Identifying Translation Effects in English Natural Language Text

Gerard Lynch

May 7, 2013

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

_____

Gerard Lynch

Summary

With the rise in popularity of applying machine learning methods to problems in textual stylometry, the increased availability of machine-readable corpora and the emerging benefits of research on corpora of translated text in the field of machine translation, there has been a corresponding increase in interest in the analysis of translated text by computational linguists, a subject which until recent years remained the preserve of translation studies scholars. This thesis details the state-of-the art in research comprising the fields of computational linguistics, translation studies and the digital humanities and describes experiments carried out using machine-learning tools on a selection of comparable corpora of translations in English with regard to three main research questions: defining markers of translated vs. original text in the same genre, obtaining source language markers in literary translations and the detection of the stylistic traces of a literary translator.

Supervised learning experiments are carried out on a number of comparable corpora of translated text, with a focus on identifying features which capture the range of translation effects mentioned. The features used in this thesis are ngram-based, consisting of ngrams of words and parts-of-speech, and document-level, which consist of the frequencies of a class of textual items and various other metrics including type-token ratios and readability scores.

Chapter 4 describes experiments on two sets of comparable corpora in English, the Europarl corpus and a corpus of translated and original articles from the online version of the New York Times, with the goal of mining features of translated language, or *translationese*. Support Vector Machines are used along with Naive Bayes and Simple Logistic classifiers on these corpora, with the task of classifying the translated side of the corpora from the non-translated side. Classification accuracy was circa. 80% for the Europarl corpus and slightly less for the NYT corpus, using a mixed feature set of the features mentioned above. The different genres of the corpora resulted in generally non-intersecting distinguishing feature sets for each corpus, however there were a small number of features in common. Classifiers which were trained on Europarl and tested on the NYT corpus reported poor results, which corroborated results from the literature by Koppel and Ordan (2011) on different dialects of *translationese*.

Chapter 5 tackles the question of source language detection in translations as examined in the Europarl corpus by van Halteren (2008). The corpus focused on here is a corpus of literary text from the nineteenth century, comprising of texts translated from German, French and Russian, with English original texts also included in the experiments. Using comparable

experimental methodology to the previous chapter, classifiers were trained on the corpus, with the task of classifying the source language of a text, a four or three class classification problem. Accuracy results varied from 99% using a feature set of the 500 most distinguishing word unigrams to 85% for a feature set containing document metrics, POS bigrams and common words. This classifier was also tested on a separate but comparable set of texts from the same literary period in order to examine the classifier performance on unknown data, a drop of ca. 20% in classification accuracy was observed in the three-language experiment and the four-language experiment, although results were still significantly higher than the baseline in both cases.

Chapter 6 focused on the question of mining distinguishing features of translator style using the same approach as previous chapters, both in parallel translations of the same text and in a corpus of translations of different texts from the same playwright by each of the translators examined. This represented a novel approach towards detecting stylistic characteristics of a translator's writing. The playwright in question was the Norwegian nineteenth century author Henrik Ibsen and the two translators were William Archer and R. Farquharson Sharp. High accuracy ($\geq$ 90%) was obtained using feature sets containing only one feature-type in ten-fold cross validation experiments on the parallel translations. A classifier consisting of document-level feature sets only was trained on the larger corpus of non-parallel translations and tested on the parallel translation set. 80% accuracy was obtained for the task of determining the translator of each of the two parallel translations of the same play, indicating that each translator maintained a distinguishable textual style across all of his translations of the playwright in question. Features included the frequency of contracted forms in English, the use of different verb forms in the translation of stage directions, and metrics such as average sentence length and type-token ratio. Sharp used the word *because* and a number of other common words significantly more than Archer in his parallel translation, these were investigated with reference to the original source coupled with a diachronic English corpus, to determine whether this phenomenon was a marker of the style of a particular translator or had other origins.

Chapter 7 focuses on commonalities over the three experiments, including document-level and ngram features which are found to be distinguishing in more than one experiment, such as the Coleman-Liau index and the ratio of nouns to total words, along with suggestions for future experimentation.

Acknowledgements

I would first and foremost like to thank my supervisor Prof. Carl Vogel for his tireless dedication and support over the course of this thesis, and for all his inspiration and support in the preceeding years also.

I am very grateful to my parents for their undying support, both moral and nutritional, over these past four years and wish to dedicate this thesis to them.

Sincere thanks are also due to the Science Foundation Ireland via the Centre for Next Generation Localisation whose generous funding allowed me to complete my research in a timely manner, unfettered by issues of a financial nature.

Much gratitude must be extended to the Green Corridor cohort of Hector, Liliana, Stephan, Anne, Roman, Oscar, Erwan, Baoli, Alfredo, Martin and Francesca, each of whom contributed in their own unique fashion to the development of a fertile research environment within the CLG group and surrounding groups, where ideas could be exchanged, honed, hammered out and fine-tuned over lively lunches and creative coffee breaks.

Thanks are due to Francesca, Kevin, Nina, Andrea and David, for excellent advice and friendly companionship during the course of the past years, with particular special gratitude due to Dr. Niamh Nic Ghabhann for all of her kind support during the pre-transfer period in particular.

Muito obrigado por *minha gatinha* Dra. Silvia Dantas for all her support and welcome distraction during the ultimate and most stressful year.

An extra special mention must be given to the good Doctors Isemann and Wetterling, without whose Teutonic tutelage and Common Room consulations I would have been sorely unprepared to tackle the many vagaries of *academe*.

Vielen herzlichen Dank meine Herrschaften!

Finally, *mille grazie* to Prof Silvia Bernardini, Dr. Federico Gaspari and Prof. Federico Zanettin for extending such a warm welcome to me during my stay at the Department of Interdisciplinary Studies on Translation, Languages and Cultures at the University of Bologna at Forli, Italy during March and April of 2011 and correspondingly a million thanks to Prof. Saturnino Luz at TCD for organizing the exchange.

”Tradutorre, traditore”

Translator, traitor.

Italian proverb

”El original no es fiel a la traducción.”

The original is unfaithful to the translation.

Jorge Luis Borges

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The research carried out in this thesis can be classified as pertaining to the discipline of *translation stylometry*, a subfield of computational stylometry which synthesises research conducted in the fields of computational linguistics, translation studies, literary stylistics and corpus linguistics. The experiments carried out within investigate stylistic properties of translated text using *supervised learning* methods drawn from the literature on text classification and textual stylometry. The original contribution to the field of knowledge will be the identification of textual features which distinguish different properties of translated text, including differences between translated text and non-translated text in the same genre, the detection of the source language of a translation and also profiling the textual features which distinguish a translator's style.

Translationese, an implied subclass of a language which consists of text translated from another language, is an important framing notion. The use of the term stems from work by Gellerstam (1986) on the interference from Swedish in novels translated from English into Swedish; this concept is explained in more detail in Section 1.1 below. The work focuses on English as a target language in its own empirical foray. Addressing the question necessarily involves explanation and improvement of extant methods for text classification.

The question of different dialects of *translationese* has been considered by Koppel and Ordan (2011) and this notion shall be explored within the thesis, first examining dialects within comparable corpora of translated and original text in two different genres, then proceeding to the analysis of stylistic characteristics of translations in the same genre, but with different source languages. Finally, stylistic characteristics of two literary translators will be examined, both in parallel translations of the same source text, and also in translations of different works by the same author.[1]

1.1 Motivation

Various terms have been used to describe the subset of a target language which consists solely of translations, the most popular of which is the aforementioned term *translationese*. The question remains as to what makes this dialect of a language objectively identifiable, and indeed whether these generalisations hold across translations from a number of source languages. Moreover, it is also of interest to consider aspects of this dialect such as stylistic preferences of one translator compared with another, together with terms and turns-of-phrase which may represent word-for-word or indeed highly close transfer from the source language.

With the rise in availability of machine-readable corpora and the prevalence of machine-learning techniques for text classification and analysis, the practice of mining monolingual translation corpora for features specific to translated language has been given more attention. As *translationese* often has negative connotations refering to the perceived quality of a trans-

[1]This scenario can be summarised as follows: translator A translated texts 1,2 and 3 from author X, and translator B translated texts 4,5 and 6 from author X, but both produced a translation of text 7 from author X.

lation, having a system which analyses and ultimately detects the presence of this linguistic style could be of benefit to translation agencies, reviewers and even individual translators themselves.

Studies on translationese to date have focused on both lexical and grammatical features of the language, the former refering to nouns and other content words and the latter referring to the occurrence of different parts of speech and grammatical structures. For example, one possible grammatical marker of translationese could be described as translations from Romance languages into English containing more instances of the preposition *of* than non-translations, based simply on the fact that the construction *noun of noun* is more common in Romance languages where such a phrase might correspond to a *noun noun*[2] compound in English.

On the other hand, one may imagine that a translation from German concerning German current affairs might refer to structures such as the *constitutional court*[3] more often than an article concerning the current affairs of a country such as the US or the United Kingdom, however this term will be also used in original-language news reports about Germany and is therefore not a robust marker for translationese. If the translator had decided to translate the notion of a *constitutional court* in a less literal sense which would be identifiable to UK or US readers (e.g. Supreme Court[4]), this string would not count as translationese as it occurs in contexts referring to judgments in the US also. Another example would be the word *prefecture* which is used often in referring to regional areas in Japan and China but seldom in other contexts.

In translation studies there is much interest in the concept of the *visibility* or *transparency* of translators and this issue should also be mentioned in relation to *translationese*. Venuti (1995) deals with this topic and is concerned with the prevailing nature for translations in the Anglo-Saxon publishing world for to be acceptable only if they read fluently in English, i.e. are indistinguishable from an original text. Using the example of *constitutional court* above, this particular phrase would possibly be unfamiliar to readers in the UK or the US and thus may be replaced with the equivalent local judicial body, be it *Supreme Court* or *High Court* or otherwise. Thus, the question of identifying *translationese* can well be interpreted in another fashion with regard to the visibility of the translator of a text, indeed Venuti himself regards the term *translationese* and variations such as *translatorese* as:

> ...perjorative neologisms designed to criticize translations that lack fluency, but also used, more generally to signify badly written prose... (Venuti, 1995, p.4)

When comparing open-class with closed-class words, it is closed-class words which are

[2]One example could be *carnet de identidad* in Spanish which is equivalent to *identity card* in English but could be translated as *card of identity* by a less experienced translator.

[3]from the German *Verfassungsgericht*.

[4]Given that Germany and the United States have different legal systems and processes, a translation like this would not normally be done, this is simply used to illustrate the notion of *translationese* as the dialect of a language which occurs within translations.

usually identified as more robust indicators of the provenance (translated or original) of a text in the literature, examples of this are taken from work by Baroni and Bernardini (2006) who found that when inspecting their classifiers used in experiments on differentiating translations from non-translations in Italian, the frequency of pronouns in translated text proved to be a distinguishing feature, which leads the authors to conclude that as Italian is a *pro-drop*[5] language, translators tend to over-represent the personal pronouns in translations from languages which do not share this property. Other studies such as work by Kurokawa, Goutte, and Isabelle (2009) also provide evidence for this phenomenon in their study on French and English translated language, finding that the English text translated from French contained more prepositions than the original English text, reflecting the greater proportion of this particular part-of-speech in the source language.

The question of individual translator style is also a concern of translation studies scholars and it is a fertile area for research using machine-learning tools. This research seeks to investigate stylometric patterns of a translator's style, both on a document-level and on an ngram level, and also compare these to their original writing in their own native tongue. This assumes the availability of original texts in comparable genres by the translators in question, unfortunately it is not always the case that such a corpus will be readily available. One feature which is not investigated in the experiments is any form of parsed representation of a text which contains more syntactic information than a part-of-speech tag.

Baker (2000) identifies stylistic variation in the use of certain verb forms in the translations of two literary translators translating from different languages, but suggests that the source language may play a role in determining the frequency of these features in the translations.

An attempt is made to reduce the confounding variables at play in this type of study by carrying out experiments on parallel translations of the same work from the same source language by different translators. These experiments focus on two different types of textual features, ngrams of words and POS and document statistics[6] and these allow for the coverage of a wide range of phenomena in the corpora examined in this thesis. Although machine-learning methods have been employed in tasks investigating translated language (Baroni & Bernardini, 2006; Koppel & Ordan, 2011; Ilisei, Inkpen, Corpas Pastor, & Mitkov, 2010; Ilisei & Inkpen, 2011), and detecting the source language of a literary translation (van Halteren, 2008), there has been little work on applying these methods to investigating the stylistic choices of a literary translator, an experiment which is carried out in Chapter 6.

However, an in-depth qualitative analysis is not performed on the target texts[7] and in general the source text is not engaged with to this extent either, due to a lack of sufficient

[5]In languages such as Spanish and Italian, the personal pronoun is often omitted, as it is effectively redundant in most contexts, the person information being conveyed by morphological variations to the verb stem. *(yo) no soy marinero* I am not a sailor vs *(tu) no eres marinero* you are not a sailor.

[6]Features which give an overview of some property of a section of text, for example the ratio of unique token types to the total number of tokens, also known as the *type-token ratio*.

[7]Some examples of this would be translation of metaphor or cultural issues regarding translation of certain taboo terms and themes.

command of the source languages in question. A number of attempts are made to determine the origins of possible source language effects and translator style markers where possible.

1.2 Research Question

The research question which will be investigated in this thesis can be summarised as follows:

> How does *translationese* manifest itself in English as a target language, across a variety of genres and source languages?

This question is then broken down into a number of sub-questions, which are addressed in separate chapters:

1. How does translated text distinguish itself from original text across two textual genres in English ?
2. What features distinguish the source language of a literary translation?
3. How does translator style manifest itself in parallel translations of the same literary text?

1.3 Structure

Chapter 2 of this thesis details the state-of-the-art of research into the stylistic properties of translated text, separating work on translated language in the field of translation studies and computational linguistics, with details of related work in computational stylometry on a number of different topics that may provide relevant methodological suggestions. Chapter 3 describes the metrics and software used in the analyses.

The three core chapters of the thesis, Chapters 4, 5 and 6, present the experimental results pertaining to the three sub-questions in Section 1.2 above. Each chapter will describe a set of experiments carried out on a particular comparable corpus which aims to answer one of the three questions which are framed in Section 1.2.

Chapter 4 details experimental results from two comparable corpora, a subsection of the English language Europarl corpus of parliamentary proceedings[8] and a corpus of translated and original articles assembled from the New York Times online edition.

Chapter 5 describes experiments on a corpus of literary translations in English which seek to identify textual features that reveal the source language of the translation under examination.

Chapter 6 investigates stylistic features which can distinguish between two parallel translations by different translators from the same author, and proceeds to verify how these features generalise to other translations of the same author by the translators in question.

[8] See (Koehn, 2005) for details.

Chapter 7 collates the results of the three experimental chapters and provides an overview of the results, drawing some overall conclusions and conducting further investigation into a number of textual features which have been identified as discriminatory across the three experiments. A number of proposals for future research directions are also provided.

Chapter 2

Literature Review

2.1 Introduction

The topic of *translationese* has been the subject of a number of studies in recent years, both in the fields of translation studies and corpus linguistics and also in computational linguistics. This chapter gives an overview of a number of key studies on the topic of translation effects in natural language text. Section 2.2 focuses on aspects of translation studies including so called *universals* of translation, Section 2.3 describes work in computational linguistics with a focus on translated texts, and Section 2.4 describes work in computational linguistics which does not examine translated text but can provide methodological frameworks for the studies of textual anomalies in general. Section 2.5 focuses on studies concerning the detection of features which identify the source language of a translation, along with similar work on L1 detection from text by non-native speakers of a language. Section 2.6 reviews prior work in translation studies, corpus linguistics and computational network analysis which focus on the identification of stylistic traits of a literary translator. Section 2.7 summarizes the studies examined in this chapter and motivates the research design for the experiments carried out in Chapters 4, 5 and 6, based on trends and gaps in the literature.

2.2 Work on *translationese* in translation studies

2.2.1 Introduction

This section describes some recent studies of translationese in translation studies. This is not an exhaustive list of the translation studies literature, but seeks to highlight some of the more recent studies which are pertinent to this current research project, studies which often employ statistical and/or computational techniques to some degree in their analyses. It is important to note that these studies focus on several different languages, however the target language of focus in the thesis is exclusively English.

2.2.2 Translation universals

Baker et al. (1993) sets out a framework for the use of methodology from corpus linguistics in translation studies. The discipline had generally focused on small-scale qualitative investigations of individual translations and translators in the years prior to this. Baker establishes the theory of *translation universals*, a topic of much interest in translation theory which provides interesting fodder for stylistic analyses.

These universals, *simplification*, *explicitation*, *convergence* and *levelling out* represent features of translated text which distinguish it from non-translated or *original* text.[1]

The simplification universal can be manifested in a number of features such as lower type/token ratio and shorter sentence length for translations implying that translations are less lexically rich than original texts.

[1]Throughout this thesis, the term *original* text is used to represent text which is not a translation.

Explicitation defines translations as more *explicit* than original text, arising from the assumption that all translations are longer than originals, proposing that translations are on average longer than comparable original text, and that they may use more discourse markers to elaborate certain subjects than original text.

Convergence is summarised by Pastor, Mitkov, Afzal, and Pekar (2008, p.1) as the notion that translated text should be more similar to other translated text than comparable original text, i.e. translations should have similar type-token ratios and readability scores, compared to original text in the same genre. *Levelling out* refers to the overuse of certain common features of the target language, refered to by Puurtinen (2003) as *normalization*. Pym, Shlesinger, and Simeoni (2008, p.319) critique this universal based on the fact that it is based on theories to explain trends in interpreting, not translations.

2.2.3 Corpus work on universals

Work by Laviosa-Braithwaite (1997) and Laviosa (1998) both deal with translation universals in text translated into English and are highly relevant to this current study for that particular reason. Laviosa's work is based on a comparable corpus compiled at the Centre for Translation & Intercultural Studies at the University of Manchester, the English Comparable Corpus (ECC)[2]

The earlier study focused on the newspaper section of the corpus which contained text from the British newspapers *The Guardian* and *The European*. The analysis was conducted using simple statistical tools which provide information such as type/token ratio, sentence length and word frequencies about a text. The translated sections of the corpus were found to use fewer open-class words compared with closed-class words and also had a lower average sentence length than the non-translated text and this was found to be the case independent of the source language of the translation. Other results of note from this study were that the translated texts contained higher frequencies of the present tenses of *to be* and *to have* than their source language counterparts.

The later study focused on the narrative prose genre from the ECC corpus. 14 narrative works in translation were selected, representing approx 1 million words. 18 texts were selected from the fiction section of the BNC from the timespan 1985 -1993 to correspond closely with the 1983-1994 timespan for the translations. The vast majority of each side of the corpus consisted of fictional works with a small proportion (7.22% for the translations, 17.5% for the non-translations) being made up of biographies. The BNC subcorpus contained approximately 700,000 words. Analysis showed that unlike the newspaper texts, the literary translations had on average longer sentences than their non-translated counterparts, however the translations were significantly less lexically dense than the non-translations. The tendency towards the verbs *to have* and *to be* was not significantly different in the literary texts as it had been in the newpaper corpus. Laviosa does uncover some correlations

[2]Further information about this resources is available at `http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm`, last verified May 7, 2010.

between the two corpora relative to the translated vs original text in both, translated texts in both corpora have a lower percentage of content words vs grammatical words, the proportion of higher frequency words vs lower frequency words is indeed higher in translated texts in both corpora and what she refers to as the *list head*[3] is proportionally larger in translated texts across both corpora.

Olohan (2001) examines the same corpora as Laviosa and attempts to identify patterns of *optional* usage in translations and non-translations. This is with regard to the translation universal of *explicitation* which proposes that translations are more explicit than non-translations, one manifestation of this is in the frequency of optional forms. One such form of interest is known to linguists as *complementizer that*, a simple example being the following, *He said that he was going in to town* vs. *He said he was going into town*. Both sentences are acceptable and are semantically identical. [4] Using a concordancer, Olohan counts the number of occurrences of the verbs *say* and *tell* in a number of different forms[5]. In all of the cases for *say*, the omission of *that* was more likely in the BNC (non-translated) section of the corpus than the TEC (translated) section of the corpus. For the verb *tell*, the results were comparable, with Olohan reporting that these results had statistical significance. Olohan describes further work on a number of different verbs such as *admit*, *claim*, *think* and *believe* which find similar results. The phrase *in order* occurs more frequently in the translated corpus, in such contexts as *in order to*, *in order for*, etc. Further results showed that contracted forms in English such as *where's*, *what's*, and *I'm* appear much more frequently in the BNC than in the TEC corpus.

Olohan (2008) continues work on this topic, this time investigating the frequency of contracted forms such as *I'll*, *she'll* and *who'd* in the same corpora, finding that the contractions are more frequent in the BNC than in the TEC corpus.

2.2.4 Grammatical features of translationese

Santos (1995) examines grammatical features with greater depth in her work on Portuguese. She presents several maxims of grammatical translationese, which can be summed up schematically as follows.

- A $\rightarrow$ B,C,D: Cases where one grammatical marker in the source language maps to several in the target language, as in the case of lexical selection, one tends to be favoured over the others

- A + obligatory B $\rightarrow$ A + optional B: When the source language contains a mandatory marker corresponding to an optional one in the target language, the frequency of the optional marker is generally higher in translated text.

[3]The list of the 100 most common words in the corpus, thus the most frequent 100 words in the translated texts account for more tokens than in the original

[4]Optional usage also has applications in the field of steganography which involves coding secret information in text, see Murphy and Vogel (2007).

[5]saying, said, says, tell, told, telling

- vague(A,B) → A ,B: When the source language is vague about the use of two constructions and this vagueness can not be preserved in the target language, there is a degree of translation mismatch.
- compact(A,B) → A + B, A,B: When two concepts are expressed in a compact manner in the source language, often only one of the concepts is expressed in the translation in the target language

Santos uses various examples from English and Portuguese to illustrate these rules. She concludes with a statement:

> One can not take it for granted a priori that translated text is a good representative of the target language
>
> (Santos, 1995, p.7)

She then follows this up with conclusions about the effect that language relatedness has on translationese:

> However, an apparently paradoxical property should be mentioned: If translationese stems from the fact that different languages have different systems, it is also related to language closeness: the closer the languages the larger the quantity of false friends and cognates, both in lexicon and in grammar. The closer the languages the easier to translate the surface and not the content, and therefore the more possible to 'level' the two languages, i.e., even out their differences.
>
> (Santos, 1995, p.8)

This poses interesting questions for the investigation of translationese between languages that are closely related, i.e is it the case that the closer a source language is to a target language, the more likely *translationese* is to occur, or is the converse in fact the case?

2.2.5 Translationese in Finnish Children's Literature

Puurtinen (2003) details her work on translationese in another specific domain, namely children's literature translated into Finnish. She observes a number of interesting artifacts, including the higher proportion of certain connectives in literature translated from English and the higher proportion of non-finite connectives in translations into Finnish. Puurtinen uses computational linguistic tools to access and count constructions in parallel and comparable corpora however all reasoning and comparison is done manually.

Puurtinen is particularly concerned with the universals of *simplification*, *explicitation* and *normalization*. For *simplification* she gives the following description.

