *John's*, *Labradors'*, *UN's*
PRP	Personal pronoun	includes subject, object, reflexive pronouns, and the impersonal pronoun: *I*, *me*, *it*, *himself*
PRP$	Possessive pronoun	*my*, *your*, *mine*, *yours*
RB	Adverb	includes most words ending in *ly*, degree words, post head-modifiers, negative markers and nominal adverbs: *quite*, *enough*, *never*, *mildly*
RBR	Adverb, comparative	Adverbs with the comparative ending *er* with a strictly comparative meaning: *better*
RBS	Adverb, superlative	Adverbs with the comparative ending *est* with a strictly superlative meaning: *best*
RP	Particle	(*tell*) *off*, (*run*) *up*, (*break*) *through*
SYM [106]	Symbol	Should be used for mathematical, scientific or technical symbols: >, +, =, β
TO	*To*	*to*
UH	Interjection	*uh*, *well*, *yes*, *my*
VB	Verb, base form	subsumes imperatives, infinitives and subjunctives: *sing*, *be*, *do*, *have*
VBD	Verb, past tense	includes the conditional form of the verb to be: *sang*, *was*, *did*, *had*
VBG	Verb, gerund or present participle	*singing*, *being*, *doing*, *having*
VBN	Verb, past participle	*sung*, *been*, *done*, *had*
VBP	Verb, non-3rd person singular present	*sing*, *am*, *are*, *do*, *have*
VBZ	Verb, 3rd person singular present	*sings*, *is*, *does*, *has*
WDT	*Wh*-determiner	when used as a relative pronoun: *which*, *that*
WP	*Wh*-pronoun	*what*, *who*, *whom*
WP$	Possessive *wh*-pronoun	*whose*

[106] It was decided to exclude SYM as a variable for the final classifier development, as it is not reliably tagged.

Tag	Description	Examples
WRB	*Wh*-adverb	*how, where, why*

Table A2.1.1: List of Penn Treebank part-of-speech tags

Tag	Description	Examples
APPGE	Possessive pronoun, pre-nominal	*my, your, our*
AT	Article	*the, no*
AT1	Singular article	*a, an, every*
BCL	Before-clause marker	*in order (that), in order (to)*
CC	Coordinating conjunction	*and, or*
CCB	Adversative coordinating conjunction	*but*
CS	Subordinating conjunction	*if, because, unless, so, for*
CSA	*As* (as conjunction)	*as*
CSN	*Than* (as conjunction)	*than*
CST	*That* (as conjunction)	*that*
CSW	*Whether* (as conjunction)	*whether*
DA	After-determiner or post-determiner capable of pronominal function	*such, former, same*
DA1	Singular after-determiner	*little, much*
DA2	Plural after-determiner	*few, several, many*
DAR	Comparative after-determiner	*more, less, fewer*
DAT	Superlative after-determiner	*most, least, fewest*
DB	Before determiner or pre-determiner capable of pronominal function	*all, half*
DB2	Plural before-determiner	*both*
DD	Determiner (capable of pronominal function)	*any, some*
DD1	Singular determiner	*this, that, another*
DD2	Plural determiner	*these, those*
DDQ	*Wh*-determiner	*which, what*
DDQGE	*Wh*-determiner, genitive	*whose*

Tag	Description	Examples
DDQV	*Wh*-ever determiner,	*whichever, whatever*
EX	Existential *there*	*there*
FO	Chemical and mathematical formulae	*IQR/s*
FU	Unclassified word	
FW	Foreign word	*in vivo, a priori, et cetera*
GE	Germanic genitive marker	*', 's*
IF	*For* (as preposition)	*for*
II	General preposition	*in, if, although*
IO	*Of* (as preposition)	*of*
IW	*With, without* (as prepositions)	*with, without*
JJ	General adjective	*first, green, mammary, acidic*
JJR	General comparative adjective	with the comparative ending *er* with a strictly comparative meaning: *older, better, stronger*
JJT	General superlative adjective	with superlative ending *est* and superlative meaning: *oldest, best, strongest*
JK	Catenative adjective	*able in, be able to, willing in, be willing to*
MC	Cardinal number, neutral for number	*two, three*
MC1	Singular cardinal number	*one*
MC2	Plural cardinal number	*sixes, sevens*
MCGE*[107]	Genitive cardinal number, neutral for number	*two's, 100's*
MCMC	Hyphenated number	*40-50, 1770-1827*
MD	Ordinal number	*first, second, next, last*
MF	Fraction, neutral for number	*quarters, two-thirds*
ND1	Singular noun of direction	*north, southeast*
NN	Common noun, neutral for number	*sheep, cod, headquarters*
NN1	Singular common noun	*book, girl*

[107] Please note that all terms marked with an asterisk indicate terms that were either not found in the texts or which occurred in less than 1 percent of the texts. Therefore, these terms do not form part of the data used.

Tag	Description	Examples
NN2	Plural common noun	*books, girls*
NNA	Following noun of title	*M.A.*
NNB	Preceding noun of title	*Mr., Prof.*
NNL1	Singular locative noun	*island, street*
NNL2*	Plural locative noun	*islands, streets*
NNO	Numeral noun, neutral for number	*dozen, hundred*
NNO2	Numeral noun, plural	*hundreds, thousands*
NNT1	Temporal noun, singular	*day, week, year*
NNT2	Temporal noun, plural	*days, weeks, years*
NNU	Unit of measurement, neutral for number	*in, cc*
NNU1	Singular unit of measurement	*inch, centimetre*
NNU2	Plural unit of measurement	*ins., feet*
NP	Proper noun, neutral for number	*IBM, Andes*
NP1	Singular proper noun	*London, Jane, Frederick*
NP2	Plural proper noun	*Browns, Reagans, Koreas*
NPD1	Singular weekday noun	*Sunday*
NPD2*	Plural weekday noun	*Sundays*
NPM1	Singular month noun	*October*
NPM2*	Plural month noun	*Octobers*
PN	Indefinite pronoun, neutral for number	*none*
PN1	Indefinite pronoun, singular	*anyone, everything, nobody, one*
PNQO	Objective *wh*-pronoun	*whom*
PNQS	Subjective *wh*-pronoun	*who*
PNQV	*Wh*-ever pronoun	*whoever*
PNX1	Reflexive indefinite pronoun	*oneself*
PPGE	Nominal possessive personal pronoun	*mine, yours*
PPH1	3rd person singular neuter personal pronoun	*it*