> The universal referred to as simplification means that the language of translations is assumed to be lexically and syntactically simpler than that of non-translated target language texts.
>
> (Puurtinen, 2003, p.4)

Examples of this can include a lower type-token ratio for translated texts. *Explicitation* is explained as follows:

> The explicitation hypothesis suggests that translations tend to be more explicit than target language originals or source texts. Translators may tend to repeat redundant grammatical items, such as prepositions, and overuse lexical repetition, which in turn results in a lower frequency of pronouns
>
> (Puurtinen, 2003, p.4)

Finally, *normalization* is described as:

> the exaggeration of typical features of the target language.
>
> (Puurtinen, 2003, p.4)

With regard to *normalization*, she explains that translations are normally found to be more conservative than non-translations but this can often depend on the status of translations in a particular literary environment. She theorizes that in fact some styles of poetry, such as nonsense poetry in the domain of Finnish children's literature were actually introduced into the canon in the target language from translations of such styles in the source language, styles which never existed previously in the target language. Puurtinen concedes that translations tend to be relatively highly regarded in the Finnish literary world and this may not be the case in other domains. She also believes that the degree of normalization is often related to the level of prestige of the source text, with classical literature not being normalized to the same extent as an instruction manual.

This prestige factor poses some interesting questions for translationese and the focus of this current study on cross-genre corpora may shed some light on this topic at least for the English language.

2.2.6 Conclusion

It is important to take the results of these studies from the translation studies literature into account in any future work which investigates translationese in English which employs methods from machine learning and natural language processing. These tools can be used in an attempt to verify or refute translation universals or other theories of translation studies by performing a detailed analysis of the textual features of translation corpora.

Furthermore, the results described above appear to indicate that some features are higher-order in that they are not simply concerned with the frequency or presence of individual items but rather with the distributions of families of items.

The next section describes studies in computational linguistics, often in conjunction with translation studies scholars, which focus on using computational methods from text classification and related fields to identify features which distinguish translated text from original text.

2.3 Prior work on translated language in computational linguistics

2.3.1 Introduction

This section describes studies which have greatly shaped the direction taken in the thesis, in the respect that they combine methods from text classification and machine learning in order to investigate linguistic and stylistic phenomena in translated text.

Section 2.3.2 describes an early study adopting NLP tools for the study of translationese in translated English, Section 2.3.3 gives an account of a landmark use of machine-learning methods towards the detection of translationese in Italian, with Section 2.3.4 summarizing studies on translationese which focus on the application area of machine translation. Section 2.3.5 describes studies on Spanish translationese which employ mainly document-level features such as average sentence length and type-token ratio.

2.3.2 POS distribution in translated English

Borin and Pruetz (2001) use a part of speech tagger to investigate distinguishing tokens of POS and words in comparable corpora of English newspaper text. As comparable text they use the reportage section of the Frown and Flob corpora, which are updated versions of the Brown and London-Oslo-Bergen corpus which have been augmented at the University of Freiburg. In the translated section, they use the English version of the *Invantrartidningen* publication, which is a multi-lingual[6] newspaper for immigrants to Sweden. They also take the Swedish source for reference purposes. The size of the corpora they investigate are relatively small, having an average of one hundred thousand tokens per section, translated and original.

Based on the distribution of POS ngrams in the translated corpus, a number of trends emerge. In the translated *Invantrartigningen* corpus, there are a higher frequency of sentence-initial prepositions and adverbial clauses, phenomena which are common in the Swedish language but less common, although still acceptable, in English.

Another phenomenon they observe is the relatively higher frequency of verb initial sentences in the corpus of translated English. They note that the translated English may contain more examples of readers' letters than in the non-translated corpus and this could be in fact responsible for the frequency of this construction, as the readers' letters generally contained a relatively high frequency of questions which were expressed in this particular grammatical form.

This work is important in the content of the thesis studies, with respect to the methodology used and as an early example of a combination of methods from natural language processing, corpus linguistics and translation studies. The source language effects in the English translations from Swedish are interesting to note in relation to work in Chapter 6 which

[6] 8 language versions, including English

examines translations from Norwegian, a language from the same family which shares some grammatical traits.

2.3.3 Translationese in Italian

One of the most widely cited studies in this area is Baroni and Bernardini (2006), which attempts to use machine-learning techniques to separate translations from non-translations in a comparable corpus of articles from the Italian-language geopolitical journal *Limes*. This work is notable also for the accompanying human experiment which attempts to measure how well humans can distinguish translations from non-translations.

Regarding the machine-learning aspects of the study, they believe that using a small corpus of text from a highly homogeneous source is a better idea for initial experiments than a larger corpus of mixed-genre texts which could contain a number of confounding factors.

Their corpus contains approximately 3 million words, 2 million of which are in original Italian and 900,000 of which are in translated Italian. All proper nouns are replaced with a unique placemarker, this is done to ensure that any results are based on robust textual features and not simply the mention of a personality or place which may divulge the nature of the text. The following example illustrates this

> Bill Clinton said that he (Bill Clinton) is too old to be nominated to the Supreme Court

becomes

> NPR1 said that he (NPR2) is too old to be nominated to the NPR3

They employ a number of different tokenizations to represent the documents in the SVM[7] classification, using unigrams, bigrams and trigrams of both original wordforms, a lemmatized representation and an unusual mixed representation where content rich words are replaced with lemmas and function words are left in their original form. They also build both weighted and unweighted feature vectors, weighted vectors using the TFIDF[8]representation and unweighted vectors, discarding the features that appear in over half of the texts in the experiment. These representations are designed to obtain robust markers from translationese that are based on close-class words and frequent parts-of-speech.

Classifiers are combined in both *majority voting* and *recall maximization* ensembles, the former relying on the majority of classifiers to label a document as a translation and the latter labelling a document as a translation if at least one classifier says so. The latter ensemble is used due to the large skew in the size of the untranslated class. 24 unique

[7]Support Vector Machines: A classification method which seeks to create a separating hyperplane between two classes,where documents are represented as vectors of their features (either binary or relative frequency counts), used often in text classification tasks, see Section 3.3.2.

[8]Term Frequency Over Independent Document Frequency, feature weighting technique which takes the frequency of a word in a corpus in conjunction with the frequency of the word in an individual document into account, see Aizawa (2003).

classifiers were created out of the different representations, each having a unigram, bigram and trigram section, with each section being relatively different from the others, in other words, representations tended to mix the features, e.g. word unigrams, POS bigrams and mixed trigrams, avoiding repetition of features in different sections. As regards trigrams, only POS and mixed selections were used due to data sparseness issues with word-based trigram models. Single classifiers containing only one representation were also used.

They perform sixteen-fold cross validation[9] on sixteen sections with 15 translations and 15 non-translations per section. Results for single classifiers provide a highest score of 77% in the binary decision of category assignment accuracy for word unigrams. The mixed classifiers using majority voting did not surpass this by much, the best combination including word unigrams, bigram mixed and lemma and trigram POS tags achieved also 77.5% accuracy with higher precision and lower recall than word unigrams alone.

The most surprising results turned out to be the experiments which used mixed representations and recall maximization. The worst of such classifiers still gave accuracy, precision and recall of over 80 % while the best combination[10] almost managed 87% accuracy with almost 90% recall. Removing classifiers based on content words from the mix caused the results to dip less sharply than when ones based on mixed POS/lemma representations were removed, a result which leads Baroni and Bernardini to surmise that syntactic and function word patterns are more important than lexical items.

Further research into the actual linguistic cues based on results from Puurtinen (2003) and Borin and Pruetz (2001) concludes that clitic pronouns and adverbs aid the detection of translated Italian by SVMs, based on performance decreases when these features are removed from the analysis. This work pioneered the usage of machine-learning classifiers for the task of translationese detection and as a result inspired the work carried out in this thesis to a large extent, from the classification algorithms used to the type of features employed.

2.3.4 Translationese and applications for MT

Motivation

In recent times a number of researchers have investigated translation direction in a machine translation context and found that using corpora translated in the same direction as one wishes to translate results in better or comparable results using less data than if translation direction is not taken into account. This section describes a number of these studies and the methodology they employ.

[9] *n-fold cross validation* is a text classification evaluation technique whereby the dataset is divided into *n* folds, usually ten or more, and for each iteration of the experiments, the dataset is divided into *n -1* times training and 1 test set, this is done *n* times and all results are averaged across the *n* folds.

[10] unigram lemmas with tfidf weighting, unigram mixed representation with tfidf weighting, bigram lemmas, bigram mixed representation lemmas, and trigram pos

Detecting *translationese* in the Canadian Hansard Corpus

Kurokawa et al. (2009) present results for detecting translated text in a bilingual French-English corpus of Canadian parliamentary proceedings. They use Support Vector Machines trained on either the French or English text separately or both. They obtain an accuracy of up to 90% for detecting translations. A novel part of their study involves using the source-target language information to train MT systems depending on the translation direction. They find that phrase-based SMT systems trained on the right source-target direction perform roughly the same or slightly better than their counterparts trained in the opposite direction[11] but use five times less training data.

They use a large bilingual Canadian Hansard corpus for their experiments, containing a total of approx 80 million words, roughly 30 times larger than the corpus used by Baroni and Bernardini (2006). There is an imbalance of 4:1 in terms of English original data versus French original data. Preprocessing consists of converting all text to lowercase and then running a POS-tagger over the data, producing four different versions of the corpus based on the example of Baroni and Bernardini (2006), with word, lemma, POS and mixed ngrams. Their experiments had four different main parameters: size of ngrams, representation as in the previous sentence, English source or French source and whether TFIDF was used. The best representation for detecting translations using the English side of the corpus is word bigrams, which results in an F-score of 90% on text blocks, which contain a number of sentences. They also ran experiments on a sentence level data structure where word bigrams also provided the best results with a 77% F-score. In general, using larger ngram representations decreased accuracy, this result was found to be independent from the representations used.

An examination of the actual bigrams of features in each side of the corpus reveals some interesting patterns, in the English section of the data, the original English contained a large proportion of references to political parties in the English speaking part of Canada. This result suggests that obfuscation of proper nouns and content words may be an important preprocessing step if the results are to be acceptable as strong indicators of translationese. In the experiments carried out in Chapters 4, 5 and 6, proper noun features are removed from the classifiers manually.

The section in English translated from French contains more bigrams with definite articles and propositions, features which the authors purport to be pure *translationese*, given that French text would generally contain more definite articles and prepositions than English text, a trait which is carried over into the source language by the translators.

Further work concerns an experiment on machine translation, using SVM prediction to predict which model should be used to translate what kind of data. Using the SVM to predict the model to use resulted in an improvement on standard phrase-based SMT practice which was to use the entire parallel corpus to train the model regardless of the source or target language in each case.

[11] FR-EN for translating from English to French for example

They report that the 0.6 improvement in BLEU[12] score would not necessarily make a practical difference in the quality of the translation but a more interesting result is the fact that for the case of data trained on the English original or French original subset of the data, the performance is virtually identical or in most cases slightly higher than the model trained on all the data for the corresponding translation direction. This indicates that training on the right kind of data can yield improvements in performance.

Kurokawa et al. (2009) acknowledge that their results may be due to a combination of different factors, admitting possible influence from the topic of the texts, the most distinguishing features were on the one hand lexical cues for the English original text, but on the other hand the French original text translated to English showed a high occurrence of bigrams of function words which distinguished it from its counterpart category of original English text. They mention that future work will examine different corpora such as the English-French subset of Europarl and possibly investigate monolingual corpora translated from different languages.

Modifying a language model for SMT based on translation direction

Lembersky, Ordan, and Wintner (2011) carry out similar experiments but focus on the constituent text of the *language model* in machine translation, this is the reference corpus which is used to rank the machine-translated candidate sentences produced by a statistical machine translation system. They compile separate language models composed of translated text from a particular source language into English and use this language model in the task of machine translation from that source language into English.

They measure the *fitness* of a particular language model to a test set using the *perplexity* metric which is described in Equation 2.3.4 below, where L is a language model., W a test set and N the number of words:

$$PP(L,W) = n\sqrt{\sum_{i=1}^{N} \frac{1}{P_L(w_i|w_1.....w_{i-1})}} \tag{2.1}$$

They use Europarl in their experiments, notably an English sub-section of the corpus containing translations from four source languages, German, Dutch, Italian and French, along with original English text. They also examine the Canadian Hansard corpus used by Kurokawa et al. (2009) in their work. Finally, they run experiments on their own corpus of English and Hebrew compiled from the *International Herald Tribune* and *Haaretz* newspapers.

In the Europarl experiments, they create six different language models for each source language, one mixed language model containing sentences randomly selected from each of the four source-language subcorpora plus the original English, one language model trained

[12]BLEU is an ngram based method of machine translation evaluation, sentences produced by an MT system are compared against a number of gold standard texts, see Papineni, Roukos, Ward, and Zhu (2002).

from the original English portion and one for each of the source languages in the Europarl subcorpus. They then extract approx 100,000 reference sentences for each source language to English pair[13] for use in their experiments.

Using the perplexity measure on these reference sets, in all cases the language model made up of English translated from the source language of the reference pair gave the best perplexity score, followed by the mixed language model. The original English language model gave the worst perplexity score in all cases. They cite the fact that language models made up of translated language from different sources were still a better fit than original English as clear evidence for the existence of translationese as a separate entity. This could be to some extent accepted as experimental validation of the *convergence* universal.

They go even further to validate this point, focusing on the German-English sub-corpus and removing named entities and standardising punctuation to prevent any bias from items of this nature. They create four abstracted version of the German-English language models, one with punctuation standardized, one with named entities removed, one with nouns abstracted and one where all words are represented by their part-of-speech tags. They also mention that an original English language model would need to be ten times the size of one translated from German to achieve the same results.

Their final experiment looks at MT performance using their language models and the results also align well with the hypotheses, the best BLEU scores were obtained by using the LM trained on the English translated from the source language, with the mixed representations being next in line, followed by the LM's from original English only.

They also conjecture than LMs translated from languages similar to the source language in question may be better than those where the source language is not closely related, in some of the experiments on the Dutch subcorpus, the LM made up of English translated from German performed better than the LM made up of English translated from French, for example.

Modifying phrase tables based on similarity to translationese

Lembersky, Ordan, and Wintner (2012) extend this work to focus instead on the internal phrase tables in a statistical machine translation system, as opposed to the language model as examined in the previous study. The phrase tables contain aligned phrases which have been learned from a parallel corpus of translations. In this foray they focus on the Canadian Hansard corpus as used in Kurokawa et al. (2009) and their own previous works. They successfully replicate the results of Kurokawa et al. (2009) and provide a more detailed explanation of why these results occurred.

Their hypothesis is that the phrase tables trained in the correct translation direction contain more unique source phrases and less translations per source phrase than a phrase table trained on text in the opposite direction or a mixed training corpus. They quantify this using two further metrics, the first being the *entropy* of a phrase table:

[13] Including two FR-EN corpora, one from the Hansard and the other from Europarl

> Given a source phrase s and a phrase table T with translations t of s whose probabilities are $p(t|s)$, the entropy H of s is:

(Lembersky et al., 2012, p.4)

$$H(s) = -\sum_{t \in T} p(t|s) \times log_2 p(t|s) \tag{2.2}$$

the second being *cross entropy (*CE), which is defined as for a text $T = w_1, w_2 ..... w_n$ and a language L:

$$CE(T,L) = -\frac{1}{N} \sum_{i=1}^{N} log_2 L(w_i) \tag{2.3}$$

Their hypotheses are upheld on the Hansard data, the $S \rightarrow T$ phrase tables have lower cross entropy and entropy than the the mixed tables. The former set of tables result in the same BLEU scores as original-only tables using a tenth of the data, while mixed tables do still provide a small gain in performance.

They focus on adapting the phrase tables using phrases from both types of tables, by defining a metric which measures how close each phrase pair is to a model of *translationese*. They build this metric into the decoder and also use cross-entropy as a tuning feature, running two separate experiments with the same training data, a mixed set of sentences from the $S \rightarrow T$ and $T \rightarrow S$ sections of the corpus. The systems with the cross-entropy augmentation and the perplexity ratio augmentation result in an increase in BLEU scores over the baseline.

They also perform a qualitative analysis on a number of sentences from a $S \rightarrow T$ based system and a baseline mixed system, finding that the quality of the system trained on the same direction tended to be better, translating certain phrases in a more culturally sensitive manner.[14]

Distinguishing machine translated text from human translated text

Related work by Carter and Inkpen (2012) focuses on the task of distinguishing machine-translated text from human-translated text. As a corpus, they also use the Canadian Hansard, similar to (Kurokawa et al., 2009; Lembersky et al., 2011, 2012), along with parallel data from Canadian Federal Government websites and website of the Government of Ontario. They focus on the task of filtering raw machine-translated text from language models which are generated from web data, a growing problem for training statistical machine translation systems. Poor quality machine-translated text in a language model can adversely affect any machine translation system which uses this language model, and this is something to be avoided when training such a system. The task is comparable to classifying translated and original text and they use a similar approach to Baroni and Bernardini (2006) and Kurokawa

[14] An example of this was the translation of times from the French 24 hour system to the English 12 hour representation.

et al. (2009), training a supervised learning system on large corpora of translated French and machine-translated French from the same source and translated English together with a machine translated version from the same source.

They posit two hypotheses based on work by Carpuat (2009) who found that machine translations from SMT systems tended to have more lexical consistency than human translations and also often exhibited strange terminological lexical choice errors.[15] They believe that these factors should enable automatic idenfication of machine-translated text.

They use Microsoft's Bing translator system to carry out the machine translation in both directions, citing usage limits and also translation quality[16] as reason for this choice. They used a SVM classifier with unigrams, type-token ratio and unigram length as features, which they trained on the individual corpora, Hansard, Government of Canada and Government of Ontario corpora. The Hansard is considered clean, i.e. it should not contain any machine-translated text, a property which cannot be confirmed for the other two corpora. The corpora are large, 949 documents for training and 58 for test for the Hansard corpus, the Government of Canada corpus contained approx. 20,000 documents but around the same amount of text as the Hansard corpus, approx 230 megabytes. The Government of Ontario corpus was kept as an out-of-domain test set for the models trained on the Government of Canada corpus, also containing a comparable amount (204 megabytes) of text.

Classification results on the Hansard corpus were high, with 99.89% from 10-fold cross validation on the training set, and 100% on the test set. Classification results on the Government of Canada corpus using 10-fold cross validation were also high, averaging 98% for the four class problem, translated English vs. machine-translated English vs. machine-translated French vs. human-translated French. They mention that the features they select are "common" features, although they fail to present examples in their paper.

However, when they used their models trained on the Government of Ontario corpus, the results were not as successful in detecting machine-translated text. They ran the experiment on a test set consisting of human translated text only, and the system classified a significant proportion of this text as machine-translated, leading the authors to conclude that their model did not generalise well on unseen or out-of-domain data.

2.3.5 Translationese in Spanish medical and technical translations

Pastor et al

Pastor et al. (2008) investigated corpora of Peninsular Spanish divided into three subcategories, medical translations by professionals, with a comparable set of original texts, medical translations by students of translation and a comparable set together with technical translations by professionals together with a comparable set of originals. All translations had

[15] One example given in (Carpuat, 2009) was *organic daughter*, from the French *fille biologique*, which should have been translated as *biological daughter*.

[16] Interestingly, they wanted a lower quality MT system as to represent better the type of machine translated text which might be found on government websites from the past number of years.

US or British English as the source language, were translated between the years 2005 and 2008 and contained between 1-2 million tokens for each corpus subdivision[17].

An interesting feature in their work is the decision not to use mostly ngram based metrics as per Baroni and Bernardini (2006), Kurokawa et al. (2009) and van Halteren (2008), preferring instead to use features such as the proportion of grammatical words in texts and the proportion of grammatical words to lexical words similar to the work in translation studies (Laviosa-Braithwaite, 1997; Laviosa, 1998).

Their work draws further on theories of *translation universals*, in particular the universals of *convergence* and *simplification* with the latter characterised by average sentence length and lexical density measures including readability scores and the former typified by a higher frequency of shared POS ngrams in a set of translations than in a set of comparable originals[18]. The features used provide a basis for the feature types used in this thesis, such as average sentence length, type-token ratio, various readability scores and POS ngrams. The methodology used in Pastor et al. (2008) reflects the trends in corpus linguistics and translation studies: formulate hypotheses based on previous results from the literature, process feature sets and then use a t-test to confirm or reject the existence of significant differences between the average frequencies of the various feature types or POS ngrams.

Breaking down results over the three sub-categories, they find that the corpus of technical translations conforms to their outlined hypotheses to the highest degree, translations have a lower lexical density, lower average sentence length and lower readability score than their comparable originals, although in the case of sentence length they hypothesised that the opposite would be in fact the case, although this is indeed of interest when compared to results in Olohan (2001) which found newspaper *translationese* to have a lower average sentence length than comparable original text but literary *translationese* to have a higher average sentence length, which may indeed suggest that this feature is genre-specific, although taking into account the two different target languages in these studies. The two corpora of medical translations contain less significant differences, with the least divergence found in the corpus of medical literature translated by student translators, for which a statistically significant difference was only found for one readability metric[19], although a number of divergences in discourse marker ratio, average sentence length and proportion of multi-clause sentences to sentences with one clause only were identified between the original and comparable sections of the corpus of professionally translated medical texts.

Ilisei et al

Recent work by Ilisei et al. (2010) examines translationese in Spanish texts using machine learning methods. This work is an expansion and continuation of experiments detailed in Pastor et al. (2008).

[17] i.e. divided into comparable and translation sub-sections, so two for each sub-category.

[18] This is examined by grouping the six sub-divisions into two larger corpora of translated and original text and investigating the degree of internal homogeneity

[19] the Coleman-Liau Index, see Chapter 3, Section 3.5.2.

They concern themselves also with the universal of *simplification*[20] and propose more textual features to capture this phenomenon such as sentence length, parse tree depth, proportion of simple and complex sentences, word length as the proportion of syllables per word, lexical richness and the proportion of lexical words to tokens. The classifiers they use are Jrip, Decision Tree, Naive Bayes, BayesNet, SVM, Simple Logistic and one combination classifier which considers the output of Decision Tree, Jrip and Simple Logistic. These classifiers are standard classification algorithms implemented in the WEKA toolkit. Jrip is a propositional rule learning algorithm and Simple logistic uses the method of simple logistic regression for classification, while Bayesian Networks are represented as an interconnected graph of probabilistic assumptions which affect the outcomes of one another, unlike Naive Bayes classifiers which assume that the co-occurence of variables do not have an effect on their neighbours in a classification task.

The data they examine are in the technical and medical domains and the translations are carried out by both professionals and students (Pastor et al., 2008). The authors consider features such as lexical richness and sentence length indicative of the *simplification* universal. The best results are given by the SVM classifier using the simplification features on the technical dataset, they achieve 97.62% accuracy using the simplification features, however in the medical domain the highest result is 82.35%. The simplification features provide an average of 5% gain over classifying without those features.

The authors analyse the features using Information Gain and χ^2 to find the features which account best for the classification. The two metrics return comparable results with lexical richness and grammatical vs. lexical word ratio being the two top features in the classification, followed by the ratio of finite verbs, the ratio of numerals, the ratio of adjectives and then sentence length.

The authors state that their features are language independent however they do not carry out research on other languages to see if this is the case, indeed the influence of genre was already seen to play a role in their study which suggests that language may also play a role in which features distinguish translated text from original text.