Tag	Description	Examples
PPHO1	3rd person singular objective personal pronoun	*him, her*
PPHO2	3rd person plural objective personal pronoun	*them*
PPHS1	3rd person singular subjective personal pronoun	*he, she*
PPHS2	3rd person plural subjective personal pronoun	*they*
PPIO1	1st person singular objective personal pronoun	*me*
PPIO2	1st person plural objective personal pronoun	*us*
PPIS1	1st person singular subjective personal pronoun	*I*
PPIS2	1st person plural subjective personal pronoun	*we*
PPX1	Singular reflexive personal pronoun	*yourself, itself*
PPX2	Plural reflexive personal pronoun	*yourselves, themselves*
PPY	2nd person personal pronoun	*you*
RA	Adverb, after nominal head	*else, galore*
REX	Adverb introducing Appositional constructions	*namely, e.g.*
RG	Degree adverb	*very, so, too*
RGQ	*Wh-* degree adverb	*how*
RGQV	*Wh*-ever degree adverb	*however*
RGR	Comparative degree adverb	*more, less*
RGT	Superlative degree adverb	*most, least*
RL	Locative adverb	*alongside, forward*
RP	Preposition adverb, particle	*about, in*
RPK	Preposition adverb, catenative	*about in, be about to*
RR	General adverb	*quite, enough, never, mildly*
RRQ	*Wh-* general adverb	*where, when, why, how*
RRQV	*Wh*-ever general adverb	*wherever, whenever*

Tag	Description	Examples
RRR	Comparative general adverb	Adverbs with the comparative ending *er* with a strictly comparative meaning: *better, longer*
RRT	Superlative general adverb	Adverbs with the comparative ending *est* with a strictly superlative meaning: *best, longest*
RT	Quasi-nominal adverb of time	*now, tomorrow*
TO	Infinitive marker	*to*
UH	Interjection	*oh, yes, um*
VB0*	*Be*, base form	subsumes imperatives and subjunctives: *be*
VBDR	*Were*	*were*
VBDZ	*Was*	*was*
VBG	*Being*	*being*
VBI	*Be*, infinitive	*to be or not, it will be*
VBM	*Am*	*am*
VBN	*Been*	*been*
VBR	*Are*	*are*
VBZ	*Is*	*is*
VD0	*Do*, base form	subsumes imperatives and subjunctives: *do*
VDD	*Did*	*did*
VDG	*Doing*	*doing*
VDI	*Do*, infinitive	*I may do, to do*
VDN	*Done*	*done*
VDZ	*Does*	*does*
VH0	*Have*, base form	subsumes imperatives and subjunctives: *have*
VHD	*Had*, past tense	*had*
VHG	*Having*	*having*
VHI	*Have*, infinitive	*have*
VHN	*Had*, past participle	*had*
VHZ	*Has*	*has*
VM	Modal, auxiliary	*can, will, would*

Tag	Description	Examples
VMK	Modal, catenative	*ought, used*
VV0	Base form of lexical verb	*give, work*
VVD	Past tense of lexical verb	*gave, worked*
VVG	*ing* Participle of lexical verb	*giving, working*
VVGK	*ing* Participle, catenative	*going in, be going to*
VVI	Infinitive	*to give, it will work*
VVN	Past participle of lexical verb	*given, worked*
VVNK	Past participle, catenative	*bound in be bound to*
VVZ	*s* Form of lexical verb	*gives, works*
XX	Analytic negation	*not, n't*
ZZ1	Singular letter of the alphabet	*A, b*
ZZ2*	Plural letter of the alphabet	*A's, b's*

Table A2.1.2: List of UCREL CLAWS7 part-of-speech tags

A2.2. Punctuation marks

Tag	Punctuation mark
<colon>	colon *:*
<excl_mark>	exclamation mark *!*
<f_stop>	full stop .
<q_mark>	question mark *?*
<semi_colon>	semi-colon *;*

Table A2.2.1: List of punctuation tags

A2.3. Quotations

Tag	Description
<quote_int>	Quotations integrated with text
<quote_int_foreign>	Non-English quotations integrated with text
<quote_nonint>	Quotations non-integrated with text

Tag	Description
<quote_nonint_foreign>	Non-English quotations non-integrated with text
<s_q>	Sentence counts for non-integrated quotations

Table A2.3.1: List of quotation tags

A2.4. Nominalisations

Tag	Nominalisational suffix-singular	Tag	Nominalisational suffix - plural
<ism>	*ism*	<isms>	*isms*
<ity>	*ity*	<ities>	*ities*
<ment>	*ment*	<ments>	*ments*
<ness>	*ness*	<nesses>	*nesses*
<tion>	*tion*	<tions>	*tions*

Table A2.4.1: List of nominalisational suffixes and their respective tags

A2.5. Text statistics

1. word count;
2. word length in characters;
3. long words (> 6 characters);
4. type/token ratio;
5. sentence count;
6. sentence length in words;
7. paragraph count;
8. paragraph length in sentences;
9. readability score: the Automated Readability Index and the Flesch Reading Ease scores; and
10. number of references (tagged <reference>).

A2.6. Key function words

also	fundamentally	merely	seemingly	through
although	furthermore	more	significantly	thus
amongst	hence	moreover	similarly	to
arguably	highly	necessarily	simply	toward
as	however	nonetheless	socially	towards
because	importantly	not	specifically	ultimately
between	in	of	such	upon
concerning	increasingly	often	than	whether
consequently	indeed	perhaps	that	which
conversely	infact	previously	the	whilst
despite	inherently	primarily	their	within
due	itself	purely	themselves	yet
essentially	largely	rather	therefore	
established	lastly	regarding	these	
firstly	many	secondly	this	

Table A2.6.1: List of the key function words of the top 1000 key words

A2.7. Most frequent words in the BNC

a	by	if	she	was
all	for	in	so	we
an	from	is	that	were
and	had	it	the	when
are	has	more	their	which
as	have	not	there	who
at	he	of	they	will
be	her	on	this	with
been	his	one	to	would
but	I	or	up	you

Table A2.7.1: List of the top fifty words in the written section of the BNC

A2.8. Prepositions

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

Table A2.8.1: List of simple prepositions

Tag	Preposition	Tag	Preposition
<according_to>	according to	<nearer_to>	nearer to*
<ahead_of	ahead of	<next_to>	next to
<along_with>	along with	<on_to>	on to
<apart_from>	apart from	<out_of>	out of
<as_for>	as for	<outside_of>	outside of
<as_of>	as of	<owing_to>	owing to
<as_per>	as per*	<preliminary_to>	preliminary to*
<as_to>	as to	<preparatory_to>	preparatory to*
<away_from>	away from	<previous_to>	previous to*
<because_of>	because of	<prior_to>	prior to
<close_to>	close to	<pursuant_to>	pursuant to*
<contrary_to>	contrary to	<regardless_of>	regardless of
<devoid_of>	devoid of	<save_for>	save for*
<due_to>	due to	<subsequent_to>	subsequent to*
<except_for>	except for*	<thanks_to>	thanks to
<exclusive_of>	exclusive of*	<up_against>	up against
<inside_of>	inside of*	<up_to>	up to
<instead_of>	instead of	<upwards_of>	upwards of*
<irrespective_of>	irrespective of	<void_of>	void of