2.3.6 Translationese in Romanian newspaper text

A further study on translationese in Romanian newspapers by Ilisei and Inkpen (2011) uses almost identical methodology to Ilisei et al. (2010), although this time focusing on a comparable corpus of Romanian newspaper articles. Extra features are used in this study to represent possible language-specific traits of *translationese* in Romanian, including the proportion of interjections, proper nouns and commons nouns, together with values for proportions of verbs in first, second and third person forms together with various moods such as the subjunctive, imperative and infinitive forms. Altogether they use thirty five document-level statistics in their experiment.

[20] See Section 2.2

They obtain high accuracy on distinguishing between classes, 98.6% with an SVM classifier using ten-fold cross validation on the training set, which consists of 416 original texts and 223 translations. As before they obtain results with and without the simplification features, achieving gains in accuracy of ca. 3% when these features are included, again providing evidence for the *simplification* universal this time in translated Romanian.

Ranking features as in the previous work showed that the information load, noun ratio, preposition ratio and lexical richness measures[21] were amongst the best discriminators between translated and original text in Romanian.

2.3.7 Dialects of translationese

Recent work by Koppel and Ordan (2011) examines similar corpora to the work in Chapter 4 and seeks to identify patterns of *translationese* across two relatively different corpora, Europarl and a comparable corpus of texts from the International Herald Tribune.

Their first experiment uses a subset of Europarl with a focus similar to van Halteren (2008), in which they try to guess the source language of English translations with Finnish, German, French, Italian and Spanish as source languages. Using their method of Bayesian regression, they obtain 97% accuracy on this corpus, which consists of ca. 500,000 words per source language, with the English comparable section containing five times this amount of text. One difference between their study and Van Halteren's work, apart from the fact that they only examine one target language, is that they use only the frequency of 300 function words as features.

They proceed to another set of experiments which seeks to investigate the different dialects of translationese based on each source language, by training classifiers on a comparable corpus of translations from one source language only with original text and testing on a test set of translations from a different source language together with original texts. They also perform experiments testing on the translations from the same language as the test set and compare the results.

The results show that training and testing on related languages, for example Italian and Spanish, are better than the results for training on Italian and testing on German. The classifiers trained on Finnish performed poorly on all test data except the Finnish test set. Nevertheless, the worst results of approx. 60% were still better than the baseline, which leads the authors to surmise that certain features of translationese are not highly corrolated with the source language in question. They find pronouns and what they refer to as *cohesive adverbs*[22] to be more frequent in translated Europarl text than in original Europarl text, regardless of the source language.

They investigate the same questions on the International Herald Tribune corpus which consists of translated Greek, Hebrew and Korean text, and is roughly half the size of their Eu-

[21] See Section 3.4 for descriptions of these.

[22] These include tokens such as nevertheless, indeed, furthermore and moreover, refered to in this thesis as *discourse markers*.

roparl corpus. Their best result for source language identification is 86.5% which is weaker than the results on the Europarl corpus but still highly significant, given that the corpus contains a mix of more diverse source languages and is smaller in size. They distinguish translationese from original English with a similar accuracy of 86.3% in the IHT corpus. The final set of experiments involve training on one corpus, i.e. Europarl and testing on the other corpus. Classification accuracy was low for these experiments, training on Europarl and testing on the IHT resulted in 64.8% accuracy while training on the IHT corpus and testing on Europarl resulted in a lower accuracy of 58.8%. They conclude that this provides evidence for different dialects of translationese.

However, their final experiment provided interesting results, they mixed chunks of the IHT corpus with the Europarl corpus, 200 texts from each of the eight source languages with the original side comprising of 1000 texts from Europarl and 600 from the IHT. They obtain 90% accuracy in this classification experiment, which again highlights the fundamental differences between translated and original text in English, regardless of style or genre.

2.3.8 String kernels for translationese detection

Popescu (2011) uses a different approach in experiments on detecting translationese using a literary translations corpus similar to the corpus used in Chapter 5. 214 books were collected, 108 by British and American authors and the other 106 divided between 30 translations from German authors and 76 works in translation by French authors. This corpus was collected from *Project Gutenberg* so due to copyright restrictions stems mostly from the nineteenth century, as with the corpus used in Chapter 5. Unlike the corpus used in Chapter 5, Popescu (2011) used multiple works by the same translators and authors.

The only processing carried out on the texts was the normalization of whitespace, as the string kernel method functions on a character level. The concept of string kernels is introduced in the experiments, in particular the *p-spectrum string kernel*, which measures the similarity of two strings based on the number of substrings of length p that these two strings have in common with one another. For two strings, s and t and an alphabet $\sum$, where $s,t \in \sum^*$, the p-spectrum kernel is defined as:

$$k_p(s,t) = \sum_{v \in \sum^p} num_v(s) num_v(t) \tag{2.4}$$

with $num_v(s)$ representing the number of occurrences of string v as a substring in s.

They use a normalized version of this kernel so that strings of different lengths can be compared:

$$\hat{k}_p(s,t) = \frac{k_p(s,t)}{\sqrt{k_p(s,s)k_p(t,t)}} \tag{2.5}$$

The first experiment sought to detect whether a text was a translation or an original. Initial results using a string kernel of length 5 and SVM classifier with the entire set of 214 texts obtained accuracy of 100% on a ten-fold cross validation test, which the author found

suspicious.

The next experiment trained a classifier on French translations and British originals and tested on German translations and American originals, which gave results of 45.83%, essentially worse than the baseline of approx 50% as the classes were relatively balanced. Examining the features from the first experiment more closely, it was found that character strings of French proper names were the most discriminatory features, resulting in serious overfitting on the training set. To counteract this effect, he collected the source of the French translations and removed any substrings in the target which occurred in the source reference corpus.

Repeating the previous experiment of training on French and British and testing on American and German, he obtained a higher accuracy score than before, at 77.08%, which again is evidence for universal elements of translationese in the target language which are not directly related to the source language.[23] A second experiment used a mix of British and American text in the original side of the training corpus with French and German kept in the training and testing phase as before, resulting in an accuracy of 76.88%, a non-statistically significant difference from the first experiment.

This work presents some interesting new approaches to the task of detecting translationese, however this approach fell foul of the issue of source language proper names as distinguishing features, a precaution which is taken in the experiments in Chapter 5 by manually removing all proper name features in the classifier. Another issue with the approach which the author acknowledges in the paper is that it can be more difficult to make sense of the distinguishing features which emerge, as they are not words but sequences of characters. Work in Chapter 5 will use a similar, albeit smaller corpus, with the task of source language detection in mind, however English original texts are also included in the experiments which allows for examination of features distinguishing original and translated text by treating original English as another source language.

2.3.9 Summary

The studies in Section 2.3 have strongly informed the design on experiments carried out in this thesis. Ngram features (Baroni & Bernardini, 2006; Kurokawa et al., 2009; Koppel & Ordan, 2011) are combined with document-level features (Pastor et al., 2008; Ilisei et al., 2010; Ilisei & Inkpen, 2011) in order to describe fully the comparable corpora which are used in the experiments.

The studies on *translationese* in the domain of machine translation described in Section 2.3.4 are important to note on a number of levels. Firstly, the fact alone that the machine translation community is interested in this problem will raise the profile of the topic of translation stylometry, and this is indeed welcomed. Secondly, researchers in these field will

[23] Although as Lembersky et al. (2011) and Koppel and Ordan (2011) found, the performance for translationese detection using different source language corpora depended on the linguistic closeness of the source languages in question.

bring a large number of technologies to bear on this task, which will benefit future work in computational linguistics and indeed translation studies.

Lembersky et al. (2011) continued work by van Halteren (2008) on Europarl and found that different types or indeed to quote Koppel and Ordan (2011), "dialects" of translationese can be identified in the corpus, these results are promising with regard to the experiments on source language detection from literary text described in Section 2.5. Kurokawa et al. (2009) also identify a higher proportion of prepositions in English translated from French, this result also provides a basis for document-level features for the same task.

Although metrics such as *cross-entropy* and *perplexity* as in Lembersky et al. (2011) and Lembersky et al. (2012) are not used in the experiments carried out in the later chapters, future work on any of the main questions dealt with in Chapters 4, 5 and 6 of this thesis would surely benefit from additional metrics such as these.

2.4 Related work in computational linguistics

This section describes a number of studies in computational linguistics which address similar topics to the detection of translated language. Investigation and examination of the methods listed here may offer alternative approaches for the task of analysing translated language.

2.4.1 Finnish learners of English

Lauttamus, Nerbonne, and Wiersma (2007) use POS trigrams to examine language variation amongst different groups of Finnish immigrants to Australia. They examine the language of two groups, which were compiled at the University of Joensuu, Finland. These were as follows: the *Adults* were all Finnish native speakers born in Finland who were over 18 when they arrived and the *Juveniles* were all Finnish native speaking children of these immigrants who were under the age of 18 when they arrived in Australia. The corpus was made up of transcribed interviews with 62 interviews from the adults and 28 interviews from the children.

The texts are tagged, and 200 most statistically-significant POS trigrams are extracted and then analysed with respect to the literature on language acquisition. The tagset used was one which was constructed by linguists[24], in order to capture more finely-grained categories of part-of-speech. The total number of POS trigrams are then collected with the frequency of each trigram in the two corpora. This vector is then inspected to determine where certain trigrams caused the distributions to skew in statistically significant fashion. Permutation tests are used in the analysis, these are explained in Nerbonne and Wiersma (2006) as follows: measure the difference between two sets with a distance metric, then combine the two sets and extract two random subsets from the combined set, repeating this process and measuring the amount of times the extracted subsets are more different than the original two sets. If the

[24]The TOSCA-ICE tagset was used, see Garside, Leech, McEnery, et al. (1997).

extraction process is carried out a large number of times, it can be then calculated how much the original division of classes differ from a chance division.

A number of statistically significant differences were observed between the *Juvenile* and *Adult* group. The adults were found to exhibit more hesitation(this had been marked up in the transcribed text as false starts, pauses and broken speech). The adults were more likely to use the discourse marker *you know* in their speech, the juveniles used more varied forms such as *you see*, *you mean*, etc. The juvenile group also used phrasal verbs where the adults did not, such as *I ran out of money*. The adults demonstrated misuse of articles, which Lauttamus et al. (2007) mention as characteristic of learners whose L1(Finnish in this case) has no articles, other examples include leaving out prepositions such as *to* with verbs of motion, and deviant word order with regard to adverbials(*I don't watch any more that one*).

The work by Lauttamus et al. (2007) exhibits many parallels with the work on translation universals, in the sense that detailed linguistic second-language acquisition universals were used as features for computational classification. The combination of computational methods and in-depth linguistic analysis is used to full effect in this study. The use of a large number of POS tags combined with transcribed speech also worked well as an experimental construct.

2.4.2 Authorship Profiling

Argamon, Koppel, Pennebaker, and Schler (2009) describe methods for *authorship profiling*. i.e. automatically identifing characteristics of an author from their writing. The different characteristics they mention include *gender*, *age*, *personality* and *native language*. Their method employs text classification with machine learning but they also employ a novel form of taxonomy for function word classification.

An example provided for personal pronouns is displayed in a tree formation, with the node personal pronoun leading off into two categories, *interactant* and *non-interactant*, with *interactant* divided into *singular* and *plural* for example. These are then used to make up the feature sets in the classification with a normalized count of each of the nodes of the tree used in the feature vectors for classification. Content-based features are also used in the analysis. They also consider only the top 1000 discriminating words in the corpus determined using the information gain metric for discriminating between the different classes.

Argamon et al. (2009) are cautious about using content words as markers as these may indeed be context dependent. As their corpus consists of tagged blog posts, it is not surprising that the content words reflect this, words which distinguish between people in their teens and those who are older include *haha*, *school*, and *lol*, which would not be likely found in a more formal text. Nevertheless, combining content words and function words improves accuracy in a number of cases.

Argamon et al. (2009) report promising results for their experiments, on a corpus of 19,320 blog authors, marked for age and gender and normalized for intervals of below 20, between 20 and 30 and 30+. Accuracy for gender is 72 % for style based features, 75.1%

for content based features and 76.1% for a combination of the two. Age, which had three classes, has accuracy of 66.9 % for style, 75.5% for content and a combined accuracy of 77.7% for both.

The next two experiments were carried out on different corpora, The International Corpus of Learner English and a corpus of stream-of-consciousness essays written by undergraduates at the University of Texas for the experiments on personality detection. Native language, with five classes has the highest distinction between style and content with the former giving 65% accuracy and the latter 82.3%, with the combined set giving 79.3%. Personality detection, with the distinction of *neurotic* vs *non-neurotic*[25] proved to be the most difficult task with an accuracy of 65.7% for style features and only 53% for content, which is barely above the baseline of 50%. The combined score in this case was 63.1%. Some stylistic features which were found to be more likely in neurotics were more frequent use of pronouns in subject position in a sentence, tendency to refer to themselves and more frequent use of propositional phrases such as *for* and *in order to*. Non-neurotics tended to be less concrete and use more indefinite terms such as *a* and *a little*. In support of the low classification accuracy, the authors mention results from an unpublished doctoral thesis by S Vasire at the University of Texas, Austin in 2006 which indicate that humans managed only a 69 % classification of neuroticism of acquaintances they had known for several years.

2.4.3 Personality detection

Luyckx and Daelemans (2008) report on similar work on personality detection from text, this time using their hand-built Personae corpus made up of personal undergraduate essays in Dutch on a specific topic, *Artificial Life*. The corpus contains 200,000 words and the authors chose a non-personality related topic in order to minimize the effect of *awareness of personality* which might bias the students' writings. All students sat a Meyers-Briggs Type Indicator test which provides scores on four axes:

1. *Introversion* vs *Extraversion*, quiet and reflective vs outgoing and impulsive,
2. *Intuition* vs *Sensing*, trusting abstract theories vs requiring concrete information
3. *Feeling* and *Thinking*, making decisions emotionally or based on logic and reason
4. *Judging* and *Perceiving*, preference for a structured life vs preference for change.

Syntactic features are extracted from the text and the TiMBL classification suite is used to classify the texts. TiMBL is a *k-nn* -memory-based classifier and was developed at the University of Tilburg[26].

Classification accuracy is not particulary high for the first two scales, 65% for *Introversion* vs *Extraversion* and 62% for *Intuition* vs *Sensing* . The latter two categories fare better

[25] Described simplistically by the authors as 'tendency to worry'.

[26] See Daelemans, Zavrel, van der Sloot, and van den Bosch (2003) for details about the TiMBL classifier.

with 73.79% for *Feeling* and *Thinking* and 82.07% for *Judging* and *Perceiving*. These results are based on four classification tasks where each task had to assign either one of the two labels for each axis. Word trigrams and POS trigrams perform well in the classification on the four scales.

Interesting methodological points to glean from these experiments include the separation of content and style-based markers, both of which prove to be useful as distinguishing variables. In the work on translation stylistics, content words were often avoided, however it may be of interest to examine their role in determining translations from non-translations.

2.5 Detecting the source language of literary translations

2.5.1 Introduction

This section describes work related to the task of detecting the source language of a literary translation. This topic has not been the focus of a large body of research, however this section will also summarise work towards answering an analogous question in textual stylometry:

Is it possible to detect the L1 of a non-native speaker from their L2 writing?

It is argued of course that the task of source language detection from literary translations is a more difficult one, as one is usually presented with texts of a high linguistic quality, which means that mispellings are not features that can be drawn upon in this work, unlike in the case of L1 detection from non-native writing, which often contains such features.

Section 2.5.2 describes research on detecting the source language of Europarl, inspiring work by Lembersky et al. (2011) and Koppel and Ordan (2011) on clustering of translations from the same or similar source languages as discussed in Section 2.3. Section 2.5.3 describes work on answering the question described above, the detection of an author's native language based on their writing in a second language.

2.5.2 Source language detection from Europarl

Work by van Halteren (2008) on the Europarl corpus provided the main framework for the experiments towards the detection of the source language of literary translations detailed in Chapter 5. The purpose of this study is to attempt to determine the source language of translated text from the Europarl corpus using machine-learning methods. 1000 medium length speeches of between 280 and 2500 words were used, in six of the languages from the Europarl corpus, English, Dutch, French, Spanish, German and Italian, 6000 texts in total. Some differences between this and the Baroni study are that translations from and into all of the six languages are considered, thus resulting in a richer set of cross-linguistic information for determining the original source language. van Halteren (2008) relies on XML tags in the corpus to gain information about the source language of a text, tags which are not always present or accurate.

In order to focus on language use rather than content words, tokens which occur in less than 10 % of the texts were replaced with a common placeholder *[X]*, which remain to facilitate the creation of non-contiguous word ngrams, in which one degree of skip is allowed. Four different text classification techniques are used, a simple count-based system they call marker-based classification which focuses on overuse of certain features in translations from different source languages, linguistic profiling which focuses on both over- and underuse of features and Support Vector Classification and Support Vector Regression

Fifteen binary classifications are carried out for each text for each possible pair of the six languages and then these individual binary classifications are used to provide a six-way classification for the source language.

Take a file whose source language is Spanish and whose target language is also Spanish, i.e. a Spanish original text. This text will be compared with all other Spanish target language texts with fifteen different comparisons referring to the target language of the text, SP + EN, SP + DE, SP + IT, SP + NL, SP + FR, EN + DE, EN + IT, EN + FR, EN +NL, DE + IT, DE + FR, DE + NL, IT + NL, IT + FR, and NL + FR. Then for each possible source language, the classification results are added up and the winner is the language with the overall highest score.

Averaging over each of the six languages, the Support Vector Regression method performed the best with an accuracy of 96.7 % , this stemming from a summation over results for the text available in translation into all of the other languages, i.e the system was able to tell the source language of a text with a higher accuracy when the translations into several languages are available. Translations into only one target language gave a lower accuracy, but this was still between 81.5 % for translations into Italian and 87.5 % for translations from Spanish.

The final section of the study by van Halteren (2008) sketches out an initial analysis of source language markers in texts translated into English. These were found by examining the results for the marker based classification which looks at what clusters are more frequent in translations from particular source languages. This approach of course does not examine items which are under-represented in translations. Examples of high-frequency ngrams include *framework conditions*[27] occurring in 22 German source-language texts as compared to 2 English SL texts and 1 French SL texts, and the bigram *certain number*[28] occurring in 25 French SL texts and 1 German, 2 Italian and 2 Dutch SL texts.

2.5.3 L1 detection from text

Somewhat analogous to the task of L1 detection in translation is the task of detecting the L1 of a non-native speaker writing in a foreign language, and Wong and Dras (2009, 2011) use sentence parses and ngram features[29] to detect syntactic idiosyncracies in non-native speaker

[27] from the German *Rahmenbedingungen*

[28] from the French *certain nombre*

[29] character ngrams, POS ngrams, function word frequencies.

text, reporting 80% classification accuracy for seven different L1 types[30] using sentence parses and ca. 70% accuracy using ngram features only on a corpus of learner essays. In this case the corpus was highly comparable, consisting entirely of learner essays in English.

Kochmar (2011) adopts a similar approach to Wong and Dras (2011) in the task of L1 identification of non-native speaker English text, focusing on a number of two-class classification problems, including broader categories such as Romance languages (French, Italian, Catalan, Spanish, Portuguese) vs. Germanic languages (German, Swiss German, Dutch, Swedish and Danish) and more finely grained classifications including Spanish vs. Catalan, for example. Features used include word ngrams, POS ngrams and character ngrams, together with more complex syntactic features such as phrase structure rules and frequencies of different error types. Kochmar (2011) also obtains 84% classification accuracy for the Germanic vs. Romance task using a combination of character unigrams, bigrams and trigrams, POS unigrams, bigrams and trigrams and word unigrams as features and does not perform any multiclass classification experiments in her study, unlike the experiments in this thesis which attempt to classify four different source languages. The target language here was also English. As mentioned previously, the features based on error types do not pertain to the corpus of literary texts used in Chapter 5, due to the fact that they contain a higher quality of language than non-native learner essays, and in the case of the corpus examined in this thesis, are likely to have been subject to an editorial review prior to publication.

Brooke and Hirst (2012) develop an alternate method for the task of L1 classification from non-native English text, they obtain word-for-word translations of word trigrams from a large blog corpus of Chinese, French, Japanese and Spanish text and use these features and also subsets of same (bigrams, unigrams, POS ngrams) as training data for classification of an author's native-language in corpora such as the aforementioned International Corpus of Learner English and other similar collections of text. They report results above the baseline of 25% (48% using word bigrams on the ICLE test corpus) however conclude that the results are not accurate enough to advocate using their method as the sole metric for L1 classification.

2.6 Identifying markers of translator's style

2.6.1 Introduction

This section deals with the literature from corpus linguistics and translation studies which investigates the stylistic properties of a translator. Section 2.6.2 describes an early work in establishing the methodological framework in translation studies, Section 2.6.3 describes early work in corpus linguistics on translated Finnish, with Section 2.6.4 bringing methods from the literary stylometry to bear on the task. Section 2.6.5 surveys studies in the translation studies literature which deal with translations where English and Chinese were either

[30] Bulgarian, Czech, French, Russian, Spanish, Chinese, and Japanese

L1 or L2, and Section 2.6.6 gives an account of studies which deviate from the methodology established by Baker (2000), along with a study on diachronic language change in time-separated translations which is useful to bear in mind when investigating style markers in parallel translations which were carried out a number of years apart.

2.6.2 Baker's framework for investigations into translator's style

In the field of translation studies, Baker (2000) established a framework for investigations into the style of a literary translator using corpora and computational tools. Employing methods from corpus linguistics including type-token ratio (TTR) which is used as a measure of linguistic richness in a text, Baker's framework draws on the field of forensic linguistics concerning textual features which are beyond the conscious control of a translator. Of course, type-token ratio does have its limitations, and indeed text length is a factor in the efficacy of this metric.

She then poses the following questions:

> (a) Is a translator's preference for specific linguistic options independent of the style of the original author?
>
> (b) Is it independent of general preferences of the source language, and possibly the norms or poetics of a given sociolect?
>
> (c) If the answer is yes in both cases, is it possible to explain those preferences in terms of the social, cultural or ideological positioning of the individual translator?

(Baker, 2000, p.8)

She then addresses these questions by focusing on the work of two different British translators, Peter Clark and Peter Bush, the former translating from Arabic, the latter from Brazilian Portuguese and several varieties of Spanish. The lack of a parallel translation of the same work by both translators means that it is difficult to draw conclusions when comparing frequencies of words found in the target text.[31] Baker concludes by proposing the study of several parallel contemporaneous translations of the same text, though acknowledges that these are not regularly available.

After Baker, there have been numerous studies which seek to use statistical methods to produce an overview of a translator's style.

2.6.3 Translator's stylistic markers in translated Finnish

Mikhailov and Villikka (2001) use statistical measures from corpus linguistics in an attempt to quantify a translator's stylometric fingerprint. Their corpus consists of a number of literary

[31] The author compares frequencies of the verb *say* for both translators, while acknowledging that the frequency of the equivalent word *qaal* in Arabic may be indeed higher than in English or the other source languages.

translations from Russian into Finnish with several authors and translators, with only one set of parallel Finnish translations of the same Russian source text, Fyodor Dostoyevski's *Notes from The Underground*. They report that the values of R, K and W (see equations below) are "almost identical" for these two translations.[32]

They summarise the R, K and W metrics as follows: the R quotient reflects the number of hapax legomena[33], H in the equations, the K quotient reflects the number of high frequency words in the text and the W quotient is a form of lexical richness measure, representing the number of unique words in the text. In the following equations, U represents the total number of unique words in a corpus, with T representing the total words.

$$R = \frac{100LogT}{1 - \left(\frac{H}{U}\right)} \tag{2.6}$$

$$K = \frac{10^4 \left(\sum_{i=1}^{\infty} i^2 H - T\right)}{T^2} \tag{2.7}$$

$$W = T^{U^{-0.172}} \tag{2.8}$$

Investigating the translational choices with regard to the source text, focusing on the Russian word *kazhetsja* 'it seems to be', they report that one translator in particular favoured the Finnish translation *taitaa* for this word over all other alternatives such as *mielestäni* or *ilmeisesti* . They also report similarities between texts translated by the same person in relation to the ratios of words, paragraphs or sentences in the original text to the equivalent textual unit in the source text. They conclude that translator style may manifest itself in the use of grammatical items such as modals and the expansion or shortening of the length of the target text, among other defining features.

2.6.4 Translator's style and Burrow's *Delta*

This topic of translator stylometry has been considered to some extent in the digital humanities, notably in the work of Burrows (2002a) who examines fifteen different translations of the Roman poet Juvenal's *Tenth Satire* into English from the original Latin with a chronological span from 1646 to 1967, four of which were prose translations, the rest composed in verse. Burrows uses his own Delta metric (Burrows, 2002b) which has been used in studies on stylometry of character contributions in literary text and authorship attribution exercises.

> A delta-score, as I propose to term entries like those in L4 and Q4, can be defined as the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text.

(Burrows, 2002a, p.5)

[32] R: 1038.76 vs. 1034.74, K: 40.03 vs. 40.94 W: 8.54 vs. 8.48

[33] words occurring only once in the text.

which in mathematical form resembles the following:

$$\sum_{i=1}^{n} |(z(X_i) - z(Y_i)| \tag{2.9}$$

with the equation for calculating the z-score here:

$$z = \frac{\text{Raw score} - \text{Population mean}}{\text{Standard deviation}} \tag{2.10}$$

Stein and Argamon (2006) combine the two equations to create a definitive formula for Burrow's Delta:

$$\sum_{i=1}^{n} |\frac{X_i - Yi}{\sigma_i}| \tag{2.11}$$

Using the top twenty most frequent words to calculate the delta score, Burrows identifies some patterns of interest among the translations, in particular the translation by seventeenth century English author Thomas D'Urfey which appeared as the translation most similar to all the others using the values of Delta as a comparison. Burrows remarks that D'Urfey's style in his own writing may echo the style of Juvenal in a fashion that even translators who have not heard of the author in question may subconsciously try to imitate when translating the Latin verse into English.

Burrow's work is novel in two particular ways: for his use of the Delta metric in his experiments and the fact that he examines a large number of translators. Although he does take a closer look at individual word frequencies and samples within a number of translations in order to illustrate certain conclusions, the Delta metric as it is used here can only predict the similarity or divergence in style of a pair of texts based on a small sub-class of frequent words only and does not necessarily identify a wide variety of distinguishing features, such as sentence length or lexical richness measures.

The Delta metric is also adopted by Rybicki (2006) who investigates the idiolects of character contributions in two temporally-separated translations of the Polish author Henryk Sienkiewicz's trilogy. Multidimensional scaling plots are created which show tight groupings between character idiolects from each translation based on the 250 most frequent words.

Our own earlier work detailed in Lynch and Vogel (2009) describes a similar task of investigating the clustering of character idiolects based on parallel translations of the plays of Henrik Ibsen into English and German. This work investigated the internal homogeneity[34] of character contributions using the χ^2 statistical metric and word unigram features. Using multidimensional scaling, closer patterns were observed in character idiolects in the parallel contemporaneous translations of the play *Ghosts* by Henrik Ibsen, translations which are

[34]In this case, how distinguishable was the speech of one character from the speech of the others, and how was this preserved in the translations.

revisited in Chapter 6 where stylometric differences between these parallel translations and also a corpus of translations of other other Ibsen plays by the same translators are investigated using supervised learning methods. The difference in focus however in this study is that the stylistic patterns of the individual translators is of interest, as opposed to the homogeneity of character contributions.

Bolstered by the success of the metric in this task and on a number of other studies in the domain of translation stylometry, Rybicki (2012) examines a corpus of translations by a number of different translators[35] in an endeavour to identify stylistic features which distinguish translators. The corpus consist of both translations from English to Polish and Polish to English, in the latter case focusing again on the work of Henryk Sienkiewicz. He uses the Delta method again to cluster texts together based on a measure of the 5000 most frequent words, and observes works clustering by original author rather than translator.

He concludes that his results corroborate Venuti's theory of a translator's invisibility (Venuti, 1995), although recognises the shortcomings of the Delta metric and the practice of only focusing on a number of frequent words. One could indeed argue that a larger range of features may provide more evidence for stylistic idiosyncracies, based on observations in Chapter 6.

2.6.5 Translator's style in translations from Chinese to English and English to Chinese

There have been a number of studies in translation stylistics on corpora of English translated from Chinese and Chinese texts translated into English.

Li, Zhang, and Liu (2011) use similar methods to Baker (2000) and Mikhailov and Villikka (2001) in a study of two English translations of the Chinese epic novel *Houggloumeng*, expanding on the initial statistical analysis with an interpretation on the reasons behind the differences in TTR and other statistical metrics based on the socio-cultural environments of the translators. They examine two translations of this Chinese epic, Hawkes and Minford's 1974 translation titled, *The Story Of The Stone*, and a later translation by Xianyi and Gladys Yang between 1978 and 1980 with the title, *A Dream Of Red Mansions*. They focus on these two versions out of a possible nine English translations. Socio-cultural issues are high on their agenda and they utilise corpus linguistics methods to illustrate these. Reporting on the STTR[36] and sentence length differences between the two translations, they find that Hawkes and Minford's translation uses longer sentences and more words than the Yangs' version, with the Yangs' version displaying a wider range of vocabulary (higher STTR) than Hawkes and Minford's version.

They then turn to explaining these differences, arguing that due to Hawke's status as a non-native Chinese speaker and Sinologist, he tended to paraphrase and explain Chinese

[35] Including his own translations from English to Polish of several authors such as John Le Carré and Douglas Coupland

[36] standardised type-token ratio: average type-token ratio per thousand words.

cultural concepts in a more verbose manner than perhaps the Yangs felt was necessary. As a result, his version eschewed footnotes, where the Yangs embraced them. Li et al. (2011) present evidence for this difference, including the manner of translation[37], cultural issues regarding the translation such as the location of the translators[38] and cultural differences in translation styles between mainland China and Hong Kong. They cite translation audience as a factor in the difference in STTR, giving the example of the Chinese phrase *cloud and rain* which refers to sexual intercourse. This is translated in a straightforward fashion by Hawkes as *making love* compared with the Yangs' version where they use their own more literal rendition, *rain-and-cloud-games*. Many of their theories are based on personal statements from both sets of translators in various publications.

Although their work is an in-depth study which marries corpus linguistic methodology with qualitative background knowledge, this thesis will not focus in detail on the cultural and personal backgrounds of the translators examined in Chapter 6, as this is not within the scope of the current dissertation. However the distinguishing factors of average sentence length and standardised type-token ratio are marked as noteworthy features of translators style, with a view to examining the nature of these same features in the corpus examined in Chapter 6.

Wang and Li (2012) examine Chinese translations of Joyce's *Ulysses* by translators Xiao and Jin with a focus more akin to the work in Chapter 6, attempting to identify features of translator's style differentiating between features reflecting lexical choice by the translator and features which represent source language effects. One example of a source language effect in Chinese is given as the translator and author Xiao's tendency to post-position adverbial clauses in his translation, which they interpret as transfer from English syntactic norms. They also report on the stylistic differences between Xiao's translation and his counterpart Jin's version, focusing on common words such as the verb *to know* which is rendered as *xiaode* by Xiao and *zhidao* by Jin, the former representing a more colloquial form of the word as used in the Shanghai dialect. Xiao also exhibits a stylistic preference for the word *duo*[39] to a significant extent, translating a total of twenty-four different verb forms in the original English using this verb *duo* in combination with an adverbial modifier to indicate the preferred meaning.

Comparing the work by Wang and Li (2012) with work by Li et al. (2011), aside from the difference in source and target languages[40], the former employs more statistical measures and focuses less on cultural aspects, although still proposing reasons for a preference for one particular lexical item over another. The work in Chapter 6 will focus even less on cultural aspects than Wang and Li (2012), although future work could indeed benefit from

[37]The Yangs translated orally, with Xianyi rendering the original into rough English and his wife Gladys smoothing the result into a more fluent form.

[38]Hawkes was a university professor based primarily in the UK, Yang was a high-ranking translator employed by an official Chinese government translation agency tasked with the translation and publication of important Chinese literary works in world languages.

[39]to stroll, or to saunter

[40]Wang focuses on English to Chinese, with Li's work focusing on Chinese to English

collaboration with a more experienced scholar of translation or literary studies in particular.

2.6.6 New approaches towards detecting a translator's style

A more complex approach to the identification of translator style is employed by El-Fiqi, Petraki, and Abbass (2011), who adopt methodology from network theory to identify stylometric patterns in two translations of the Holy Q'uran into English. They identify non-contiguous[41] sequences of words in the text of the Q'uranic verses and use a set of these to train a Fuzzy Lattice Reasoning Classifier which obtains a classification accuracy of 70% in detecting the translator of a segment. They do not provide detailed descriptions of the relevant features extracted in their study, their use of motifs refers only to a document-level structural representation of a word sequence, rather than the word sequence itself.

Their work presents a novel approach towards the detection of stylistic properties of a translator, however their abstract representation of word sequences does not facilitate interpretation of the results. One can also argue that using methods of text classification coupled with ngram and document-level statistics can provide a valid enough description of certain aspects of translator style and perhaps a more coherently interpretable one to researchers more familiar with text classification and corpus linguistics. This is the approach which is taken in Chapter 6, with comparable, if not clearer results than El-Fiqi et al. (2011). Nevertheless, the work is interesting due to its choice of a more complex representation of translator style.

2.6.7 Language change investigation from time-separated translations

Although the main focal point is not translatorial style but rather language change over time, work by Altintas, Can, and Patton (2007) on measuring language change in Turkish using parallel translations of the same texts is important to consider as this is another confounding factor in the identification of translator style markers. They investigate two corpora of translated Turkish, one from the period 1940-1957 and one from the period 1990-1997, with Russian, French and English as source languages and each translation in the first corpus paired with a corresponding modern translation in the second corpus. In their experiments, they use average length of word stems and suffixes[42] and a number of corpus linguistic metrics such as TTR and lexical richness measures. For word stems, they find a statistically significant difference in the TTR, with the newer translations having a lower TTR than the older works[43]. They are aware of the limitations of TTR and use same-size samples of 1000 words in their experiments. For stem and word lengths, using logistic regression analysis, they determine that stem lengths have decreased over time but in fact word length in general has increased, due to an increase in the length of suffixes in the language. As a caveat,

[41] they use non-adjacent word ngrams in their representation - n-skipgrams

[42] As Turkish is an agglutinative language, the distribution of word lengths and suffix lengths signifies different phenomena as it would in English, for example.

[43] 14.867 to 12.516.

they mention the *translationese* concept, accepting that their corpus of Turkish may indeed represent a particular dialect of the language and thus any trends within this dialect may not generalise to the language as a whole. For the experiments carried out in this thesis, is it important to acknowledge this change in language over time as a potential confounding factor in the stylistically discriminating features between two parallel translations of the same work. In Chapter 6 this potential issue is addressed by referring to a diachronic corpus of English when examining certain features in detail.

2.7 Conclusion

Table 2.1 summarises the experimental setup over a number of key studies examined in Chapter 2. There are a range of different corpora examined from parliamentary proceedings to student essays to current affairs texts. Seven different European languages[44] are examined in the various studies. Perhaps most interesting is the diverse size of the corpora used in the experiments, which range from a few hundred thousand words to eighty million words in the case of Kurokawa et al. (2009). In general, papers focused on binary classification tasks such as translated vs non-translated text and native vs non-native text, however source language detection has a number of languages to choose from, work by Argamon et al. (2009) examined user gender, age, personality and native language using different corpora and Luyckx and Daelemans (2008) were concerned with four different axes of personality type.

Table 2.2 collects features, accuracy and classification methods across experiments, Support Vector Machines occur frequently as the top-performing classification method, features include a mix of POS and word ngrams in the majority of the studies, perhaps unsurprising as these are generally the most popular in the text classification literature. Important also are the document-level features which until recent times remained the preserve of translation studies research, these are shown to give excellent accuracy in classification in (Ilisei et al., 2010; Ilisei & Inkpen, 2011). In studies where accuracy was measured, the accuracy for classifying translations from non-translations was quite high, compared with the studies by Argamon et al. (2009) and Luyckx and Daelemans (2008) on detecting other information such as personality type, age and gender from text.

The experiments carried out in this thesis seek to compare and combine both document-level features and ngram features, an experimental setup which is not common in the literature, with the possible exception of the work by Pastor et al. (2008). Where necessary, average frequencies of document-level metrics will also be examined. Different corpora are used in Chapters 4, 5 and 6, this was initially done due to the different categories of text required in each chapter. Europarl provided adequate texts which were annotated by source language, however translator information is not present, which limits its usefulness for translator style analysis. Due to the readily available information on source language, Europarl

[44] English, Spanish, Italian, Dutch, Romanian, French, German.

would also have been a valid corpus for source language detection but this question has been dealt with extensively in studies by van Halteren (2008) and Koppel and Ordan (2011). For this reason, a corpus of literary texts was compiled for the source language detection experiments, in order to validate the methods on a corpus of texts which is not as stylistically coherent as Europarl. Again, this corpus could have proven interesting for translator style experiments, although parallel translations of the same text were excluded from the corpus on grounds that they might introduce confounding factors for the source language detection task.

The translations of Ibsen's Ghosts had been examined in earlier work (Lynch & Vogel, 2009) which focused primarily on the preservation of character idiolects in translation, however Ibsen's drama proved a useful set of text for translator style experiments as it was possible to obtain parallel translations of the same drama by different translators, who had also translated other plays by Ibsen without much temporal separation between translations.[45] Investigating different corpora for each experiment also enables cross-genre comparisons to be carried out, which can be of use when investigating translation universals.

Structural parses of texts have not been used in the experiments due to a lack of experience with these representations and the software to create them and measures such as perplexity and entropy from the machine learning literature are not implemented either, as this thesis seeks to primarily investigate the efficacy of document-level metrics due to their descriptive purposes and established nature in the literature, although the use of information-theoretic measures should not be ruled out entirely in future experiments.

The experiments in this thesis adopt the *supervised learning* approach, which involves *training* a classifier on a training set and then *testing* this classifier on a test set from the same distribution, in order to identify features which robustly distinguish different categories from one another. The categories of text examined here are: translated vs. original text, translations from a number of source languages or indeed translations of the same work by different translators.

Source language detection from translation has not been a fertile area of research in previous years, indeed with only a handful of studies focusing on this topic in isolation, however the experiments in Chapter 5 seek to investigate this phenomenon in literary translation and indeed it is hoped that the promising results in this experiment will lead to further investigations on this topic, perhaps also side-by-side with the literature on L1 detection from non-native text which is a closely related task in computational linguistics.

Until recently, there has been little research investigating the topic of translator style using supervised learning methods as per (Baroni & Bernardini, 2006; van Halteren, 2008; Ilisei et al., 2010). The work in Chapter 6 sets out a framework for future studies of parallel contemporaneous translations and translator influence over authorial style using these methods in particular and the same feature set used in Chapters 4 and 5. x

[45] Section 2.6.7 describes how translations from fifty years apart can be used to investigate language change in Turkish, a large temporal gap between translations could add confounding elements to any studies of translator style.

Author	Language	Genre	Quantity	Task
Laviosa98	En	Fiction	2m words	T not(T)
Baroni06	It	Current Affairs	3m words	T not(T)
VanHalteren08	6 EU	Europarl	6000 texts	Which SL?
Koppel11	En	Europarl/IHT	5m,2.5m words	Which SL,T not(T)
Kurokawa09	En,Fr	Hansard	80m words	T vs not(T)
Ilisei10	Es	Med,Tech	600 texts	T vs not(T)
Popescu11	En	Literary	214 novels	Which SL?
Ilisei11	Ro	News	630 texts	T vs not (T)
Lauttamus07	En	Spoken	305K words	N vs not(N)
Luyckx08	Nl	Essays	200K words	Various
Argamon09	En	Essays/Blogs	>140K words	Various

Table 2.1: Experimental setup summary

Author	Features	Accuracy	Method(Best)
Laviosa98	Document-level	n/a	n/a
Baroni06	Mixed ngrams	86.7	SVM
VanHalteren08	Mixed ngrams	96.7	SVM,SVR
Kurokawa09	Word 2-grams	90(En)	SVM
Koppel11	Word unigrams	97%(EP),87.5%(IHT),90%(Both)	BMR
Ilisei10	Document-level	97(tech),83(med)	SVM
Popescu11	String kernels	77%	SVM
Ilisei11	Document-level	98.6	SVM
Lauttamus07	POS 3-grams	n/a	Perm test
Luyckx08	POS,word 3-grams	65(IE)62(IS)73.8(FT)83(JP)	k-NN
Argamon09	POS,word ngrams	76(G)77.7(A)82.3(NL)65.7(P)	BMR

Table 2.2: Features and results summary

Chapter 3

Methods

3.1 Introduction

This chapter explains in more detail the software packages and statistical metrics used in this thesis. A mixture of off-the-shelf packages and custom code[1] was used to generate the feature-sets which are then used in the experiments in Chapters 4, 5 and 6. Section 3.2 describes the machine learning toolkits and NLP tools used in the experiments, with Section 3.4 describing in detail the document-level statistics used.

3.2 Software packages

3.2.1 WEKA

The main software package used to carry out the experiments described in this thesis was the WEKA machine-learning toolkit, an open-source Java-based workbench for carrying out experiments using a number of machine-learning algorithms, developed by the University of Waikato in New Zealand, (Hall, Frank, Holmes, Pfahringer, Reutemann, & Witten, 2009).