Tag	Preposition	Tag	Preposition
<near_to>	near to		

Table A2.8.2: List of two-word complex prepositions

Tag	Preposition	Tag	Preposition
<as_far_as>	as far as	<in_need_of>	in need of
<at_the_expense_of>	at the expense of	<in_place_of>	in place of
<at_the_hands_of>	at the hands of	<in_process_of>	in process of*
<at_variance_with>	at variance with*	<in_quest_of>	in quest of*
<by_dint_of>	by dint of*	<in_reference_to>	in reference to
<by_means_of>	by means of	<in_regard_to>	in regard to*
<by_virtue_of>	by virtue of	<in_relation_to>	in relation to
<by_way_of>	by way of	<in_respect_of>	in respect of*
<for_sake_of>	for sake of*	<in_respect_to>	in respect to*
<for_the_sake_of>	for the sake of	<in_return_for>	in return for
<for_want_of>	for want of*	<in_search_of>	in search of
<from_want_of>	from want of*	<in_spite_of>	in spite of
<in_accordance_with>	in accordance with	<in_the_face_of>	in the face of
<in_addition_to>	in addition to	<in_the_light_of>	in the light of
<in_aid_of>	in aid of*	<in_the_process_of>	in the process of
<in_back_of>	in back of*	<in_view_of>	in view of*
<in_case_of>	in case of*	<on_account_of>	on account of*
<in_common_with>	in common with*	<on_behalf_of>	on behalf of
<in_comparison_to>	in comparison to	<on_ground_of>	on ground of*
<in_comparison_with>	in comparison with*	<on_grounds_of>	on grounds of*
<in_compliance_with>	in compliance with*	<on_pain_of>	on pain of*
<in_conformity_with>	in conformity with*	<on_the_ground_of>	on the ground of*
<in_consequence_of>	in consequence of*	<on_the_grounds_of>	on the grounds of
<in_contact_with>	in contact with*	<on_the_matter_of>	on the matter of*
<in_exchange_for>	in exchange for	<on_the_part_of>	on the part of

Tag	Preposition	Tag	Preposition
<in_face_of>	in face of*	<on_the_strength_of>	on the strength of*
<in_favour_of>	in favour of	<on_top_of>	on top of
<in_front_of>	in front of	<with_reference_to>	with reference to
<in_lieu_of>	in lieu of*	<with_regard_to>	with regard to
<in_light_of>	in light of	<with_respect_to>	with respect to
<in_line_with>	in line with	<with_the_exception_of>	with the exception of

Table A2.8.3: List of three-word complex prepositions

A2.9. Reporting verbs

acknowledge	comment	exclaim*	proclaim	submit
add	complain*	explain	promise	suggest
admit	concede	forecast*	pronounce*	swear*
affirm	confess*	foretell*	protest	testify*
agree	confide*	guarantee	remark	vow*
allege*	confirm	hint	repeat	warn
announce*	contend	insist	reply	write
argue	convey	maintain	report	
assert	declare	mention	retort*	
boast*	deny	object	say	
claim	disclose	predict	state	

Table A2.9.1: List of public factual verbs

accept	determine	forget	mean	reflect
anticipate	discern	gather	note	remember
ascertain	discover	guess*	notice	reveal
assume	doubt	hear	observe	see
believe	dream	hold	perceive	show
calculate	ensure	hope	presume	signify
check	establish	imagine	presuppose*	suppose*
conclude	estimate	imply	pretend	suspect

conjecture*	expect	indicate	prove	think
consider	fancy*	infer	realise	understand
decide	fear	insure*	reason	
deduce	feel	judge	recall	
deem	find	know	reckon*	
demonstrate	foresee*	learn	recognise	

Table A2.9.2: List of private factual verbs

allow	demand	intend	prefer	resolve
arrange*	desire	move	pronounce	rule
ask	enjoin*	ordain*	propose*	stipulate*
beg*	entreat*	order	recommend*	urge
command	grant	pledge*	request	vote
decree*	instruct	pray*	require	

Table A2.9.3: List of suasive verbs

confirm	discuss

Table A2.9.4: List of miscellaneous reporting verbs

seem	appear

Table A2.9.5: List of perception verbs

A2.10. Conjunctions

again	eventually	instead	nonetheless	similarly
alternatively	finally	last	notwithstanding	subsequently
altogether	first	likewise	originally	then
consequently	firstly	meanwhile	otherwise	therefore
conversely	furthermore	moreover	overall	third
e.g.	hence	namely	rather	thirdly
else	however	nevertheless	second	thus

equally	i.e.	next	secondly	viz.*

Table A2.10.1: List of conjunctive adjuncts/conjuncts/linking adverbials

Tag	Conjunct	Tag	Conjunct
<as_a_consequence>	as a consequence	<in_comparison>	in comparison
<as_a_result>	as a result	<in_conclusion>	in conclusion
<by_comparison>	by comparison*	<in_consequence>	in consequence*
<by_contrast>	by contrast	<in_contrast>	in contrast
<first_of_all>	first of all*	<in_other_words>	in other words
<for_example>	for example	<in_particular>	in particular
<for_instance>	for instance	<in_sum>	in sum*
<in_addition>	in addition	<in_summary>	in summary
<in_any_case>	in any case	<in_the_first_place>	in the first place
<in_any_event>	in any event*	<in_the_second_place>	in the second place*

Table A2.10.2: List of multi-word conjunctive adjuncts/conjuncts/linking adverbials

although	once	unless	whereas	whilst
as	since	until	whereby	
because	though	when	whereupon*	
if	till	whenever	while	

Table A2.10.3: List of subordinating conjunctions

and	but	or

Table A2.10.4: List of coordinating conjunctions

A2.11. Downtoners

almost	just	partially	scarcely	virtually
barely	little	partly	simply	
but	merely	practically	slightly	
enough	nearly	quite	somewhat	

hardly	only	rather	sufficiently	

Table A2.11.1: List of downtoners

A2.12. Stance adverbs

accordingly	frankly*	mainly	reportedly*	typically
confidentially*	generally	speaking	technically	
figuratively*	honestly*	strictly	truthfully*	