WEKA expects datasets in its own ARFF[2]format, which consists of a header containing the attributes present in the file and their types(numeric, String etc) followed by a *@Data* tag which precedes the values of these attributes for each instance. A sample ARFF file should resemble the following:

```
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class
{Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```

Figure 3.1: Sample WEKA ARFF file, from http://cahitarf.sourceforge.net/arff.html, last verified May 7, 2013

[1] Appendix A.2 contains the code used to calculate the document-level metrics for this thesis.
[2] Attribute-Relation File Format.

As observed in Figure 3.1, the header contains *@ATTRIBUTE* tags followed by the names and types of the attributes. Notice the class attribute contains a list of strings, separated by commas. This is refered to as a nominal attribute type, which basically provides a list of all the possible values of a certain attribute. The class variable is usually nominal in an ARFF file. Some classifiers cannot handle *STRING* attribute types and WEKA provides a filter to convert these to nominal format.

WEKA provides both a reasonable fully-functional GUI interface for carrying out experiments in an interactive fashion coupled with a command-line mode which can be useful for running batch-type experiments.

Of course using a ready-made one-size-fits-all software tool for a quite specific task can have disadvantages, and WEKA is no exception to this. It does not contain any native processing filters for natural language processing tasks which could facilitate the creation of the most basic feature sets used here which consist of word ngram frequencies.

Due to the fact that it is written entirely in the Java programming language, it also suffers from limitations inherent in Java itself, such as the allocation of large heap sizes on 32bit machines[3], a limitation which presents itself when dealing with large feature sets, such as the aforementioned word bigrams and unigrams, which depending on the corpus itself can be larger than 200,000 individual features.

However, the large number of algorithms and classifiers available coupled with the relatively intuitive GUI provided make up for the shortcomings with regards to NLP specific issues. There are third-party NLP add-ons available for use with WEKA and the one which proved most workable for the purposes of this study is the TagHelperTools package, which is described in Section 3.2.2.

3.2.2 TagHelperTools

TagHelperTools is an NLP processing add-on for WEKA which was designed for use by social scientists for the analysis of survey and coded results from qualitative studies.

Due to the fact that the software is aimed at this particular group, it does not provide many options which would be of use to the average computer scientist. One particular drawback is the lack of an actual command-line-only mode, coupled with input issues and file format limitations[4].

A number of workaround scripts were written to allow the type of corpora used in this thesis to be analysed by this particular piece of software, including one script which transforms a directory of text files into a tab separated CSV file, with a category and text column, the former containing the category name for classification and the latter containing the content of the text file itself.

Despite these initial drawbacks, the software interfaces well with WEKA and provides

[3]Java only allows heap sizes of up to 2 gigabytes on 32bit machines.

[4]The default input is a tab-separated Excel file, it offers no support to input a directory of text files, for example

a number of options for creating WEKA compatible ARFF files including a number of feature types, word unigrams, word bigrams and part-of-speech bigrams.[5]. The POS tagging is handled by the Stanford tagger, and it supports data in German, English, Spanish and Chinese.

TagHelperTools provides inbuilt switches for stopword removal and lemmatization, however none of these were used in our study, due to the fact that some stopwords are precisely the tokens which are of most interest in the detection of stylistic patterns.

3.2.3 TreeTagger

The TreeTagger (Schmid, 1994) was developed at the IMI institute in the University of Stuttgart, Germany. It is a probabilistic part-of-speech tagger which uses the Penn-treebank tagset which is the same tagset as the Stanford POS tagger which comes as part of the TagHelperTools package described in Section 3.2.2. This tagger is used for the generation of the POS tags used to calculate the metrics described in Section 3.4. The two taggers utilise the same tagset which is useful when cross-referencing features.

3.3 Classification metrics

3.3.1 Naive Bayes

The Weka implementation of a Naive Bayes classifier is used in the experiments, this classifier can be explained as follows:

A Naive Bayes classifier assumes that each variable in a set of variables is independent from one other, hence the *naive* in the name of the classifier. The classifier computes the likelihood of each class from the dataset, based on the frequency of each class in the dataset. This forms part of the classification probability, which is normally refered to as the *prior* possibility. Next, the probability of occurrence in each class is computed for all of the possible features in the dataset.

A Naive Bayes classification function is then computed, example from (Zhang & Su, 2004):

$$[h]C_{nb}(E) =_c \arg\max \mathrm{p}(C) \prod_{i=1}^{n} p(a_i|c) \tag{3.1}$$

where E refers to an example, C is a class and c represents the actual value of the class. To summarise in brief terms, the most probable class assignment for an example is that which maximizes the *prior* likelihood for the class multiplied by the probability of each of the i elements(a) in E occurring in the class.

There are some disadvantages to using the Naive Bayes classifier for NLP tasks, one obvious one being the independency assumption, as the frequency of occurrence of certain word types are not independent and in fact closely related to one another.

[5] Although curiously no support for POS unigrams

Although simple however, Naive Bayes classifiers can be robust and produce surprisingly good results:

> This method is important for several reasons, including the following. It is very easy to construct, not needing any complicated iterative parameter estimation schemes. This means it may be readily applied to huge data sets. It is easy to interpret, so users unskilled in classifier technology can understand why it is making the classification it makes. And, particularly important, it often does surprisingly well: It may not be the best possible classifier in any given application, but it can usually be relied on to be robust and to do quite well.

Wu and Kumar (2009, p.163)

3.3.2 Support Vector Machines

Support Vector Machines or SVM's are a highly popular form of machine-learning classifier proposed first by Cortes and Vapnik (1995) as a solution to a two class classification problem.

> For a two-class linearly separable learning task, the aim of SVC is to find a hyperplane that can separate two classes of given samples with a maximal margin which has been proved able to offer the best generalization ability. Generalization ability refers to the fact that a classifier not only has good classification performance (e.g., accuracy) on the training data, but also guarantees high predictive accuracy for the future data from the same distribution as the training data.

(Wu & Kumar, 2009, p.38)

They define an equation for the optimal hyperplane in 3.3.2, with w as the weight vector and b as the bias

$$w^T x + b = 0 \tag{3.2}$$

The distance r from a boundary sample x to the hyperplane as shown in Figure 3.2 from Wu and Kumar (2009) is given as follows.

$$r = \frac{g(x)}{||w||} \tag{3.3}$$

where $g(x) = \mathbf{w}^T x + b$, also known as the discriminant function of x.

Thus, a *maximal margin classifier* tries to find optimal values for $\mathbf{w}$ and b such that ρ(See Figure 3.2) or the *margin of separation* defined by the shortest geometrical distances(r^* in Figure 3.2) from each class boundary to the hyperplane, is maximised.

Letting the functional margin equal one, they then define for a training set $\{\mathbf{x}_i, y_i\}_{i=1}^n \in \mathbf{R}^m \times \{\pm 1\}$

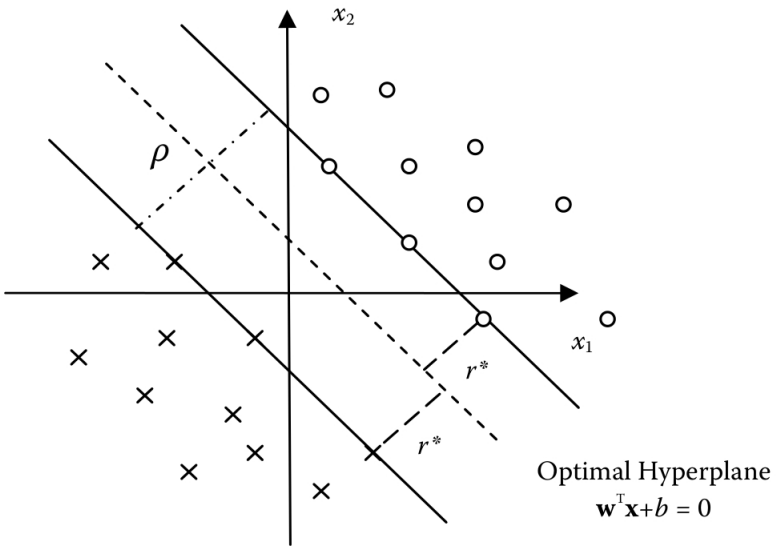

Figure 3.2: Diagram displaying maximum margin classifier for two-class linearly-separable problem

$$\begin{aligned} \mathbf{w}^T\mathbf{x}_i + b \geq 1 \quad for \quad y_i = +1 \\ \mathbf{w}^T\mathbf{x}_i + b \leq 1 \quad for \quad y_i = -1 \end{aligned} \tag{3.4}$$

Data points $\{\mathbf{x}_i, y_i\}$ are the so-called *support vectors*, these are the data points in Figure 3.2 which are closest to the hyperplane. The geometrical distance r^* from the support vector x^* can be defined as follows:

$$r^* = \frac{g(\mathbf{x}^*)}{||\mathbf{w}||} = \begin{cases} \frac{1}{||\mathbf{w}||} & if \quad y^* = +1 \\ -\frac{1}{||\mathbf{w}||} & if \quad y^* = -1 \end{cases} \tag{3.5}$$

According to Figure 3.2, ρ, also refered to as the *margin of separation* is defined as:

$$\rho = 2r^* = \frac{2}{||\mathbf{w}||} \tag{3.6}$$

Thus, a *support vector classifier* can be defined as a maximisation problem on ρ with respect to $\mathbf{w}$ and b:

$$\begin{aligned} &max_{w,b} \frac{2}{||\mathbf{w}||} \\ s.t. \quad &y_i(\mathbf{w}^T x_i + b) \geq 1, \quad i = 1, ..., n \end{aligned} \tag{3.7}$$

which is equivalent to:

$$\begin{aligned} &min_{w,b} \frac{||\mathbf{w}||^2}{2} \\ s.t. \quad &y_i(\mathbf{w}^T x_i + b) \geq 1, \quad i = 1, ..., n \end{aligned} \tag{3.8}$$

Unfortunately, real problems are not so easily linearly separated, and this is where an extra step is required.

Tan, Steinbach, Kumar, et al. (2006, p.270) describe the process for non-linear support vector machines. The most obvious solution is to create some nonlinear transformation Φ to project the data into a new feature space where it will be linearly separable.

However this can run the risk of encountering the so-called *curse of dimensionality*. The optimal solution is to define a similarity function known as a *kernel function* which when computed for a pair of vectors in the original space is equivalent to the dot product of these vectors in a higher dimensional space. Computing this function is computationally less intense than transforming the set of attributes using Φ and then defining the separating hyperplane on the transformed data.

One such kernel function is the polynomial function:

$$K(\mathbf{x},\mathbf{y}) = (\mathbf{x}.\mathbf{y}+1)^p \tag{3.9}$$

which is used as the kernel function in the default SVM implementation in Weka.

Joachims (1998) pioneered the usage of Support Vector Machines in text classification tasks and since then they have been used in a wide variety of tasks involving textual corpora including, but not limited to: the detection of male/female language identification (Koppel, Argamon, & Shimoni, 2002), author profiling (Argamon et al., 2009), debate position classification (Thomas, Pang, & Lee, 2006) and personality detection (Mairesse & Walker, 2008) from text. They are also used by (Baroni & Bernardini, 2006), (Kurokawa et al., 2009), (van Halteren, 2008) and (Ilisei et al., 2010) in their experimentation on comparable corpora of translated and original text.

3.3.3 Simple Logistic Regression

The Simple Logistic Regression classifier in Weka implements a logistic regression model. *Logistic regression*, as opposed to *linear regression* seeks to predict a number of discreet values based on a set of input variables. In the case of the experiments in this thesis, one might wish to develop a model for predicting whether a text is in fact a translation or original. Similar to an SVM classifier, this method is ideal for binary classification.

Witten, Frank, and Hall (2011, p.126) give an account of logistic regression as implemented in the Weka toolkit. They assume two classes, and an original target variable $[Pr[1|a_1,a_2,....,a_k]$. This function cannot be approximated by linear regression. Instead, the transformation function, known as the *logit* function (see Figure 3.3) is computed for the variable

$$\frac{log[Pr[1|,a_1,a_2,....,a_k]}{1-[Pr[1|a_1,a_2,....,a_k]} \tag{3.10}$$

This transforms the output of the regression function from $\{0,1\}$ to $\{-\infty,+\infty\}$.

This is usually expressed as a linear function similar to linear regression:

$$[Pr[1|a_1,a_2,....,a_k] = \frac{1}{1+exp(-w_0-w_1a_1....-w_ka_k)} \tag{3.11}$$

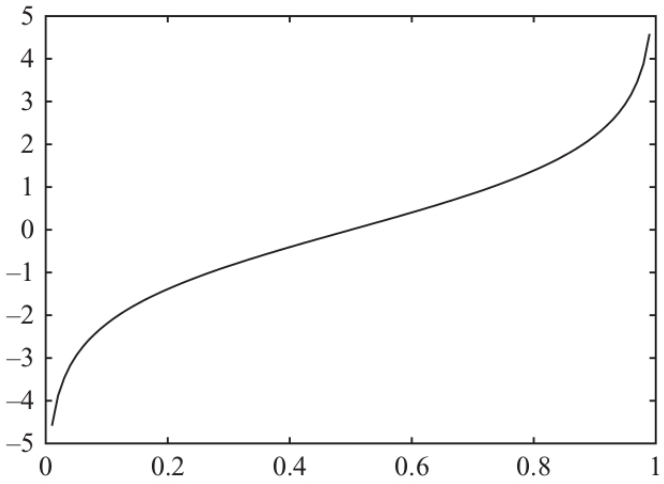

Figure 3.3: Logit transformation curve

To find weights that fit the data, the *log-likelihood* ratio is introduced:

$$\sum_{i=1}^{n}(1-x^{(i)})log(1-Pr[1|a_1^{(1)},a_2^{(2)},....,a_k^{(k)}])+x^{(i)}log(Pr[1|a_1^{(1)},a_2^{(2)},.....,a_k^{(k)}]) \quad (3.12)$$

where x^i is equal to zero or one. They propose a simple solution to this problem by iteratively a sequence of weighted least-squares regression problems until the log-likelihood ratio converges to a maximum.

3.3.4 Decision Tree Classifier

The J48 decision tree classifier in Weka is an implementation of the C4.5 decision tree algorithm formulated by Ross Quinlan (Quinlan, 1993, 1996).

> The following algorithm generates a decision tree from a set D of cases:
>
> - If D satisfies a *stopping criterion*, a tree for D is a leaf associated with the most frequent class in D. One reason for stopping is that D contains only cases of this class, but other criteria can be formulated. (see below)
> - Some test T with mutually exclusive outcomes T_1; T_2.....T_k is used to partition D into subsets D_1; D_2.....D_k, where D_i contains those cases that have outcome T_i . The tree for D has test T as its root with one subtree for each outcome T_i that is constructed by applying the same procedure recursively to the cases in D

(Quinlan, 1996, p.2)

The idea of a *splitting criterion* is then introduced:

The default splitting criterion used by C4.5 is gain ratio, an information-based measure that takes into account different numbers (and different probabilities) of test outcomes. Let C denote the number of classes and p(D; j) the proportion of cases in D that belong to the jth class. The residual uncertainty about the class to which a case in D belongs can be expressed as:

$$Info(D) = -\sum_{j=1}^{C} p(D,j) \times log_2(p(D,j)) \tag{3.13}$$

$$Gain(D,T) = InfoD^C p(D,j) \times log_2(p(D,j)) \tag{3.14}$$

(Quinlan, 1996, p.2)

Ramakrishnan (2009) describes the basic C4.5 process in an algorithmic fashion:

C4.5(D)
Input: an attribute-valued dataset D
Tree = {}
if D is "pure" OR other stopping criteria met then terminate
end if
for all attribute a $\in D$ **do**
compute information theoretic criteria if we split on a
end for
a_{best} = Best attribute according to above computed criteria
Tree = Create a decision node that tests a_{best} in the root
D_v = Induced sub-datasets from D based on a_{best}
for all D_v **do**
$Tree_v$ = C4.5(D_v)
Attach $Tree_v$ to the corresponding branch of Tree
end for
return Tree

More advanced features of the algorithm included different pruning methods for generating the most optimum trees, Ramakrishnan (2009) provides detailed descriptions of these.

The main reason for using the Decision Tree classifier in Chapter 6 was to provide an easily interpretable representation of the distinguishing values of the various document-level features, and this is where this particular classifier can be very useful, with an example from Ramakrishnan (2009) in Figure 3.4.

This decision tree has been induced from 14 instances of data about the weather as displayed in Table 3.1.

```
outlook = overcast: Play (4.0)
outlook = sunny:
      | humidity <= 75 : Play (2.0)
      | humidity > 75 : Don't Play (3.0)
outlook = rain:
  |windy = true: Don't Play (2.0)
  |windy = false: Play (3.0)
```

Figure 3.4: Decision tree for golf dataset

Day	Outlook	Temperature	Humidity	Windy	Play Golf?
1	Sunny	85	85	False	No
2	Sunny	80	90	True	No
3	Overcast	83	78	False	Yes
4	Rainy	70	96	False	Yes
5	Rainy	68	80	False	Yes
6	Rainy	65	70	True	No
7	Overcast	64	65	True	Yes
8	Sunny	72	95	False	No
9	Sunny	69	70	False	Yes
10	Rainy	75	80	False	Yes
11	Sunny	75	70	True	Yes
12	Overcast	72	90	True	Yes
13	Overcast	81	75	False	Yes
14	Rainy	71	80	True	No

Table 3.1: Data set for golf decision tree

3.4 Document-level features

This section describes the different statistical measurements[6] which were calculated during the experiments. Table 3.2 details the eighteen features used in the analysis.

These metrics are used by Pastor et al. (2008) and by Ilisei et al. (2010) and Ilisei and Inkpen (2011) in their studies on translated and original Spanish and Romanian text, which have informed the experimental approach taken in this thesis. Pastor et al. (2008) quotes the work of Douglas Biber on linguistic variation in English (Biber, 1988, 1995, 2003) as inspiration for a number of the features used, and claim to have come up with the idea for some of these from their own analyses, although they provide the following caveat:

> Some of these features have been adopted from Biber (1995), Biber (2003); other such as the type of sentences, are our own proposals. It is worth noting that the set of stylistic features is language dependent.

(Pastor et al., 2008, p.3)

Indeed in relation to the aspect of language dependence many of the readability scores are based on English text and used by Pastor et al. (2008) on Spanish text, however as the

[6] Generally referred to as *document-level features*.

feature	Description
avgsent	Average sentence length
typetoken	Ratio of word types to total words
lexrichness	Ratio of lemmas to total words
infoload	Ratio of open-class words to total words
avgwordlength	Average word length
nounratio	Ratio of nouns to total words
ARI	Readability metric
CLI	Readability metric
grammlex	Ratio of open-class words to closed-class words
conjratio	Ratio of conjunctions to total words
pnounratio	Ratio of pronouns to total words
simplecomplex	Ratio of simple to complex sentences
complextotal	Ratio of complex sentences to total sentences
numratio	Ratio of numerals to total words
fverbratio	Ratio of finite verbs to total words
prepratio	Ratio of prepositions to total words
dmarkratio	Ratio of discourse markers to total words
simpletotal	Ratio to simple sentences to total sentences

Table 3.2: Document-level features

corpora in this thesis consist entirely of English text, one can be more confident in the accuracy of the scores, although the main aim is not necessarily obtaining the readability score for a particular text but in fact examining the relative difference in readability between textual corpora. One set of features which are not implemented are the features based on the depth of a particular sentence parse, however these features are of interest in any future experimental work on the topic. Ilisei et al. (2010) use a classifier made up of a list of different document metrics and this is the procedure which is followed also in this thesis although the mean and standard deviation of a number of metrics is also examined in isolation for each experiment to examine how individual features differ between translated and original text.

3.4.1 Average sentence length

Average sentence length[7] is calculated as text length divided by number of sentences. In actual terms, using the POS-tagged tokens of a text, this manifests itself as total number of tokens - number of sentence markers divided by number of sentence markers. In the extreme case that a chunk of text contains no sentence marker, the number of sentences is set to one.

$$\frac{\text{Total number of tokens}}{\text{Number of sentences}} \tag{3.15}$$

In the literature, Laviosa-Braithwaite (1997), Laviosa (1998) and Mikhailov and Villikka (2001) investigate average sentence length. Laviosa (1998) finds that the average sentence

[7]Often referred to during the thesis in tables as *avgsent*.

length of translated text in a corpus of news articles is significantly lower than the average sentence length for non-translations but that the average sentence length for literary translations is actually higher than for original texts.

Pastor et al. (2008) find that in their comparable corpus of technical texts, there is a statistically significant difference between the average sentence length of the translated section and the average sentence length in the original section. The translated section had a longer average sentence length (27.29 vs 18.2) than the original section.

Li et al. (2011) find statistically significant differences between the average sentence length of the two translators they examine.

3.4.2 Type/token ratio

The *type-token ratio* of a text, often referred to as TTR, is a document-level statistic that calculates the diversity in the vocabulary of a piece of writing. It is calculated by dividing the total number of tokens by the number of unique token types.

$$\frac{\text{Total number of token types}}{\text{Number of tokens}} \tag{3.16}$$

There are some limitations to this metric, the most obvious one being that the TTR gradually declines over the length of a text, meaning that it is not directly comparable for texts of different lengths. In the experiments in this thesis each of the textual segments are kept the same length within each experiment, to ensure that this artifact does not affect the results.

Pastor et al. (2008) refers to this measure as measuring the *lexical density* of a text, and find a statistically significant difference in the lexical density of a corpus of technical translations in Spanish and a corpus of comparable technical texts in Spanish. They also find a significant difference using this test for a comparable corpus of medical translations and originals in Spanish. In both cases the translated section of the corpus had a lower type-token ratio than the non-translated side. Li et al. (2011) found statistically significant differences in standardized type-token ratio between the work of the two translators examined in their study.

3.4.3 Lexical richness

Lexical richness is defined in Pastor et al. (2008) as the total number of tokens divided by the number of lemmas. In this case a lemma refers to the base form of a word, for example *student* and *students* have the same lemma, *student*. The lemma is obtained from the output of the TreeTagger (See Section 3.2.3).

$$\frac{\text{Total number of lemmas}}{\text{Total number of tokens}} \tag{3.17}$$

The difference in average lexical richness values between the two corpora in the work

by Pastor et al. (2008) mentioned in Section 3.4.2 above is also found to be statistically significant .

3.4.4 Information load

Information load is described in Ilisei et al. (2010) as the proportion of lexical words to total tokens. This is calculated by dividing the lexical words (nouns, verbs, etc) by the total tokens. Ilisei et al. (2010) define the lexical words as verbs, nouns, adjectives, adverbs and numerals. The same delineation is used when calculating the ratio.

$$\frac{\text{Total number of lexical tokens}}{\text{Total number of tokens}} \tag{3.18}$$

3.4.5 Average word length

The average word length is calculated by obtaining the total word length and then dividing it by the number of words. Ilisei et al. (2010) calculate word length by number of syllables but in this thesis the number of letters in the words are counted and the average value of this is taken instead. This feature plus average sentence length are also used in the readability metrics which are described in Section 3.5 below.

$$\frac{\text{Total length of all tokens}}{\text{Total number of tokens}} \tag{3.19}$$

3.5 Readability Metrics

This section describes two readability metrics which are used as features in the classification experiments. These tests were developed to predict the US grade level required to understand a piece of writing.

3.5.1 ARI

The *ARI*, or Automatic Readability Index was developed by the US military in the 1960's (Smith & Senter, 1967) as an enhancement of existing readability metrics such as the Flesch readability test , (Flesch, 1948), and Dale and Chall's formula, (Dale & Chall, 1948). It differed from many predecessors as it was the first readability metric which was designed to be used on an electric typewriter to provide feedback on the readability of a text being typed. Previous readability scores such as Flesch's and Dale and Chall's formulae counted the number of syllables in a word as word length, but the ARI differed in this case as it counted the number of characters in a word instead, which made automated data collection easier.

The use of characters to count word length is still useful nowadays even with modern programming languages and methods as it removes the need for a library of syllabification rules when writing a script to calculate the metric.

The ARI formula was derived based on technical documentation in English, and this is one caveat that should be taken into consideration when using the metric to predict reading level. However the experiments in this thesis deal with relative values of this metric calculated on comparable corpora of translated and original text in the same genre, so this is not a major cause for concern.

$$\text{ARI} = 4.71(\text{Average word length}) + 0.5(\text{Average Sentence Length}) - 21.43 \tag{3.20}$$

Pastor et al. (2008) report significant differences for the average ARI score for their comparable corpus of Spanish technical translations and original texts.

3.5.2 CLI

The *CLI* or Coleman-Liau Index is a readability metric developed by psychologists Meri Coleman and T Liau in the 1960's to detect the minimum US grade level required to successfully interpret a piece of writing. The test is similar to the ARI test described in Section 3.5.1 above in that it calculates word length as number of characters and is designed to be used by an automated system. However where the ARI is calculated using technical documentation as a reference, the CLI was calculated on a reference set of educational materials. As with the ARI, detecting the US grade level of a text is the not the main concern, instead the focus is on comparing relative values for this metric on different corpora, so although the metric was not trained on the same genres of text[8], it is believed that using this metric will still provide some insight between these different textual styles.

$$\text{CLI} = 5.89(\text{Average word length}) + 29.5\frac{\text{Number of Sentences}}{\text{Number of Words}} - 15.8 \tag{3.21}$$

Pastor et al. (2008) report statistically significant differences between the CLI values for the translated and original section of their corpus of medical texts, the translated side of which has been translated by student translators.

[8]In the case of the current research, parliamentary proceedings, newspaper articles and world literature

3.6 Sentence ratios

3.7 Introduction

This section describes the ratios which are calculated based on different sentence types. To calculate these, the text was first tagged by the Treetagger and then split into an array of sentences. Once in this form, the number of verbs in each sentence was counted and it was classified as either a simple sentence (one finite verb) or a complex one (more than one finite verb). Finite verbs are defined here as those tagged with the Penn tags *VBZ*, *VBD*, and *VBP*, corresponding to the 3rd person singular present form, past tense form and the non-3rd person singular form in English.

3.7.1 Ratio of simple sentences to complex sentences

$$\frac{\text{Number of simple sentences}}{\text{Number of complex sentences}} \tag{3.22}$$

This ratio quantifies the proportion of sentences with only one finite verb compared with sentences which contain more than one finite verb.

3.7.2 Ratio of simple sentences to total sentences

$$\frac{\text{Number of simple sentences}}{\text{Number of sentences}} \tag{3.23}$$

This ratio quantifies the ratio of sentences with only one finite verb compared to the total number of sentences. Pastor et al. (2008) report a statistically significant difference between the average value for this ratio between the translated and original sections of their professionally translated corpus of medical text and their professionally translated corpus of technical text in Spanish.

3.7.3 Ratio of complex sentences to total sentences

$$\frac{\text{Number of complex sentences}}{\text{Number of sentences}} \tag{3.24}$$

This ratio quantifies the proportion of sentences with more than one finite verb to the total number of sentences.

3.8 Other ratios

This section describes a number of ratios which are part of the document-level feature set.

3.8.1 Ratio of grammatical words to lexical words

This ratio can be described as the ratio of closed-class[9] words to open-class words in a text. As mentioned in Section 3.4.4 above, nouns, verbs, adverbs, adjectives and numerals are classified as members of the open class or lexical words and everything else is considered to be a member of the closed class or a grammatical word.

$$\frac{\text{Total number of grammatical words}}{\text{total number of lexical words}} \tag{3.25}$$

3.8.2 Ratio of prepositions to total words

This ratio counts the proportion of prepositions to total words. A preposition is defined here as any word with the tag *IN* from the Penn Treebank tagset which is used by the Treetagger. The preposition *to* is not counted in this ratio, as it is given special dispensation in the Penn tagset.

$$\frac{\text{Total number of prepositions}}{\text{total number of words}} \tag{3.26}$$

In their study on the Canadian Hansard corpus of translated French and English, together with original writing in both languages, Kurokawa et al. (2009) found that English translated from French contained a higher proportion of prepositions than original English.

3.8.3 Ratio of numerals to total words

This ratio counts the proportion of numerals in the text.

$$\frac{\text{Total number of numerals}}{\text{total number of words}} \tag{3.27}$$

3.8.4 Ratio of finite verbs to total words

This ratio counts the proportion of finite verbs in the text. Section 3.7 describes the finite verb tagging process.

$$\frac{\text{Total number of numerals}}{\text{total number of words}} \tag{3.28}$$

[9]The class of words for which it is generally impossible to add to, prepositions and determiners in English are an example, compared with the case of nouns and verbs where new members are regularly added.

3.8.5 Ratio of discourse markers to total words

This ratio calculates the ratio of a number of common English discourse markers to the total words in the corpus. These discourse markers counted are: *therefore*, *as a result*, *consequently*, *moreover*, *furthermore*, *in addition*, *however*, *nevertheless*, *on the other hand*, *while*, *whereas*, *with regard to*, *as regards* and *as for*.

$$\frac{\text{Total number of discourse markers}}{\text{total number of words}} \tag{3.29}$$

3.8.6 Ratio of pronouns to total words

$$\frac{\text{Total number of pronouns}}{\text{total number of words}} \tag{3.30}$$

This ratio quantifies the proportion of pronouns to total words.

3.8.7 Ratio of nouns to total words

$$\frac{\text{Total number of nouns}}{\text{total number of words}} \tag{3.31}$$

This ratio quantifies the proportion of nouns to total words.

3.8.8 Ratio of conjunctions to total words

$$\frac{\text{Total number of conjunctions}}{\text{total number of words}} \tag{3.32}$$

This ratio quantifies the proportion of conjunctions to total words.

3.9 Conclusion

This chapter has described the software packages used in the experimental chapters, along with the classification algorithms and features employed in the analysis. The relation between the metrics used in the experiments is important to note, the readability scores for instance use average word length and average sentence length, so one can imagine a relationship between these.

The ratio of grammatical words to lexical words is related to ratios of conjunctions to total words and prepositions to total words, as these items are members of the same class. Type-token ratio and lexical richness are related in the sense that one is similar to the other but perhaps more finely tuned, in the sense that the latter is concerned with the distribution of lemmas, where a number of inflected types are collapsed into one lemma, whereas the former is a simple version which counts plurals of nouns and past tense of verbs as different types from their singular or present tense counterparts.

Information load to some extent is inversely related to the ratio of grammatical to lexical items, as one ratio seeks to quantify the amount of information processing power necessary

to comprehend a text based on the proportion of content words, and the other one quantifies the amount of closed-class items in a text, a higher value for the latter would imply a lower value for the former. This can be observed in Chapter 4, where the original sections of both corpora have a higher mean value for information load and a lower mean value for the ratio of grammatical to lexical items.

Of course, the relationship between document-level features and ngram features is also of interest, high frequencies of certain POS bigrams will also have an effect on the proportional frequencies of broader categories, for example. Future work will implement more features such as the frequency of contractions in English, along with frequencies of textual phenomena such as the use of passive voice.

Chapter 4

Comparing translated and original text

Information	Translated	Original
Words	886694	839104
Texts	963	708
Source Languages	6	1

Table 4.1: Europarl subset

4.1 Introduction

This chapter describes experiments on a two comparable corpora of translated and original text in English. The features and classifiers detailed in Chapter 3 are used towards this end. Ngram feature sets and document-level feature sets are combined in order to examine their performance on the two corpora.

4.2 Corpora

4.2.1 Europarl

The Europarl corpus, (Koehn, 2005) is a parallel corpus in 11 languages which consists of transcripts of the proceedings of the European parliament. The corpus has been used in linguistics for machine translation research and systems training however it has also been used in some studies on translated text such as the work by van Halteren (2008) which attempts to detect the source language of a translated text given its translation into multiple languages. After extracting the XML markup from the corpus, an English subsection was selected which comprised of parliamentary contributions from the year 2005 with both original English and translations from Greek, Czech, Danish, Spanish, German and Finnish.

4.2.2 New York Times corpus