Table A2.12.1: List of non-factual stance adverbs

actually	certainly	indeed	never	really
always	definitely	inevitably	obviously	undoubtedly

Table A2.12.2: List of factual stance adverbs

Tag	Adverb	Tag	Adverb
<in_fact>	in fact	<without_doubt>	without doubt
<of_course>	of course	<no_doubt>	no doubt

Table A2.12.3: List of two-word factual stance adverbs

apparently	most cases*	possibly	roughly
evidently	most instances*	predictably*	sort of
kind of	perhaps	probably	

Table A2.12.4: List of likelihood stance adverbs

amazingly*	disturbingly*	ironically	sensibly	wisely*
astonishingly*	hopefully*	regrettably*	surprisingly	
conveniently*	fortunately	rightly	unbelievably*	
curiously*	importantly	sadly	unfortunately	

Table A2.12.5: List of attitudinal stance adverbs

A2.13. Stance adjectives

acceptable	curious	hopeful*	odd*	thankful*
adamant*	depressed	horrible*	okay*	tragic
advisable*	desirable	hurt	paradoxical	typical
afraid	disappointed*	imperative	peculiar	unacceptable
alarmed*	dissatisfied*	incidental*	pleased	unaware
amazed*	distressed*	inconceivable	preferable	uncomfortable
amazing*	disturbed	incredible	reassured*	understandable
amused*	dreadful	indisputable	relieved*	unfair
angry	embarrassing	interesting	ridiculous	unfortunate*
annoyed*	encouraged	ironic	sad	unhappy
annoying*	essential	irritated*	satisfied	unlucky*
anomalous	extraordinary	lucky	sensible	unthinkable*
appropriate	fitting	mad	shocked	untypical*
astonished*	fortunate	natural	shocking	unusual
aware	frightened	neat	silly*	upset
awful*	funny*	necessary	sorry*	upsetting*
careful	glad*	nice*	strange	vital
concerned	good	notable	stupid	wonderful*
conceivable	grateful*	noteworthy	sufficient	worried
critical	great	noticeable	surprised*	
crucial	happy	obligatory	surprising	

Table A2.13.1: List of attitudinal stance adjectives

accepted	confident	false	positive	sure
apparent	convinced	impossible	proved	true
certain	correct	inevitable	plain	well-known*
clear	evident	obvious	right	

Table A2.13.2: List certainty/factual stance adjectives

doubtful	probable
likely	unlikely
possible	

Table A2.13.3: List of likelihood stance adjectives

apt	likely
certain	prone
due	sure
guaranteed	unlikely
liable	

Table A2.13.4: List of certainty stance adjectives

able	disposed	hesitant*	prepared	sufficient
anxious	doomed	inclined	quick	unable
bound	eager	insufficient	ready	unwilling
careful	eligible	keen	reluctant	welcome
competent	fit	loath*	set	willing
determined	greedy*	obliged	slow	

Table A2.13.5: List of ability/willingness stance adjectives

afraid	concerned	embarrassed*	impatient*	puzzled*
amazed*	content	free	indignant*	relieved*
angry	curious	furious*	nervous	sorry*
annoyed*	delighted*	glad*	perturbed*	surprised*
ashamed*	disappointed*	grateful	pleased	worried
astonished*	disgusted*	happy	proud	

Table A2.13.6: List of personal affective stance adjectives

difficult	pleasant
easier	possible
easy	tough*
hard	unpleasant

impossible	

Table A2.13.7: List of ease/difficulty stance adjectives

awkward*	criminal	important	reasonable	useless
appropriate	cumbersome	improper	right	unreasonable*
bad	desirable	inappropriate	safe	unseemly*
best	dreadful	interesting	sick	unwise*
better	essential	logical	silly*	vital
brave	expensive	lucky	smart*	wise
careless	foolhardy*	mad	stupid	wonderful*
convenient	fruitless	necessary	surprising	worse
crazy*	good	nice*	useful	wrong

Table A2.13.8: List of evaluation stance adjectives

A2.14. Nouns

assertion	discovery	knowledge	realization	statement
conclusion	doubt	observation	realisation	
conviction	fact	principle	result	

Table A2.14.1: List of factual stance nouns

assumption	feeling	indication	probability	thesis
belief	hypothesis	notion	rumour*	
claim	idea	opinion	sign	
contention	implication	possibility	suggestion	
expectation	impression	presumption	suspicion	

Table A2.14.2: List of likelihood stance nouns

comment	proposition	requirement
news	remark	
proposal	report	

Table A2.14.3: List of non-factual stance nouns

ground	view
hope	thought
reason	

Table A2.14.4: List of attitudinal stance nouns

agreement	determination	opportunity	reluctance	threat
authority	duty	plan	responsibility	wish
commitment	failure	potential	right	willingness
confidence	inclination	promise	scheme	
decision	intention	proposal	temptation	
desire	obligation	readiness	tendency	

Table A2.14.5: List of controlling nouns

APPENDIX 3

Data preparation and annotation

A3.1. Data cleaning

afraid	curious	importantly	perhaps	this
amongst	desirable	impossible	possible	through
angry	desire	in	promise	to
as	despite	indeed	reason	unlikely
at	doubt	likely	right	vital
between	dreadful	lucky	simply	which
but	due	mad	stupid	whilst
by	essential	merely	sufficient	with
careful	for	more	surprising	within
certain	from	necessary	than	worried
claim	good	not	that	
concerned	happy	of	the	
confirm	hope	on	their	

Table A3.1.1: List of terms with multiple occurrences in the word lists

A3.2. SVMTool

The usage for SVMTool is:
SVMTagger [options] <model> (Giménez & Màrquez, 2006:29). This research project used: ./bin/SVMTagger -V 1 -S LRL -T 4 SVMT.eng/WSJTP < "file to be tagged" > "tagged file", making use of the following options:

- V (verbose) was set to 1,
- S (tagging direction) was set to LRL (left-to-right and right-to-left),
- T (tagging strategy) was set to 4, which is claimed to be robust against unknown words (Giménez, & Màrquez, 2006:30).

The choice of these options was informed by Giménez (2006); however, many other options are available, which can be viewed in Giménez and Màrquez (2006).