The NYT corpus is a small hand-assembled[1] comparable corpus of New York Times Opinion-Editorial articles spanning the period 1993-2010. The articles were taken from the contributors to the Opinion pages of the NYT, rather than regular columnists. The translated side of the corpus contained texts which were translated from 16 source languages: 22 from Spanish, 17 from German, 15 from Russian, 14 from French, 8 from Hebrew, 6 from Italian, 3 each from Polish and Japanese, 2 each from Czech, Chinese and Arabic and 1 each from Uighir, Dutch, Farsi, Icelandic and Korean. The articles cover a range of topics on global affairs. When compiling the comparative side of the corpus, one of the main compilation criteria was to try and avoid any topic-based classification bias where possible, articles were chosen which described global affairs rather than topics which were overly US-centric. A list

[1] A search was carried out on the NYT homepage for the text "* translated by *", as no easier method could be found to search for translated articles only.

Information	Translated	Original
Words	90278	87773
Texts	101	101
Source Languages	16	1
Unique Authors	69	97
more than 1 article	16	5

Table 4.2: NYT corpus

of article titles, original publication dates and author and translator information is available in Appendix A.1.

There are several issues to bear in mind regarding the NYT corpus. Firstly, it is quite small, the Europarl section under investigation contains in the order of ten times as much text. Secondly, the fact that the text is drawn from opinion articles raises some issues regarding bias on issues, different viewpoints, etc. At the same time, care was taken to have comparable articles on similar topics and from similar viewpoints, and on the subject of topic it can be argued that it is in fact more heterogeneous than taking translations and non-translations from a source such as Europarl, where there is a clear geo-political line drawn between those texts which are native English, representing contributions from UK and Irish members of the European Parliament, and translations which represent the views of parliamentarians from other member states.

4.3 Experimental setup

The experiments compare results for document-level and ngram-based features on both datasets. The document-level features were described in Section 3.4. In order to keep file sizes balanced, which is important for calculating metrics such as type-token ratio, a 2k sample was taken from each file in the two datasets.

Word unigrams, bigrams and POS bigrams were computed for the datasets using the TagHelperTools package. The frequency of each of these tokens was reduced to a binary value, either the feature occurs in a textual chunk and thus has value 1 or the feature does not occur and has value 0. This is done automatically by TagHelperTools.

In this experiment, the goal of the classifier is to detect if a textual chunk is or is not a translation. Results of the experiments carried out on the Europarl corpus are in Section 4.4.1 and results for the NYT corpus in Section 4.4.2.

Ten-fold cross validation was used in all of the experiments with all feature selection being carried out on the training set within each fold, independent of any other iterations.

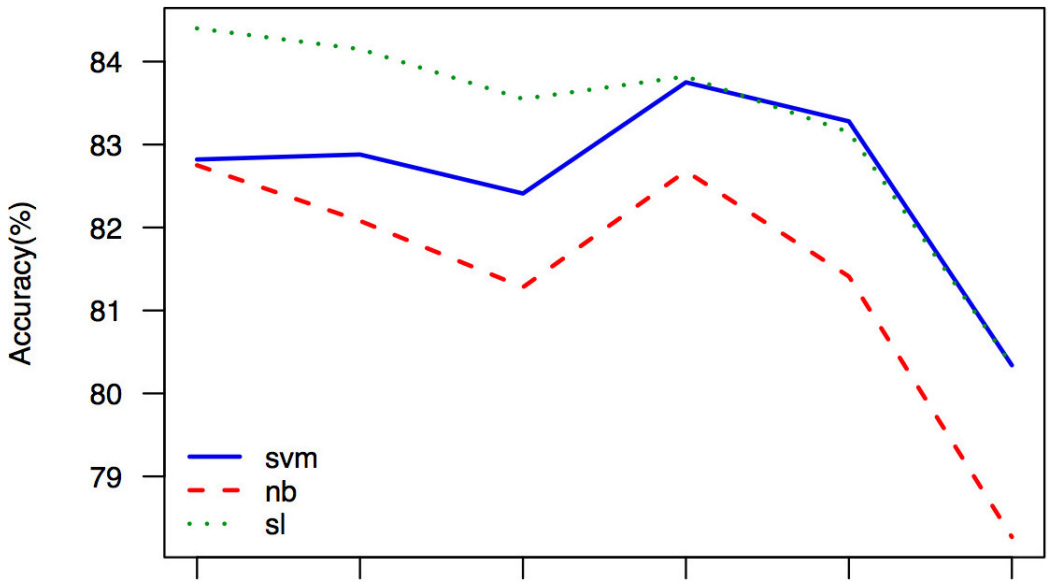

Figure 4.1: Classification results on Europarl corpus: Word unigrams(Top 500-50)

4.4 Single-feature sets

This section describes results carried out on single-feature sets. Section 4.4.1 details the results of the experiments on the Europarl comparable corpus and Section 4.4.2 describes results on the New York Times comparable corpus.

4.4.1 Results on Europarl subset

Algorithm	Features	Test Set	Accuracy
Baseline	n/a	10f cv	57%
SVM	15doc	10f cv	76%
SimpLog	15 doc	10f cv	77%
NaiveBayes	15doc	10f cv	71%
SVM	13 doc	10f cv	76%
SimpLog	13 doc	10f cv	**78%**
NaiveBayes	13doc	10f cv	71%

Table 4.3: Classification results on Europarl corpus: Document-level features

Taking the results in Table 4.3, the best performance is obtained by the Simple Logistic classifier using 13 document level features[2], The next best performance is for the Simple Logistic classifier using a subset of 12 of the original document-level features.

Results using single feature sets were more varied in nature, using word unigrams, the Simple Logistic classifier obtains the highest accuracy using the top 500 features[3] in Figure

[2]This was all of the features from Section 3.4 minus the preposition ratio, finite verb ratio and the three sentence ratios.

[3]The scale used is 500-400-300-200-100-50

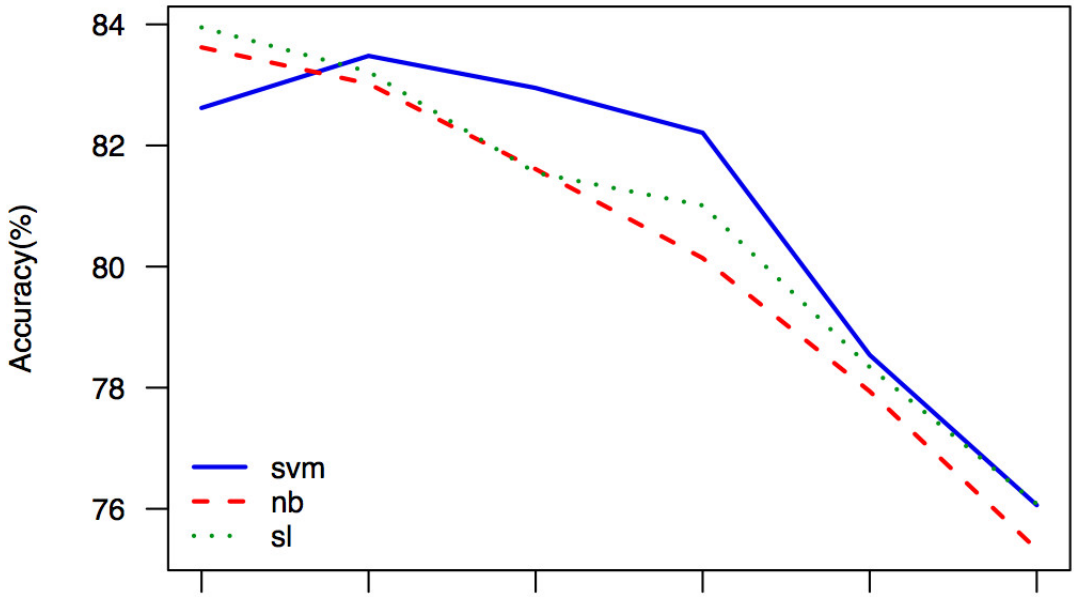

Figure 4.2: Classification results on Europarl corpus: Word bigrams(Top 500-50)

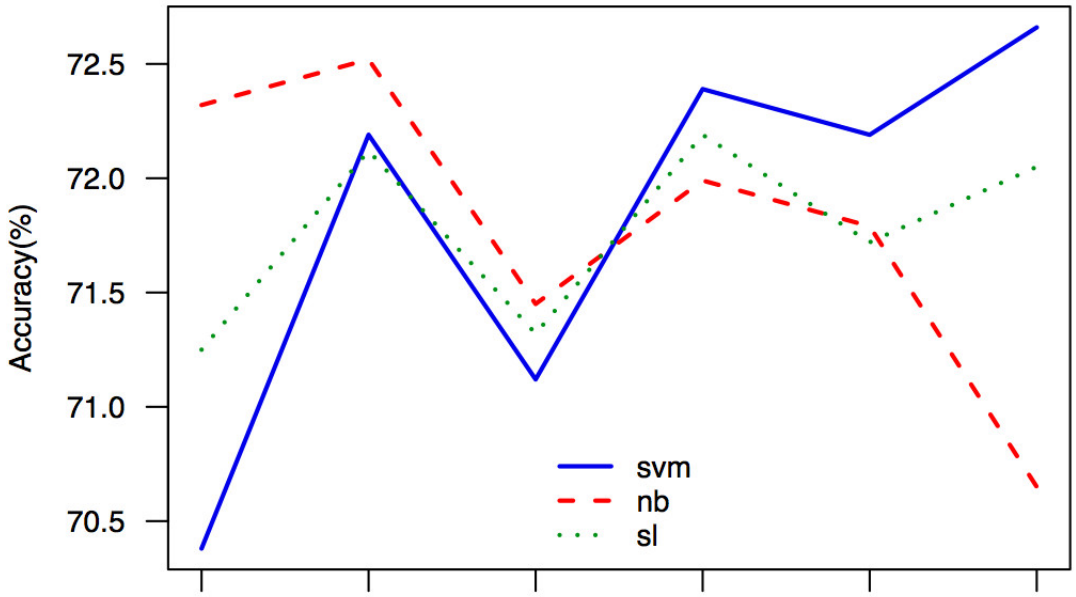

Figure 4.3: Classification results on Europarl corpus: POS bigrams(Top 500-50)

4.1, with Naive Bayes and SVM also performing well. Classification accuracy decreases when the top 50 features are used.

Word bigram features also display high accuracy scores for the Simple Logistic and Naive Bayes classifiers, also exhibiting a similar drop in accuracy when less features are used.

Classification results for POS bigrams are less accurate than those using word features, these classifiers exhibit an increase in accuracy for the SVM and Simple Logistic classifier as less features are used.

Chi	Rank	Token	Chi	Rank	Token
82.3781	1	IN-WDT	16.5416	27	TO-NNP
64.6993	2	BOL-RB	16.3362	28	BOL-CC
52.5642	3	BOL-WP	15.27	29	NN-WDT
52.1311	4	TO-WDT	15.0507	30	WDT-VBZ
48.098	5	WDT-PRP	15.0322	31	PRP-NNPS
45.5846	6	CC-IN	14.8381	32	NNS-WDT
42.2	7	WP-VBZ	14.0733	33	NN-RP
39.5747	8	PRP-TO	13.8971	34	NNP-NN
36.1709	9	IN-PRP	13.3627	35	PRP-MD
35.6263	10	WDT-DT	13.1741	36	IN-VBG
34.4358	11	NNP-NNS	13.0841	37	PRP-JJ
33.9766	12	NNPS-NNP	12.9548	38	VB-CC
33.1218	13	WP-PRP	12.822	39	MD-IN
30.4684	14	TO-PRP	12.7125	40	CD-NNP
27.5363	15	IN-WP	12.6464	41	VB-PRP
25.0598	16	NNP-CD	12.2605	42	BOL-JJ
24.5335	17	CC-RB	11.5192	43	NN-NNS
22.9346	18	VBG-PRP	11.3275	44	NNP-VBZ
21.7263	19	VBG-VBN	10.8428	45	DT-PRP
21.1651	20	NNP-RB	10.7519	46	DT-DT
20.4619	21	WDT-NN	10.6054	47	VB-VBZ
19.7712	22	VBN-RP	10.4757	48	RP-IN
18.9007	23	WDT-VBP	10.3522	49	IN-EX
18.7717	24	VBZ-PRP			
18.1436	25	PRP-IN			
16.6653	26	NNP-TO			

Table 4.4: POS bigrams: Europarl

The word bigram features in Table 4.6 display some examples of the features found by van Halteren (2008) to discriminate between different source languages. These include bigrams such as *ladies and* and *and gentlemen* and other forms of address. According to van Halteren (2008), speakers from English speaking countries address the President only, while speakers from other European countries address the congregation as a whole, hence the more frequent presence of *ladies and gentlemen* in the translated section of the corpus and *Mr President* in the original English sections. One example bigram, *this house*, occurred approximately four times as often in the translated side of the corpus.

Chi	Rank	Token	Chi	Rank	Token
107.6128	1	though	27.6135	24	ireland
87.1787	2	must	27.5415	25	strong
65.6991	3	uk	27.3325	26	including
59.5496	4	something	27.132	27	context
54.6008	5	house	26.8544	28	ladies
54.2136	6	things	26.4074	29	christian
51.2945	7	s	26.2589	30	strongly
50.64	8	which	25.8308	31	being
48.3795	9	fully	25.3697	32	recognise
40.7936	10	reason	24.9427	33	means
39.4264	11	eu	24.9136	34	say
38.6467	12	thing	24.8342	35	will
34.5427	13	gentlemen	24.6976	36	quite
32.5717	14	them	24.6064	37	group
32.5538	15	remarks	24.4142	38	makes
32.1385	16	what	23.9052	39	do
31.4694	17	colleagues	23.7423	40	nothing
31.1149	18	fact	23.5048	41	however
29.7856	19	believe	23.1998	42	social
29.5224	20	welcome	23.0906	43	2005
29.2823	21	commission	22.3993	44	look
29.0368	22	greater	21.9867	45	continue
28.9053	23	commitment	21.8105	46	make

Table 4.5: Word unigrams: Europarl

4.4.2 Results on NYT corpus

This section describes classification results on the NYT corpus. Examining Figures 4.4, 4.5 and 4.6, the highest accuracy result is 64% obtained by the Naive Bayes classifier with 300 POS bigrams as the feature set. Classification results are lower than on the Europarl corpus in all cases.

Table 4.7 shows the results using document-level features with the highest classification accuracy obtained using the six document-level features in Table 4.8 and the Support Vector Machine classifier.

Chi	Rank	Token	Chi	Rank	Token
60.34738	1	the-uk	27.82083	25	christian-democrats
55.70667	2	this-house	27.13731	26	people-s
53.45793	3	is-that	27.03993	27	that-this
52.56422	4	and-that	26.66954	28	must-be
44.12823	5	to-which	26.34364	29	for-this
40.22836	6	must-not	26.00414	31	mr-president
38.16676	7	means-of	26.00414	30	european-democrats
37.66364	8	and-so	25.39752	32	do-with
36.77482	9	by-means	25.24867	33	the-same
34.86556	10	and-gentlemen	25.12737	34	other-words
34.74922	11	believe-that	25.09149	35	that-the
34.08721	12	which-we	24.74349	36	for-instance
33.61738	13	ladies-and	24.5856	37	of-eu
32.46899	14	in-which	23.96612	38	of-course
32.27437	15	commission-will	23.6636	39	to-say
32.0673	16	that-reason	23.65825	40	in-this
31.6727	17	we-must	23.4748	41	group-of
31.45702	18	but-also	23.4078	42	the-fact
30.5371	19	recognise-that	23.36006	43	uk-presidency
29.88205	20	ready-to	23.12947	44	is-something
29.85748	21	the-group	23.0578	45	the-eu
29.46165	22	reason-that	22.77668	46	in-future
28.67174	23	like-to	22.72297	47	i-believe
28.40656	24	for-it	22.13111	48	is-where

Table 4.6: Word bigram features: Europarl

Algorithm	Features	Test Set	Accuracy
SVM	19doc	10f cross-v	66.4773%
Simplog	19doc	10f cross-v	63.63%
Bayes	19doc	10f cross-v	65.91%
SVM	6doc	10f cross-v	69.3182%
Simplog	6doc	10f cross-v	67.04%
Bayes	6doc	10f cross-v	68.75%

Table 4.7: Classification Results on NYT corpus: Doc-level feature sets

Rank	Chi	Feature
1	17.76824	nounratio
2	16.77447	grammlex
3	16.62782	infoload
4	16.11093	avgwordlength
5	16.07963	fverbratio
6	12.00307	cli

Table 4.8: Top ranked document features in 10-fold cross validation on NYT corpus

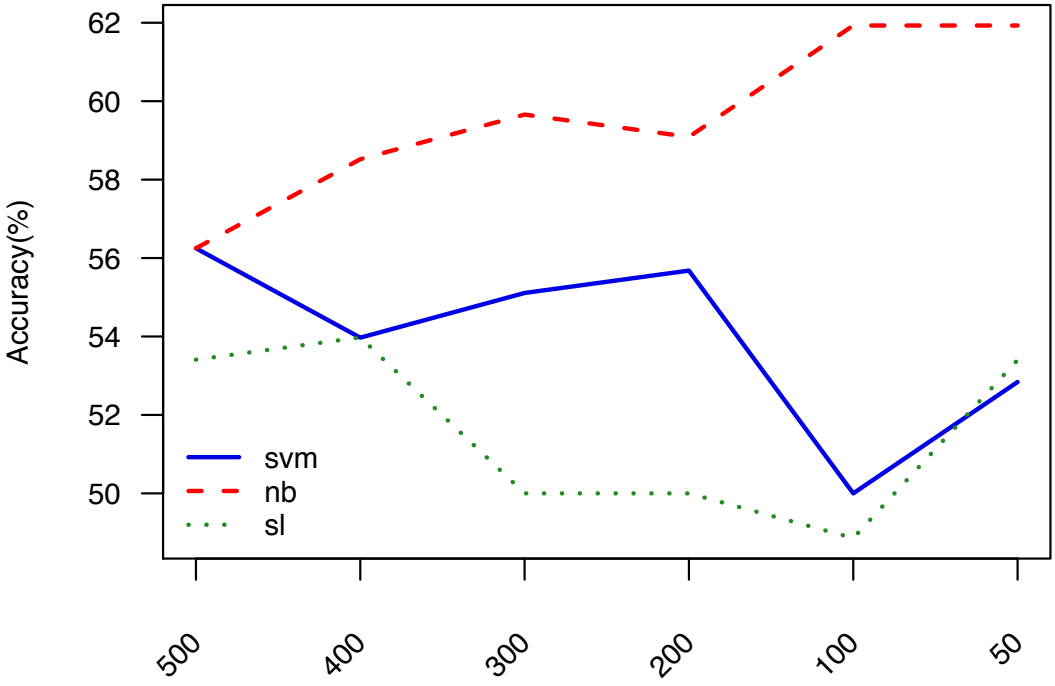

Figure 4.4: Classification results on NYT corpus: Word unigrams(Top 500-50)

4.5 Combined Feature Sets

The next step after comparing the two types of feature-sets was to combine both sets into a single classifier to examine whether this increased classification accuracy.

This section presents results of experiments conducted on combined feature sets. The first section examines the results on the Europarl Corpus and the second section examines the NYT corpus.

4.5.1 Europarl

Figure 4.7 displays the accuracy results for mixed feature sets on the Europarl corpus. The highest accuracy is obtained by the Simple Logistic classifier using 500 mixed features with just over 88% accuracy. Table 4.14 displays a number of mean values for the highly-ranked document-level features. The non-translated section has a higher CLI score, but interestingly

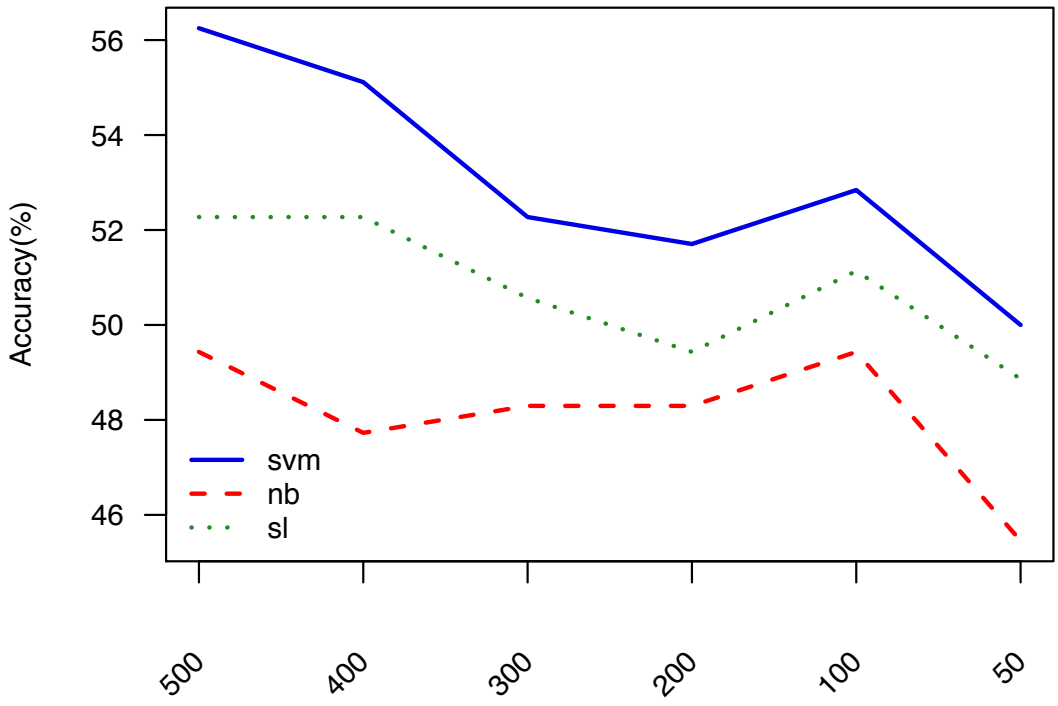

Figure 4.5: Classification results on NYT corpus: Word bigrams(Top 500-50)

a lower ARI score, a lower average sentence length, a higher value for nounratio, information load and average word length, a lower proportion of closed-class to open-class words and a lower ratio of finite verbs to total words.

4.5.2 NYT Corpus

Figure 4.8 displays classification accuracy for the mixed feature set on the NYT corpus. The best result obtained was 65% with the Naive Bayes classifier and the feature set containing 50 features only.

Table 4.18 displays mean values for a number of the document-level ratios occurring in Table 4.16. The original side of the corpus has a higher average CLI score and ARI score, a higher ratio of nouns to total words, a higher value for information load, a higher average word length, a lower ratio of closed-class items to open-class items and a lower average ratio of finite verbs to total words. In this case the ARI and CLI give a more comparable result than on Europarl however it must be remembered that the ARI and CLI metrics were both developed on different textual corpora. With the exception of the ARI value, the relationships between the mean values of the metrics for the translated and original sections of the corpus are comparable to the results on the Europarl corpus.

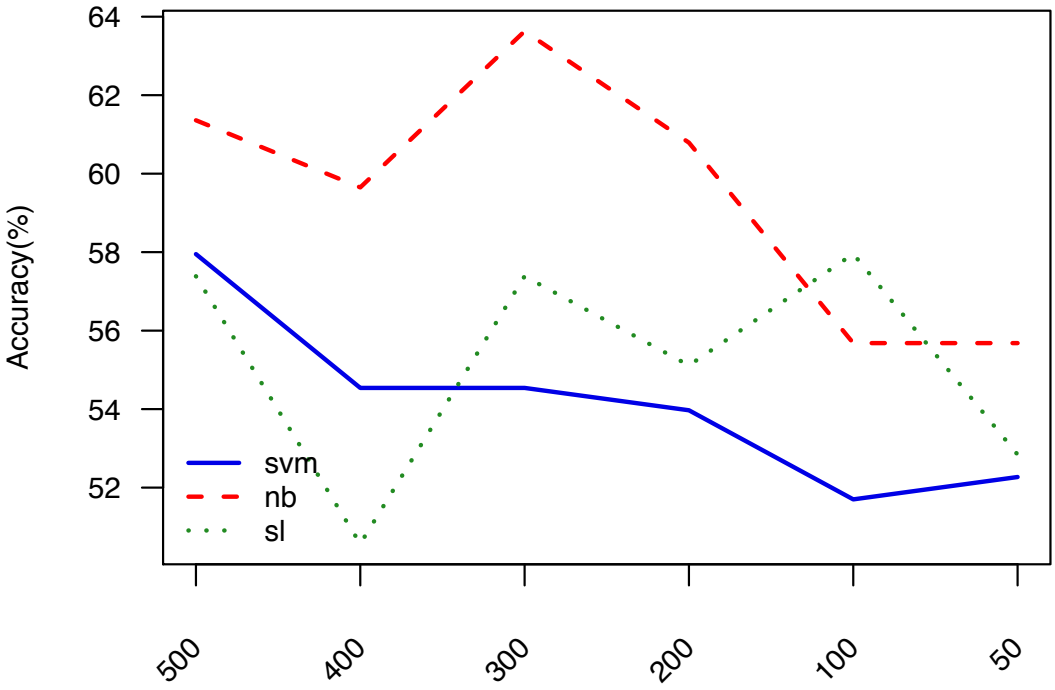

Figure 4.6: Classification results on NYT corpus: POS bigrams(Top 500-50)

Chi	Rank	Token	Chi	Rank	Token
12.7772	1	community	6.0722	17	canada
10.3642	2	decade	6.0722	18	hearts
10.3472	3	bear	5.9387	19	legitimate
9.1036	4	possiblity	5.9211	20	realized
9.0948	5	always	5.9211	21	diplomacy
8.4901	6	enemy	5.5721	22	expanding
7.6068	7	u	5.5721	23	neutral
7.5142	8	domestic	5.5721	24	desperate
7.4575	9	fund	5.5721	25	militant
7.0218	10	environment	5.5308	26	pages
6.8304	11	weather	5.2645	27	experienced
6.8304	12	perspective	5.2645	28	chaotic
6.3546	13	word	5.2645	29	tries
6.0722	14	dealing	5.0306	30	totalitarianism
6.0722	15	asia	5.0306	31	illusion
6.0722	16	channels			

Table 4.9: Word unigram features: NYT

Rank	Chi	Feature	Rank	Chi	Feature
1	9.7832	people-who	16	6.0722	because-we
2	9.2719	state-of	17	6.0722	to-him
3	8.1926	an-american	18	6.0722	our-country
4	8.0177	an-u	19	6.0722	in-september
5	7.173	and-who	20	6.0722	in-power
6	7.1261	the-possibility	21	5.9211	tries-to
7	7.1261	possibility-of	22	5.9075	has-taken
8	6.636	vote-in	23	5.9011	people-will
9	6.595	these-two	24	5.5721	the-enemy
10	6.3546	other-side	25	5.5721	rhetoric-of
11	6.3546	dealing-with	26	5.5308	a-woman
12	6.3546	fears-of	27	5.2645	and-israel
13	6.3546	made-it	28	5.2645	engaged-in
14	6.1832	they-should	29	5.2645	does-the
15	6.0722	it-seemed			

Table 4.10: Word bigram features: NYT

Chi	Rank	Token	Chi	Rank	Token
13.11472	1	JJS-NNS	4.18712	20	MD-CC
11.50689	2	RP-IN	4.00105	21	MD-NN
10.84326	3	TO-PRP	4.00105	22	MD-EX
9.27188	4	PRP-NNPS	4.00105	23	MD-CD
8.70197	5	NN-NNS	4.00105	24	MD-NNPS
8.01005	6	TO-CC	3.12218	25	NNPS-CD
7.60534	7	NN-VBZ	3.12218	26	MD-NNS
7.34187	8	CD-WP	2.98344	27	NNPS-DT
7.02185	9	JJS-VBZ	2.98344	28	NNPS-FW
6.80131	11	NNPS-NNPS	2.98344	29	MD-PRP$
6.80131	10	NNP-EX	2.98344	30	MD-PRP
6.62813	12	JJS-VBD	2.98344	31	NNP-JJ
6.62813	13	MD-IN	0	32	NNPS-VBN
6.61406	14	JJS-VB	0	33	VBN-NNPS
6.35456	15	JJS-PRP	0	34	NNPS-JJ
5.90751	16	MD-NNP	0	35	NNPS-JJR
5.90751	17	MD-JJS	0	36	JJ-WP$
5.90751	18	MD-DT	0	37	JJ-VBG
4.18712	19	MD-JJ	0	38	PRP-VBP

Table 4.11: POS bigram features: NYT

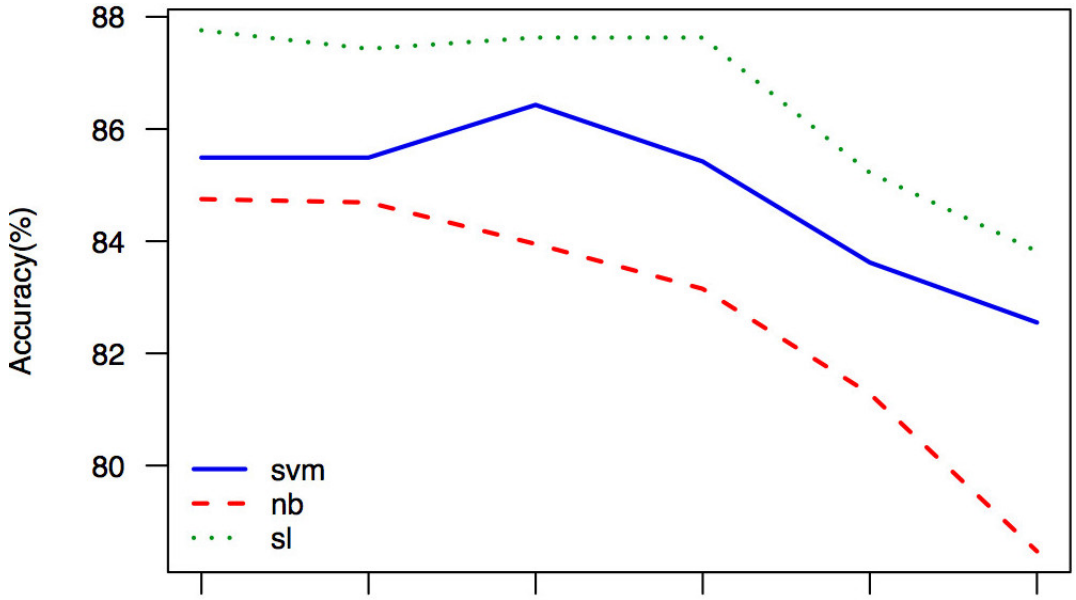

Figure 4.7: Classification results on Europarl corpus: Mixed features(Top 500-50)

Chi	Rank	Token	Chi	Rank	Token
391.7425	1	**avgsent**	64.6993	13	BOL-RB
343.6669	2	**nounratio**	60.3474	14	the-uk
183.5639	3	**avgwordlength**	59.5496	15	this-house
134.8823	4	**grammlex**	55.7067	16	house
129.388	5	though	55.0376	17	**typetoken**
122.6182	6	**infoload**	54.6008	18	is-that
107.6128	7	must	54.2136	19	things
89.1623	8	IN-WDT	53.4579	20	s
87.1787	9	**ari**	52.5642	22	which
82.3781	10	**conjratio**	52.5642	21	and-that
76.1131	11	uk	52.1311	23	TO-WDT
65.6991	12	something	51.2945	24	to-which

Table 4.12: Mixed features 1-24: Europarl

Chi	Rank	Token	Chi	Rank	Token
50.64	25	BOL-WP	37.6636	38	remarks
48.3795	26	fully	36.7748	39	PRP-TO
48.098	27	reason	36.1709	40	what
45.5846	28	WDT-PRP	35.6263	41	and-so
44.1282	29	thing	34.8656	42	by-means
43.8551	30	eu	34.7492	43	and-gentlemen
42.2	31	CC-IN	34.5427	44	IN-PRP
40.7936	32	**pnounratio**	34.4358	45	WDT-DT
40.2284	33	must-not	34.0872	46	colleagues
39.5747	34	WP-VBZ	33.9766	47	which-we
39.4264	35	gentlemen	33.6174	48	believe-that
38.6467	36	them	33.1218	49	ladies-and
38.1668	37	means-of	32.5717	50	NNPS-NNP

Table 4.13: Mixed features 25-50: Europarl

Metric	Originals	Translations
ari	13.53522	16.31690
cli	13.50766	12.69592
avgsent	23.05808	29.91967
nounratio	0.2785114	0.2572843
infoload	0.5895204	0.5717544
avgwordlength	4.545422	4.385321
grammlex	0.5741338	0.6228523
fverbratio	0.07543852	0.08192503

Table 4.14: Mean values of document-level ratios on EP corpus: Translated section vs. original section

Metric	Originals	Translations
ari	3.091036929	4.5008444035
cli	1.7400681021	1.5634050773
nounratio	0.037781662051	0.032441605396
infoload	0.0312899554770662	0.0292974401305943
avgwordlength	0.29516631897	0.27026351096
grammlex	0.0788840985043305	0.0796222951611805
fverbratio	0.0173336080	0.017330076832

Table 4.15: Standard deviations for document-level ratios on EP corpus: Translated section vs. original section

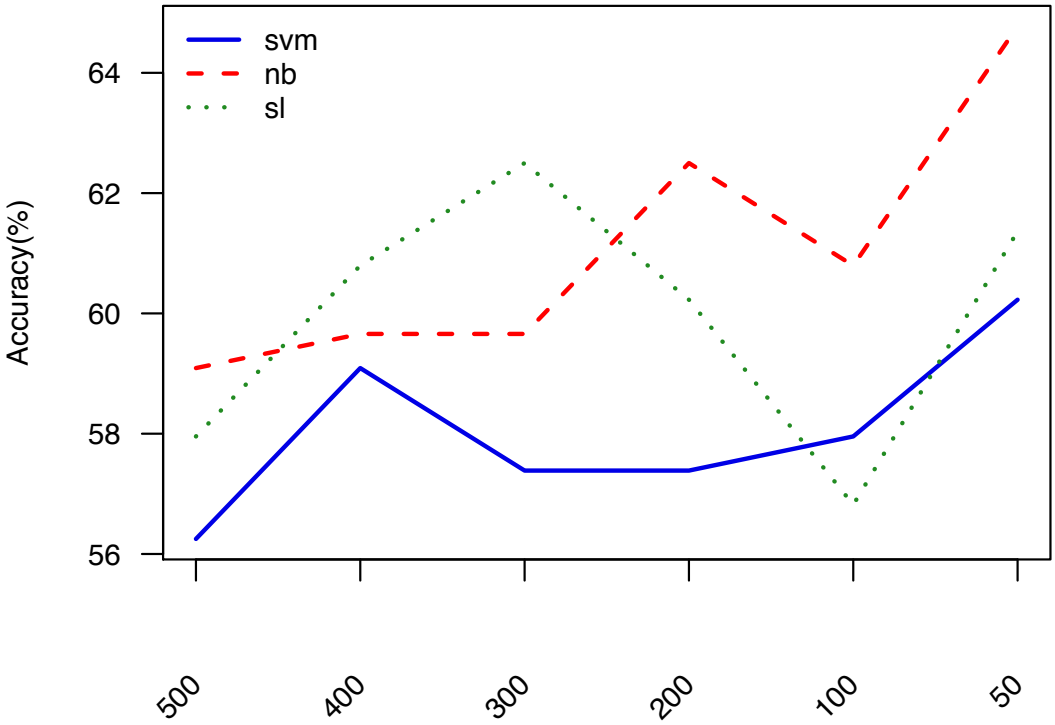

Figure 4.8: Classification results on NYT corpus: Mixed features(Top 500-50)

Chi	Rank	Token	Chi	Rank	Token
17.76824	1	**avgwordlength**	9.78316	13	always
16.77447	2	**grammlex**	9.27188	14	and-who
16.62782	3	people-who	9.27188	15	*PRP-NNPS*
16.11093	4	community	9.10359	16	enemy
16.07963	5	*JJS-NNS*	9.0948	17	*VBD-WDT*
13.11472	6	state-of	8.70197	18	*NN-NNS*
12.77717	7	decade	8.49012	19	u
12.00307	8	*RP-IN*	8.19262	20	the possibility
11.50689	9	an-american	8.01774	21	domestic
10.84326	10	the-u	8.01005	22	environment
10.36415	11	bear	7.6068	23	fund
10.34725	12	*TO-PRP*	7.60534	24	*NN-NNS*

Table 4.16: Mixed features 1-24: NYT

Chi	Rank	Token	Chi	Rank	Token
7.51422	25	perspective	6.62813	39	fears-of
7.45753	26	weather	6.61406	40	*NNS-RBS*
7.34187	27	possibility-of	6.59499	41	elsewhere-in
7.173	28	asia	6.35456	42	dealing-with
7.12612	29	dealing	6.35456	43	expanding
7.12612	30	word	6.35456	44	militant
7.02185	31	channels	6.35456	45	made-it
7.02185	32	vote-in	6.35456	46	they-should
6.83038	33	hearts	6.35456	47	desperate
6.83038	34	canada	6.1832	48	it-seemed
6.80131	35	other-side	6.07217	49	in-september
6.80131	36	legitimate	6.07217	50	because-we
6.63602	37	realized			
6.62813	38	diplomacy			

Table 4.17: Mixed features 25-50: NYT

Metric	Originals	Translations
cli	13.23897	12.32823
ari	12.71631	11.88888
nounratio	0.2983284	0.2801902
infoload	0.6221567	0.6038150
avgwordlength	4.352567	4.197375
grammlex	0.5117812	0.5549575
fverbratio	0.07543852	0.08192503

Table 4.18: Mean values of document-level ratios on NYT corpus: Translated section vs. original section

Metric	Originals	Translations
cli	1.37866508768	1.2719044426
ari	2.44964093983	2.63598611497
nounratio	0.0262955435061	0.0297954215896
infoload	0.0266335727235	0.0271317732667
avgwordlength	0.25263257874562	0.231625442660999
grammlex	0.0579202012711	0.0617685456617
fverbratio	0.0149594925618	0.013947735506

Table 4.19: Standard deviations for document-level ratios on NYT corpus: Translated section vs. original section

4.6 Cross corpus experiments

This section describes experiments similar to those described by Koppel and Ordan (2011), where one genre of text was used for training and one for testing. In these experiments, Europarl is used as the training set and the NYT training set is used as the test set. Just as in Koppel and Ordan (2011), the results are poor, hardly an improvement on the baseline, Table 4.20 shows the results for the three classifiers.

Algorithm	Features	Test Set	Accuracy
SVM	13doc	NYT	50%
SimpLog	13doc	NYT	54%
NaiveBayes	13doc	NYT	55%

Table 4.20: Results for cross-corpus experiments with Europarl as a training set

4.7 Discussion

In general, the experiments using combined feature sets for Europarl report comparable results with the work by Baroni and Bernardini (2006) on an comparable corpus of Italian current affairs articles and Ilisei et al. (2010) who examined a comparable corpus of Spanish technical text, although not quite as high as Koppel and Ordan (2011), who used the frequency of the three hundred most common words as the sole feature set in the experiments. Classification on the NYT corpus was not significantly improved with the combination of both feature types, with a subset of six document level features resulting in the highest accuracy result of 69% using an SVM classifier.

Comparing the results obtained by single-feature sets and mixed sets in Section 4.5 and Section 4.3, classification results are improved on the Europarl corpus using combined sets and not significantly improved on the NYT corpus. The difference in corpus size and consistency is likely to account for this trend, as evidenced by the studies carried out by Koppel