APPENDIX 4
Support vector machines

A4.1. The optimal hyperplane classifier and the hard margin classifier

The simplest form of SVMs make use of a special hyperplane classifier called the maximal margin classifier, which separates only linearly separable data. Consider the case where S_t is $(\mathbf{x}_1, y_1),...,(\mathbf{x}_\ell, y_\ell)$, $\mathbf{x} \in \mathbb{R}^n$, and $y \in \{+1, -1\}$, where the values of y indicate binary classification. Thus, the input $\mathbf{x}$ is assigned to a class $y = -1$ if $f(\mathbf{x}) \leq 0$, or to $y = +1$ if $f(\mathbf{x}) \geq 0$. This function, $f(\mathbf{x})$, is a linear function of $\mathbf{x} \in X$ and is given by Cristianini and Shawe-Taylor (2000:9):

$$\begin{aligned} f(x) &= \langle \mathbf{w} \cdot \mathbf{x} \rangle + b \\ &= \sum_{i=1}^{n} w_i x_i + b \end{aligned} \tag{A4.1}$$

where $(\mathbf{w}, b) \in \mathbb{R}^n \times \mathbb{R}$ are the parameters controlling $f(\mathbf{x})$; $\mathbf{w}$ is the weight vector normal to the hyperplane and b is the threshold, which moves the hyperplane parallel to its former position. The decision rule, which governs the estimate of the target function $f(\mathbf{x})$, is $\operatorname{sgn}(f(\mathbf{x}))$, where $\operatorname{sgn}(0) = 1$. The decision rule, which governs the estimate of the target function $f(\mathbf{x})$, is $\operatorname{sgn}(f(\mathbf{x}))$, $\operatorname{sgn}(f(\mathbf{x})) := \begin{Bmatrix} 0 & \text{if } \mathbf{x} < 0, \\ 1 & \text{if } \mathbf{x} \geq 0 \end{Bmatrix}$. This situation can be more easily understood if interpreted geometrically as the training space is divided into two (classes) by the hyperplane defined by Equation A4.1.

The maximal margin (optimal) hyperplane is that hyperplane that separates the data points (feature vectors) without error and such that the distance between the plane and the nearest vector is maximised. The functional margin of a training instance $(\mathbf{x}_i, y_i)$ is defined with respect to the hyperplane $(\mathbf{w}, b)$ as (Cristianini & Shawe-Taylor, 2000:11):

$$\gamma_i = y_i(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b). \tag{A4.2}$$

This definition encompasses an inherent degree of freedom in that $\mathbf{w}$ and b can be rescaled from $(\mathbf{w}, b)$ to $(\lambda\mathbf{w}, \lambda b)$, $\lambda \in \mathbb{R}^+$, without changing the function, while

making the functional margin bigger. The solution to this situation implies constraining the value of $\mathbf{w}$. The geometric margin is defined as the functional margin of a normalised[108] weight vector and it remains unchanged if $\mathbf{w}$ and b are scaled as it is scaled by $\|\mathbf{w}\|_2$, where the length of $\mathbf{w}$ is defined in terms of the Euclidean notion of length (Cristianini & Shawe-Taylor, 2000:94–95; also Schölkopf & Smola, 2002:192). In order to determine the maximal margin hyperplane the geometric margin must be maximised; this can be done by minimising the weight vector's (Euclidean) norm and by making the functional margin equal to 1.[109] The latter ensures a constraint on the value of $\mathbf{w}$; it implies that $\langle \mathbf{w} \cdot \mathbf{x}^+ \rangle + b = +1$, and $\langle \mathbf{w} \cdot \mathbf{x}^- \rangle + b = -1$, thus the geometric margin can be given by:

$$\gamma = \frac{1}{2}\left(\left\langle \frac{\mathbf{w}}{\|\mathbf{w}\|_2} \cdot \mathbf{x}^+ \right\rangle - \left\langle \frac{\mathbf{w}}{\|\mathbf{w}\|_2} \cdot \mathbf{x}^- \right\rangle \right)$$

$$= \frac{1}{2\|\mathbf{w}\|_2}\left(\langle \mathbf{w} \cdot \mathbf{x}^+ \rangle - \langle \mathbf{w} \cdot \mathbf{x}^- \rangle \right)$$

$$= \frac{1}{\|\mathbf{w}\|_2}. \qquad \text{(A4.3)}$$

Note that if $\|\mathbf{w}\|_2$ is made a unit vector and thus of length equal to 1, then the functional and geometric margins will be the same.

The above discussion leads to the actual optimisation problem here. The maximal margin hyperplane, which will successfully separate S_t by realising (A4.3), must minimise the objective function: $\Phi(\mathbf{w}) = \langle \mathbf{w} \cdot \mathbf{w} \rangle = \frac{1}{2}\|\mathbf{w}\|^2$ (A4.4)

subject to the following inequality constraint: $y_i\left(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b\right) \leq 1,\ i = 1,\ldots,\ell$ (A4.5)

(Vapnik, 2000:132).[110] This problem is known as the primal optimisation problem.

---

[108] This term refers to the division of the margin by $\|\mathbf{w}\|_2$.

[109] Hyperplanes of the functional margin equal to 1 are called canonical hyperplanes (Schölkopf & Smola, 2002:190–191).

[110] $\|\mathbf{w}\|$ is made $\frac{1}{2}\|\mathbf{w}\|^2$ (where $\|\mathbf{w}\|^2$ is the dot product $\langle \mathbf{w} \cdot \mathbf{w} \rangle$) in order to make the objective function of the optimisation problem quadratic. For a detailed discussion of optimisation theory, see Cristianini and Shawe-Taylor (2000:79–80); also Schölkopf and Smola (2002:149–184).

The primal optimisation problem can be solved by using the Lagrangian function.[111] This function is defined as the objective function, as given in (A4.4), and a linear combination of the constraints, as given (A4.5) (Cristianini & Shawe-Taylor, 2000:83). The coefficients of this linear combination are the Lagrange multipliers. The primal Lagrangian for this problem is:

$$L_p(\mathbf{w},b,\boldsymbol{\alpha}) = \frac{1}{2}\langle \mathbf{w}\cdot\mathbf{w}\rangle - \sum_{i=1}^{\ell}\alpha_i\left[y_i\left(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b\right)-1\right], \tag{A4.6}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers. The Lagrangian must be minimised with respect to $\mathbf{w}$ and b (the primal variables) and be maximised with respect to $\alpha_i \geq 0$ (the dual variables). The dual representation of the Lagrangian has the same optimal solution as the primal (the strong duality theorem, see Cristianini & Shawe-Taylor, 2000:85–86). The dual is relevant for SVMs as it has the number of variables equal to the ℓ rather than to the number of attributes, consequently reducing dimensionality (Abe, 2005:18).

Thus, in order to find the dual Lagrangian, a saddle point must be found and the primal variables eliminated. A saddle point indicates the optimal solution of both the primal and dual problems values and is called the zero duality gap, which designates zero difference in solution values for both problems (Cristianini & Shawe-Taylor, 2000:86). According to the Karush-Kuhn-Tucker (KKT) conditions, the saddlepoint is the point at which the partial derivatives of L with respect to the primal variables must vanish. This is a minimum point if the function is convex.[112] Such a point will thus ensure the minimisation of the primal variables. Following KKT, this saddlepoint is found by setting the partial derivatives of the primal variables to zero (Cristianini & Shawe-Taylor, 2000:95–96):

$$\frac{\partial L(\mathbf{w},b,\boldsymbol{\alpha})}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{\ell} y_i\alpha_i\mathbf{x}_i = \mathbf{0}, \text{ and} \tag{A4.7}$$

$$\frac{\partial L(\mathbf{w},b,\boldsymbol{\alpha})}{\partial b} = \sum_{i=1}^{\ell} y_i\alpha_i = 0, \tag{A4.8}$$

[111] For more on the Lagrangian see Cristianini and Shawe-Taylor (2000:81–88); and Vapnik (2000:134–135).

[112] For more on KKT conditions see Kuhn and Tucker (1951), Cristianini and Shawe-Taylor (2000:87), and Schölkopf and Smola (2002:166–167).

which yields the relations:

$$\mathbf{w} = \sum_{i=1}^{\ell} y_i \alpha_i \mathbf{x}_i, \tag{A4.9}$$

$$0 = \sum_{i=1}^{\ell} y_i \alpha_i. \tag{A4.10}$$

These relations in (A4.9) and (A4.10) are then substituted back into (A4.6) to yield the Lagrangian dual (L_d):[113]

$$L_p(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\langle \mathbf{w} \cdot \mathbf{w} \rangle - \sum_{i=1}^{\ell} \alpha_i \left[y_i \left(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \right) - 1 \right]$$

$$= \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle - \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle + \sum_{i=1}^{\ell} \alpha_i$$

$$= L_d(\boldsymbol{\alpha}) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle. \tag{A4.11}$$

In this way, the Lagrangian variables can be used to solve the quadratic optimisation problem, where (A4.11) is subject to the constraints $\sum_{i=1}^{\ell} y_i \alpha_i = 0$, and $\alpha_i \geq 0$, $i = 1, \ldots, \ell$; and the weight vector $\mathbf{w}^* = \sum_{i=1}^{\ell} y_i \alpha_i^* \mathbf{x}_i$ realises the optimal hyperplane with the geometric margin of Equation A4.3, $\gamma = \frac{1}{\|\mathbf{w}^*\|_2}$ (Cristianini & Shawe-Taylor, 2000:96).

One of the relations in the KKT conditions, the KKT complementarity condition (Abe, 2005:310) is given by:

$$\alpha_i^* \left[y_i \left(\langle \mathbf{w}^* \cdot \mathbf{x}_i \rangle + b^* \right) - 1 \right] = 0, \; i = 1, \ldots, \ell. \tag{A4.12}$$

This relation allows for the calculation of b, which is not solved in the dual problem. Thus by selecting any i for which $\alpha_i \neq 0$ the value of b can be found (Burges, 1998:130–131). However, for precision,[114] the mean of the calculated values for b^* (if $\alpha_i \neq 0$) for positive α's is usually computed. Cristianini and Shawe-Taylor (2000:96; also Boser *et al.*, 1992:147) calculate b^* using the primal constraints as follows:

[113] This dual problem is also known as the Wolfe dual problem (Hastie, Tibshirani, & Friedman, 2001:109). See also Schölkopf and Smola (2002:171–172) for more on this.
[114] See Burges (1998:131), Hechter (2004:46), and Abe (2005:20).

$$b^* = -\frac{\max_{y_i=-1}\left(\left\langle \mathbf{w}^* \cdot \mathbf{x}_i \right\rangle\right) + \min_{y_i=1}\left(\left\langle \mathbf{w}^* \cdot \mathbf{x}_i \right\rangle\right)}{2}. \quad \text{(A4.13)}$$

In addition to providing the value for b, the KKT condition given in (A4.12) also indicates that for $\alpha_i > 0$ the constraint is active and for $\alpha_i = 0$ it is inactive.[115] *Support vectors* (*SVs*) refer to training instances ($\mathbf{x}_i$) for which the constraint is active ($\alpha_i > 0$; that is non-zero). Vapnik (1982) shows that if the data are separated using such an optimal hyperplane that the expectation value of the probability of classification error on a test set is bounded by the ratio of expected number of *SVs* and training vectors (S_t).

The decision function for new data points (with each point $\in \mathbb{R}^N$) is thus (Schölkopf & Smola, 2002:14):

$$\operatorname{sgn}\left(f(\mathbf{x})\right) = \sum_{i=1}^{\ell} y_i \alpha_i^* \left\langle \mathbf{w} \cdot \mathbf{x}_i \right\rangle + b^*. \quad \text{(A4.14)}$$

Thus if $\operatorname{sgn}\left(f(\mathbf{x})\right) > 0$, then the data point is classified as $y = +1$; similarly, if $\operatorname{sgn}\left(f(\mathbf{x})\right) < 0$, the data point is classified as $y = -1$ (Abe, 2005:20).

A4.2. The soft margin classifier

Cortes and Vapnik (1995) introduce a solution[116] to the problem of noisy data by the introduction of the slack variables:

$\xi_i \geq 0,\ i = 1,\ldots,\ell$; and (A4.15)

relaxed separation constraints:

$$y_i\left(\left\langle \mathbf{w} \cdot \mathbf{x}_i \right\rangle + b\right) \geq 1 - \xi_i,\ i = 1,\ldots,\ell. \quad \text{(A4.16)}$$

The slack variables thus allow the constraints of the margin (A4.5) to be violated on condition that there is some increase in the value of the objective function (A4.4) at the optimal solution. Thus will allow the training set to be separated with some errors, but

[115] An equality constraint is active if the solution weight vector that satisfies the inequality constraint is equal to zero and is inactive if it does not (Cristianini & Shawe-Taylor, 2000:80).
[116] See Schölkopf and Smola (2002: 204) for more potential solutions and some rebuttals.

without allowing too many misclassification errors, whilst ensuring that the maximal margin requirements are achieved.

If ξ_i is large enough, the constraint in (A4.16) can always be met. But this may result in a trivial solution with all ξ_i taking on large values. In order to prevent this ξ_i are penalised in (A4.4). As a particular $\xi_i > 1$ is required for an error to occur, $\sum_i \xi_i$ is an upper bound on the training errors (Schölkopf & Smola, 2002:16). The simplest way of including this in (A4.4) is called the C-SV classifier; the objective function is now given by (Schölkopf & Smola, 2002:205):

$$\Phi(\mathbf{w},\xi) = \frac{1}{2}\langle \mathbf{w}\cdot\mathbf{w}\rangle + \frac{C}{\ell}\sum_{i=1}^{\ell}\xi_i,\ C>0, \tag{A4.17}$$

subject to (A4.15) and (A4.16).

For the L_2-norm case, the primal Lagrangian for (A4.17) is given by Cristianini and Shawe-Taylor (2000:105):

$$L_p(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\alpha}) = \frac{1}{2}\langle \mathbf{w}\cdot\mathbf{w}\rangle + \frac{C}{2}\sum_{i=1}^{\ell}\xi_i^2 - \sum_{i=1}^{\ell}\alpha_i\left[y_i\left(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b\right) - 1 + \xi_i\right]. \tag{A4.18}$$

Similar to the optimisation problem in Section 3.7.2, the dual is calculated by setting the value of the partial derivatives to zero to obtain the relations (Cristianini & Shawe-Taylor, 2000:105):

$$\frac{\partial L_p(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\alpha})}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{\ell} y_i\alpha_i\mathbf{x}_i = \mathbf{0}, \tag{A4.19}$$

$$\frac{\partial L_p(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\alpha})}{\partial \boldsymbol{\xi}} = C\boldsymbol{\xi} - \boldsymbol{\alpha} = \mathbf{0}, \text{ and} \tag{A4.20}$$

$$\frac{\partial L_p(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\alpha})}{\partial b} = \sum_{i=1}^{\ell} y_i\alpha_i = 0. \tag{A4.21}$$

The relations obtained in (A4.19), (A4.20) and (A4.21) are then substituted into L_p to obtain L_d (Cristianini & Shawe-Taylor, 2000:105):

$$\begin{aligned} L_p(\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\alpha}) &= \sum_{i=1}^{\ell}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{\ell} y_i y_j \alpha_i\alpha_j\langle \mathbf{x}_i\cdot\mathbf{x}_j\rangle + \frac{1}{2C}\langle\boldsymbol{\alpha}\cdot\boldsymbol{\alpha}\rangle - \frac{1}{C}\langle\boldsymbol{\alpha}\cdot\boldsymbol{\alpha}\rangle \\ &= L_d(\boldsymbol{\alpha}) = \sum_{i=1}^{\ell}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{\ell} y_i y_j \alpha_i\alpha_j\langle \mathbf{x}_i\cdot\mathbf{x}_j\rangle - \frac{1}{2C}\langle\boldsymbol{\alpha}\cdot\boldsymbol{\alpha}\rangle. \end{aligned}$$

This is equivalent to:

$$L_d(\boldsymbol{\alpha}) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} y_i y_j \alpha_i \alpha_j \left(K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{C} \delta_{ij} \right), \tag{A4.22}$$

where $\delta_{ij} = 1$ if $i = j$ and 0.[117] Maximising this problem where (A4.22) is subject to the constraints: $\sum_{i=1}^{\ell} y_i \alpha_i = 0$, and $\alpha_i \geq 0$, $i = 1, \ldots, \ell$; and the weight vector (as with the fully separable case) $\mathbf{w}^* = \sum_{i=1}^{\ell} y_i \alpha_i^* \mathbf{x}_i$ realises the maximal margin hyperplane of the geometric margin (Cristianini & Shawe-Taylor, 2000:106):

$$\gamma = \left(\sum_{i \in sv} \alpha_i^* - \frac{1}{C} \langle \boldsymbol{\alpha}^* \cdot \boldsymbol{\alpha}^* \rangle \right)^{\frac{1}{2}}. \tag{A4.23}$$

Similar to the hard margin classifier, the KKT complementarity condition necessary for calculating b as discussed in Section 3.7.2 for this problem is given by Cristianini and Shawe-Taylor (2000:106):

$$\alpha_i \left[y_i \left(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \right) - 1 + \xi_i \right] = 0, \; i = 1, \ldots, \ell. \tag{A4.24}$$

The value of b can thus be given by Abe (2005:38):

$$b = y_i - \sum_{j=1}^{\ell} \alpha_j y_j \left(K(\mathbf{x}_i, \mathbf{x}_j) + \frac{\delta_{ij}}{C} \right) \tag{A4.25}$$

The decision function is the same as that of the hard margin classifier; in kernel form, it is given by Abe (2005:38):

$$\operatorname{sgn}\left(f(x) \right) = \sum_{i=1}^{\ell} y_i \alpha_i K(x, x_i) + b. \tag{A4.26}$$

The objective function for the ν- SV classifier in (A4.4) is given by Schölkopf and Smola (2002:206):

$$\Phi(\mathbf{w}, \boldsymbol{\xi}, \rho) = \frac{1}{2} \langle \mathbf{w} \cdot \mathbf{w} \rangle - \nu\rho + \frac{1}{\ell} \sum_{i=1}^{\ell} \xi_i, \tag{A4.27}$$

subject to the constraints: $y_i \left(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \right) \geq \rho - \xi_i$, $i = 1, \ldots, \ell$, and (A4.28)

$$\xi_i \geq 0, \; \rho \geq 0. \tag{A4.29}$$

[117] The Kronecker delta (see Abe, 2005:38).

For the ν- *SV* classifier the primal Lagrangian for (A4.26) is given by Schölkopf and Smola (2002:207):

$$L_p(\mathbf{w},\boldsymbol{\xi},b,\rho,\boldsymbol{\alpha},\boldsymbol{\beta},\delta)=\frac{1}{2}\|\mathbf{w}\|^2-\nu\rho+\frac{1}{\ell}\sum_{i=1}^{\ell}\xi_i-\sum_{i=1}^{\ell}\left(\alpha_i\left(y_i\left(\langle\mathbf{w}\cdot\mathbf{x}_i\rangle+b\right)-\rho+\xi_i\right)+\beta_i\xi_i\right)-\delta\rho,$$

$$\alpha_i,\beta_i,\delta\geq 0. \tag{A4.30}$$

Again, in order to establish the dual, the primal variables $(\mathbf{w},\boldsymbol{\xi},b,\rho)$ are minimised and the dual variables maximised $(\boldsymbol{\alpha},\boldsymbol{\beta},\delta)$. In the usual method, as exemplified above, by setting each primal to zero the following relations are obtained:

$$w=\sum_{i=1}^{\ell}y_i\alpha_i\mathbf{x}_i, \tag{A4.31}$$

$$\alpha_i+\beta_i=\frac{1}{\ell}, \tag{A4.32}$$

$$\sum_{i=1}^{\ell}\alpha_i y_i=0, \text{ and} \tag{A4.33}$$

$$\sum_{i=1}^{\ell}\alpha_i-\delta=\nu. \tag{A4.34}$$

The dual is then obtained by substituting the above relations into (A4.29), which gives:

$$L_d(\boldsymbol{\alpha})=-\frac{1}{2}\sum_{i,j=1}^{\ell}y_i y_j\alpha_i\alpha_j K\left(\mathbf{x}_i,\mathbf{x}_j\right), \tag{A4.35}$$

subject to the constraints:

$$0\leq\alpha_i\leq\frac{1}{\ell}, \tag{A4.36}$$

$$\sum_{i=1}^{\ell}\alpha_i y_i=0, \tag{A4.37}$$

$$\sum_{i=1}^{\ell}\alpha_i\geq\nu. \tag{A4.38}$$

The KKT complementarity condition for this problem is given by:

$$\alpha_i^*\left[y_i\left(\left\langle\mathbf{w}^*\cdot\mathbf{x}_i\right\rangle+b^*\right)-\rho+\xi_i\right]=0,\ i=1,\ldots,\ell. \tag{A4.39}$$

In order to give b and the margin parameter p, it is necessary to make use of the KKT complementarity condition given in (A4.39) above. Using sets $S_{\pm}$ of the same size $s>0$, which contain SVs whose dual variables are $0<\alpha_i<1$ and classes $y_i=\pm 1$, gives:

$$b=-\frac{1}{2s}\sum_{x\in S_+\cup S_-}\sum_{j=1}^{\ell}\alpha_i y_i K\left(\mathbf{x},\mathbf{x}_j\right), \text{ and} \tag{A4.40}$$

$$\rho=\frac{1}{2s}\left(\sum_{x\in S_+}\sum_{j=1}^{\ell}\alpha_j y_j K\left(\mathbf{x},\mathbf{x}_j\right)-\sum_{x\in S_-}\sum_{j=1}^{\ell}\alpha_j y_j K\left(\mathbf{x},\mathbf{x}_j\right)\right). \tag{A4.41}$$

The decision function is the same as that given in (A4.26).

APPENDIX 5
Feature selection[118]

Normal variables		Non-normal variables	
***T*-test**	**Feature Selection and Variable Screening module**	**Mann-Whitney *U***	**Kolmogorov-Smirnov Two-Sample Test**
	***k*=2**		
ARI readability	vvz	EX	PDT
Flesch readability	vbdz	PDT	POS
mean word length	ARI readability	POS	VBD
DT	VBZ	VBD	VBG
JJ	to	VBG	find
jj	TO	VBP	recall
NN	JJ	agree	actually
not	a	suggest	never
RP	this	find	in fact
there	when	forget	result
this	JJS	recall	as to
TO	the	reflect	on to
to	NN	command	had
vbdz	**k=3**	signify	their
vvz	Flesch readability	similarly	they
Words >6	jj	themselves	up
when	of	thus	was
	JJ	<isms>	were
	SUCH	evident	<reference>
	full stop	obvious	ge
	mean word length	bound	io
	k=4	insufficient	ppho2
	jj	keen	pphs2
	when	unfortunately	rl

[118] Note that the CLAWS7 tags are written in small letters and the Penn Treebank tags in capitals for easy distinction. Also note that words in capitals are from the keywords list, see Appendix 2.

Normal variables		Non-normal variables	
	there	readiness	rp
	words >6	easy	vbdr
	RP	pleasant	vdd
	has	unpleasant	vdi
	VB	vital	vhd
	k=5	actually	vhn
	jj	inevitably	<quote_nonint>
	Flesch readability	never	<s_q>
	to	obviously	
	TO	in fact	
	a	result	
	WP	statement	
	ARI readability	possibly	
	mean sentence length	sort of	
	JJ	suggestion	
	when	typically	
	there	content	
	k=6	as for	
	to	as to	
	TO	on to	
	jj	by means of	
	this	for the sake of	
	mean sentence length	in favour of	
	a	in line with	
	k=7	in search of	
	jj	had	
	Flesch readability	they	
	of	up	
	this	was	
	k=8	were	
	this	eventually	

Normal variables		Non-normal variables	
	when	whenever	
	a	appear	
	readiness	<reference>	
	jj	eg	
	to	csw	
	TO	ex	
	ARI readability	fo	
	of	ge	
	nnl1	nn1	
	k=9	npd1	
	vbdz	ppho2	
	jj	ppbs2	
	RP	ppio2	
	Flesch readability	ppx2	
	to	rp	
	TO	to	
	k=10	vbdr	
	vbdz	vdd	
	when	vdi	
	jj	vhd	
	is	vhn	
	this	vv0	
	of	vvd	
	TO	vvg	
	to	<quote_nonint>	
	CC	<s_q>	

Table A5.1: List of features selected

NOTATION AND ABBREVIATIONS[119]

BOW	bag-of-words
POS	part-of-speech
SVM	support vector machine
G	generator of the data
S	supervisor
LM	learning machine
$y \in Y$	output and output space
$\mathbf{x} \in X$	input and input space
S_t	training set
ℓ	number of observations in the training set
x_i	observations
y_i	class labels of observations
γ_i	margin
$\boldsymbol{\alpha}$	Lagrange multipliers/dual variables
L_p	primal Lagrangian
L_d	dual Lagrangian
ξ	slack variables
F	feature space
$\langle \mathbf{x} \cdot \mathbf{z} \rangle$	inner (dot) product of $\mathbf{x}$ and $\mathbf{z}$
$\phi : X \to F$	mapping to feature space
$K(\mathbf{x}, \mathbf{z})$	kernel $\langle \phi(\mathbf{x}) \cdot \phi(\mathbf{z}) \rangle$
$\mathbf{w}$	weight vector
b	bias
$\|\cdot\|_p$	p-norm
$\mathbb{R}$	set of real numbers
$\mathbb{R}^+$	set of real positive numbers
$\mathbb{R}^N$	set of natural numbers
SV	support vector
C	cost parameter
RB	recall of 'bad' examples
RG	recall of 'good' examples
AR	average recall of 'bad' and 'good' examples
TLN	text-length normalised data
RAW	raw data
LOG	logarithm-transformed data
IDF	inverse document frequency data
CV	cross-validation
MGR	middle-band grades removed

[119] Bold and non-italicised text indicates a vector; italicised text indicates a scalar.