and Ordan (2011) and the experiments in Section 4.6, which show that even with document level features, poor classification accuracy is obtained by training on Europarl and testing on the NYT corpus.

Examining Tables 4.12 and 4.13 with their corresponding tables for the NYT corpus, Tables 4.16 and 4.17, it is interesting to measure frequencies across corpora.

The word bigram *believe that* is notable, occurring 57 times in the translated section of the NYT corpus and 21 times in the original section, compared with 225 occurrences for the original Europarl section and 741 times in the translated section. This bigram could be classified as a complementizer *that* construction similar to those examined by Olohan (2001), whose work is detailed Section 2.2.3. In this case the *that* is an optional item, whose removal would make no semantic or grammatical difference to a sentence. However, it is of interest to compare the top ten features in the combined experiments with the highest classification accuracy for each corpus to compare trends across corpora. Table 4.21 compares the ten highest ranked features based on aggregate ranks from 10 fold cross validation from each of the combined experiments alongside results from Ilisei et al. (2010).

Europarl		NYT		Ilisei et al 2010	
1	*avgsent*	1	**avgwordlength**	1	lexrichness
2	*nounratio*	2	**grammlex**	2	**grammlex**
3	**avgwordlength**	3	people who	3	*fverbratio*
4	**grammlex**	4	community	4	numratio
5	though	5	JJS NNS	5	adjratio
6	infoload	6	state of	6	*avgsent*
7	must	7	decade	7	*pnounratio*
8	IN-WDT	8	RP-IN	8	**avgwordlength**
9	ari	9	an american	9	simplesentences
10	conjratio	10	the u	10	zerosentences

Table 4.21: Overview of distinguishing features

Table 4.21 presents some interesting correlations. The first is that document-level features dominate the top ten features in both of the corpora examined in this experiment, although there are a number of POS bigrams and word unigrams also present. The features in bold are common across all three corpora, each of different genre and in one case, language. Two features, average word length and proportion of closed-class to open-class words are common across all three corpora. Features in italics are common to at least two of the experiments, with readability measures featuring in the Europarl top-ten.

The word *though* features in the Europarl top ten, and is actually more frequent in the translated side of the corpus, along with the POS tag *IN WDT* which corresponds to a preposition *IN* + a determiner such as *which* or *who*. This poses a number of possibilities however, as in the WSJ tagset which is used here, *IN* can refer to a subordinating conjunction or a preposition. The top ten features from the NYT corpus contain the POS bigram *RP IN* which corresponds a particle plus subordination conjunction or preposition, one example could be

of that[4]. Also the bigrams *JJS NNS* which is a superlative adjective and plural noun and *TO PRP* which is the preposition *TO* plus a personal pronoun, one example of this could be the construction *to me*.

Although there are a number of distinguishing features in both sets[5] which are related to the topics and themes contained within, it is heartening to note that many of the most robust features are in fact POS tags and bigrams of common words, indicated that topic-based classification does not in fact account for the majority of the classification of translated vs. original text.

4.8 Conclusion

This chapter has examined two distinct feature types for classifying translated text from original text. Document-level statistics and ngram features were compared over two monolingual comparable corpora of translated text in English, together with a hybrid classifier which used the highest-ranked features from each feature-set, resulting in an increase in classification accuracy on the Europarl corpus, with the combined featured set not contributing to a significant gain in classification accuracy on the corpus of New York Times Opinion Editorial Contribution articles.

Accuracy results on the NYT corpus were quite low, which may be due to the diverse nature of the articles and the various source languages in the corpus, but also could be a factor of corpus size. Document-level features performed slightly better however, indicating that higher-level trends may indeed prevail despite the small size of the corpus.

Document-level statistics featured heavily in the top-ten feature list of the combined feature set (See Table 4.21) with for each corpus as ranked by the chi-squared metric in Weka. Comparing these with the top ten features in Ilisei et al. (2010) found some correlations between the features. Average word length and ratio of open-class to closed-class words were two features common to both corpora in this study and in the work by Ilisei et al. (2010) which examined Spanish text.

Future work will use more document-level features from the literature such as *perplexity*, *entropy* and ratios of contracted items in an attempt to improve classification of translated text and to identify characteristics that distinguish translated text from original text in English. More comparable corpora shall be examined in future experiments in an attempt to ascertain features of translated text in English which are corpus-independent.

[4]Occurring 60 times in the translated section of the corpus but only 24 in the untranslated section.

[5]Examples include *UK* and *Germany* in the Europarl word unigram set and *US*, *Asia* and *Canada* in the NYT features.

Chapter 5

Source language markers in literary translations

5.1 Introduction

This chapter focuses on experimentation towards the detection of source language influence in English literary translations from the nineteeth and early twentieth century. A corpus of novels has been assembled from this time period which consists of fifteen translations, five each from Russian, German and French, and five works written originally in English[1]. Cross validation experiments are carried out on the corpus to determine robust features which identify the L1 of the texts.

Document-level metrics such as sentence length and readability scores are calculated together with ngram-based features such as the frequency of POS tags and closed-class words, features which are not directly related to the topics and themes contained within the texts. The aim is to identify features that are more general than the work in itself: the purpose of the experiments is not to correctly attribute texts to their translator or to original author so much as to L1.

In order to minimize the effect of authorial or indeed translatorial style in this study, no more than one work by the same author or translator has been selected.

Ten-fold cross validation was used to obtain accuracy results, with all feature selection performed within each fold. The tables of distinguishing features are based on aggregated rankings over each of the folds.

The temporal span of the language in the corpus is the latter half of the nineteenth century. In the case of the translations, there are several which did not appear in print until the early twentieth century. Criteria for selection were as follows:

1. text should be available in an machine-readable format and in the public domain.

2. from the previous point, this dictates that text will most likely stem from prior to the early twentieth century, due to US copyright law.

3. each text should have a unique author and in the case of translations, translator, i.e. no repeated authors or translators.

4. text should be of sufficient length, at least two hundred kilobytes in size, i.e. preferably a novel or novella.

In many cases, particular translators had translated numerous works by a single author and indeed also occasionally by several authors. Thus, it was necessary to obtain a configuration of texts which allowed each author and translator to remain unique.[2]

The list of texts is presented in Table 5.1. Texts were sourced from Project Gutenberg.[3]

[1] Henceforth the source language of the text will be referred to as the L1, borrowing from the language acquisition literature

[2] This was more complicated for Russian, for example, with the translator Constance Garnett having translated works by Tolstoy, Dosteyevsky and Turgenev, amongst others, resulting in the bypassing of a title of such repute as *Anna Karenina* for the less well-known novella *The Cossacks* by Tolstoy, due to the fact that Garnett was already represented as the sole available translator of Turgenev.

[3] www.gutenberg.org, last verified May 7, 2013

5.1.1 Corpus

Title	Author	Source	Date pub.	Translator	Translation pub.	Person
Great Expectations	Charles Dickens	English	1861	n/a	n/a	1st
The Picture of Dorian Gray	Oscar Wilde	English	1891	n/a	n/a	3rd
Jude the Obscure	Thomas Hardy	English	1895	n/a	n/a	3rd
Treasure Island	R.L Stevenson	English	1883	n/a	n/a	1st
Middlemarch	George Eliot(M. Evans)	English	1874	n/a	n/a	3rd
The Idiot	Fyodor Dostoyevsky	Russian	1869	Eva Martin	1915	3rd
The Man Who Was Afraid	Maxim Gorky	Russian	1899	Hermann Bernstein	1901	3rd
Fathers and Children	Ivan Turgenev	Russian	1862	Constance Garnett	1917	3rd
The Cossacks	Leo Tolstoy	Russian	1863	Louise and Alymer Maude	n/a	3rd
A Man of our Time	Mikhail Lermontov	Russian	1841	J.H Wisdom and Marr Murray	1917	1st
The Count of Monte Cristo	Alexandre Dumas	French	1844	Anon	1846	3rd
Madame Bovary	Gustave Flaubert	French	1857	Eleanor Marx-Aveling	1898	3rd
Fr Goriot	Honoré de Balzac	French	1853	Ellen Marriage	1901	3rd
The Hunchback of Notre Dame	Victor Hugo	French	1831	Isabel F. Hapgood	1888	3rd
Around the World in Eighty Days	Jules Verne	French	1873	George Makepeace Towle	1873	3rd
Effi Briest	Theodor Fontane	German	1896	William A. Cooper	1914	3rd
The Merchant of Berlin	Luise Mühlbach	German	1896	Amory Coffin	1910	3rd
Venus in Furs	Leopold Von Sacher-Masoch	German	1870	Fernanda Savage	1921	1st
The Rider on the White Horse	Theodor Storm	German	1888	Margarete Muensterberg	1917	3rd
Debit and Credit	Gustave Freytag	German	1855	Georgiana Harcourt	1857	3rd

Table 5.1: Texts in main corpus

To keep the corpus balanced for each source language, a random contiguous section of two hundred kilobytes of text was selected from each work in the study and this was divided up into twenty chunks of ten kilobytes each. This results in one hundred textual segments per source language. This balancing of the corpus is important when using metrics such as type-token ratio which vary with relation to text length. Eighteen document-level features are employed in this analysis. Also experiments are carried out using ngram features, in this case word-unigrams and part-of-speech bigrams. Of course, the frequency of untranslated terms and titles from the source language together with placenames and character names could prove highly useful in predicting the source language of a text, however one would expect these to vary depending on the topics and themes within the text[4].

Experiments are carried out with and without proper nouns as textual features to compare the extent to which these influence the classification accuracy.

[4]One could imagine a novel translated from French in which the action takes place in a Francophone locale containing tokens such as Monsieur, Madame, Rue, etc.

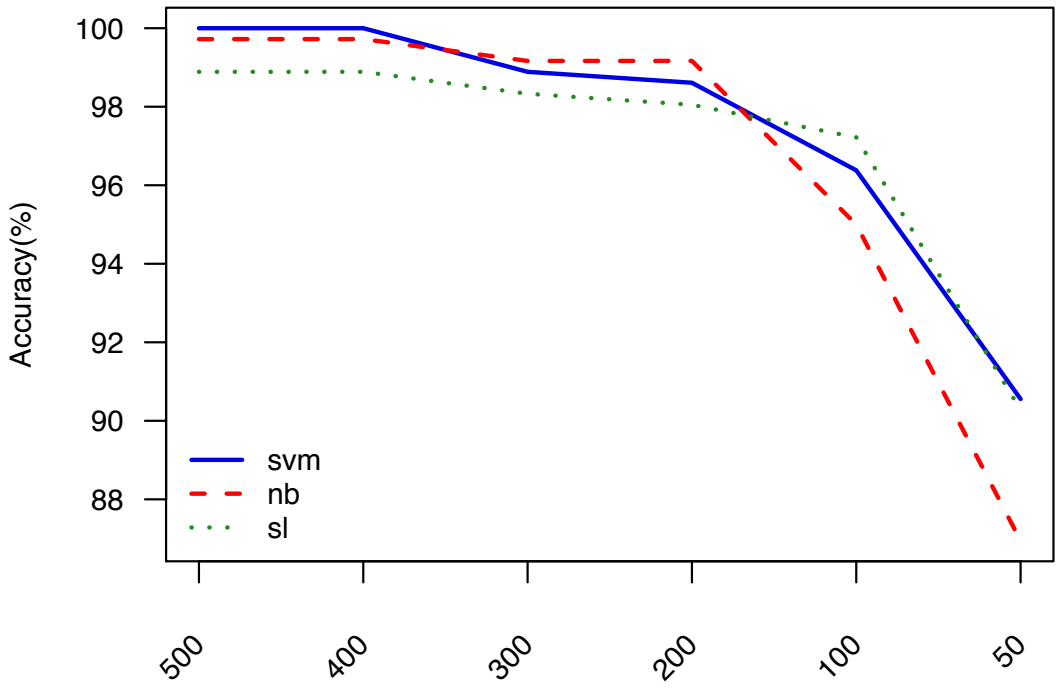

Figure 5.1: Word unigrams results : 4 languages

5.2 Translations plus originals

5.2.1 Single feature sets

Using the Support Vector Machine classifier, 66% accuracy is obtained using ten-fold cross validation over the four categories using the 18 document level statistics only. The Naive Bayes classifier performs worse, giving 54% accuracy. The Simple Logistic classifier performs the best here, with 68% accuracy using the eighteen document level features only. Given that the baseline for this task is 25%, 68% can be considered a quite promising result, considering that the features used here represent the frequencies of various parts-of-speech across an entire text segment and which should not contain any bias from themes or topics contained in the texts, although the results are lower for the hold-out set, at 62% for the Simple Logistic classifier.

Figure 5.1 displays the results for different numbers of word unigram features. SVM and Naive Bayes obtain almost 100% accuracy with the 500 feature set, with accuracy dropping off by ca. 10% on average when only 50 words are used. Figure 5.2 displays results using POS bigrams only as features, these results are considerably poorer than those using word unigrams, with the SVM classifier managing over 65% accuracy using 500 POS bigrams. It should be reiterated however that this result is still above the baseline.

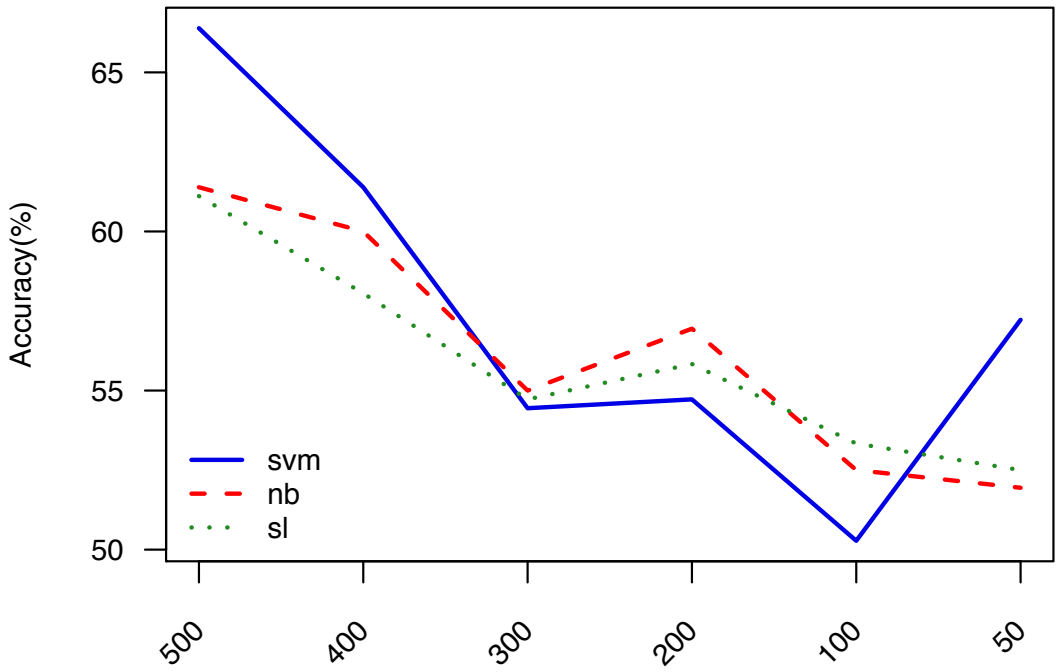

Figure 5.2: POS bigram results : 4 languages

Run	Training	Test	Classifier	Feature Set	Accuracy
1	Full	10-f cv	Baseline	n/a	25%
2	Full	10-f cv	NB	18 doc-level	54%
3	Full	10-f cv	SVM	18 doc-level	66%
4	Full	10-f cv	SimpLog	18 doc-level	**68%**

Table 5.2: Summary of classification accuracy: Full corpus

5.2.2 Combined feature sets

Figure 5.3 displays accuracy results for a mixed feature set containing ranked POS bigrams, document level features and word unigrams. The results here are comparable to those in Figure 5.1, however in this case the SVM just nudges out the Simple Logistic classifier to the top spot with ca. 99% accuracy using 400 features.

5.3 Translations only

5.3.1 Single feature sets

In order to examine the L1 prediction accuracy for the corpus of translations only, the English original texts are removed from the analysis and the same experimental setup is run again, this time with only the translations from the three source languages.

An increase in classification accuracy compared to the experiments using the four categories is obtained, the best result using document-level features in Table 5.8 is the SVM

Chi	Rank	Token	Chi	Rank	Token
191.1184	1	toward	60.2458	11	berlin
101.8571	2	de	56.4456	12	thousand
79.6687	3	von	54.1083	13	paris
78.6035	4	mr	52.0254	14	it's
78.1577	5	monsieur	50.1781	15	cossack
69.6095	6	francs	49.9458	16	rue
66.4622	7	m	49.868	17	hut
62.1622	8	**prepratio**	49.224	18	towards
62.1324	9	la	48.7354	19	**numratio**
61.1304	10	**nounratio**	48.6329	20	saint

Table 5.3: Features 1-20 for Figure 5.3

Chi	Rank	Token	Chi	Rank	Token
48.3455	21	**ari**	33.2283	36	anton
47.5911	22	**fverbratio**	33.1439	37	maryanka
47.1891	23	jude	32.2981	38	olenin
46.9136	24	lexrich	30.9333	39	foma
46.7665	25	dikemaster	27.0928	40	though
46.6164	26	bazarov	26.4912	41	hauke
43.3339	27	fink	26.2167	42	dorian
42.7951	28	dike	26.16	43	innstetten
37.8411	29	effi	25.7212	44	wanda
37.8411	30	passepartout	25.6271	45	fogg
37.8411	31	emma	25.6141	46	madame
37.8409	32	bovary	25.3143	47	mme
37.6963	33	mrs	25.2518	48	sue
36.2862	34	furs	24.1848	49	london
35.8047	35	farm	24.125	50	now

Table 5.4: Features 21-50 for Table 5.3

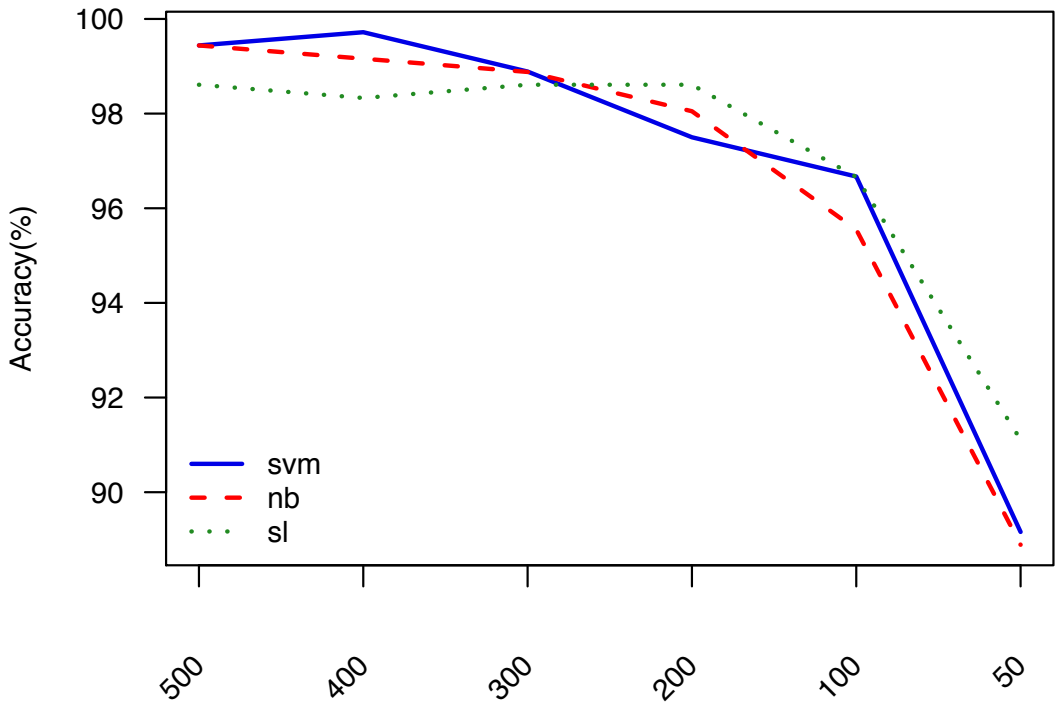

Figure 5.3: Mixed feature results : 4 languages

classifier with 78.66% using 10-fold cross validation although the result on the hold-out set is slightly lower. Interestingly, Naive Bayes performs better here. It must be taken into consideration however that in this case the baseline is of course higher, at 33%. Figure 5.4 is comparable to counterpart Figure 5.1 with Figure 5.5 a slight improvement with the Simple Logistic classifier managing an accuracy of over 75% using 500 features.

5.3.2 Combined feature sets

Figure 5.6 is comparable to its counterpart Figure 5.3 with high accuracy for all three classifiers, dipping slightly when the 50 feature set is used.

5.4 Removal of content words from mixed feature set

As can be seen in Tables 5.3 and 5.4, a good deal of the distinguishing features are proper names, place names and certain source language specific particles such as *de* and *von* which are used in French and German surnames names of noble descent.

There are still a number of document-level and POS features contained in this lineup, which bodes well for the robustness of the classifier. To investigate these features, all proper nouns were removed from a feature set of 200 features ranked using cross-validation on the 4 languages set, leaving 50 features in the set.

Table 5.7 displays the results obtained by using such a set, with the Simple Logistic classifier obtaining 85.5% accuracy using these features alone.

Chi	Rank	Token	Chi	Rank	Token
191.1184	1	toward	60.2458	11	though
101.8571	2	**prepratio**	56.4456	12	that's
79.6687	3	**nounratio**	54.1083	13	*RB-CC*
78.6035	4	thousand	52.0254	14	**conjratio**
78.1577	5	it's	50.1781	15	i'll
69.6095	6	towards	49.9458	16	*PRP-CC*
66.4622	7	**numratio**	49.868	17	i'm
62.1622	8	**ari**	49.224	18	*FW-FW*
62.1324	9	**fverbratio**	48.7354	19	*VBP-VB*
61.1304	10	**lexrich**	48.6329	20	law

Table 5.5: Features 1-20 for Table 5.7

Chi	Rank	Token	Chi	Rank	Token
48.3455	21	suddenly	33.2283	36	he's
47.5911	22	scream	33.1439	37	**avgsent**
47.1891	23	eh	32.2981	38	whispered
46.9136	24	resumed	30.9333	39	anyone
46.7665	25	*CD-CD*	27.0928	40	**typetoken**
46.6164	26	don't	26.4912	41	**complextotal**
43.3339	27	got	26.2167	42	**simplecomplex**
42.7951	28	stepped	26.16	43	**simpletotal**
37.8411	29	drink	25.7212	44	what's
37.8411	30	sense	25.6271	45	beneath
37.8411	31	passengers	25.6141	46	thought
37.8409	32	'eh	25.3143	47	there's
37.6963	33	**infoload**	25.2518	48	somewhere
36.2862	34	count	24.1848	49	ain't
35.8047	35	presently	24.125	50	you're

Table 5.6: Features 21-50 for Table 5.7

Run	**Training**	**Test**	**Classifier**	**Feature Set**	**Accuracy**
1	Full	10-f cv	NB	Mixed	80.83%
2	Full	10-f cv	SVM	Mixed	84.44%
3	Full	10-f cv	SimpLog	Mixed	85.55%

Table 5.7: Summary of classification accuracy: 4 language reduced feature set

Run	**Training**	**Test**	**Classifier**	**Feature Set**	**Accuracy**
1	Full	10-f cv	Baseline	18 doc-level	33%
2	Full	10-f cv	NB	18 doc-level	66.667%
3	Full	10-f cv	SVM	18 doc-level	**78.667%**
4	Full	10-f cv	SimpLog	18 doc-level	76%

Table 5.8: Summary of classification accuracy: Translations only

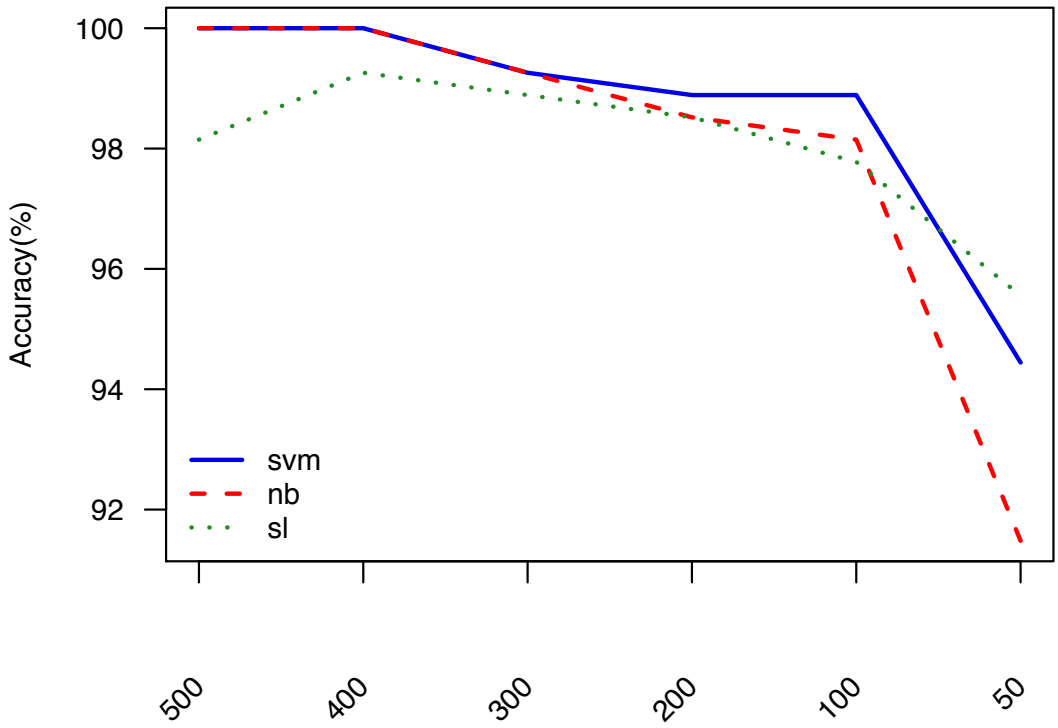

Figure 5.4: Word unigrams results : 3 languages

Metric	English	French	German	Russian
prepratio	0.04282897	0.0341565	**0.04480062**	0.04287795
nounratio	0.2023104	**0.2340025**	0.2100839	0.2076903
lexrich	0.3220333	**0.3487249**	0.321811	0.3161849
numratio	0.006568431	**0.009578647**	0.00527522	0.006415348
fverbratio	0.1071781	0.1004072	0.1075573	**0.116368**
ari	7.764139	**8.045916**	7.099629	5.889317
conjratio	0.1198757	**0.1223002**	0.1139004	0.110811

Table 5.9: Mean values for document-level features: 4 source languages

5.5 Discussion of features

5.5.1 Mean and SD values for document-level features

Examining Tables 5.9 and 5.10 which display mean and standard deviation values for a number of document-level features on the training set, it is evident that the French subsection of the corpus has the highest values for ratio of nouns to total words, lexical richness, ratio of numerals to total words, ratio of conjunctions to total words and Automated Readability Index. Although Kurokawa et al. (2009) found that Canadian Hansard text translated from French into English had a higher ratio of prepositions that original English Hansard text, in this case the subsection of the corpus with the highest preposition ratio is the section of the corpus with German as the source language. Russian had the highest proportion of finite verbs to total words among the subcorpora.

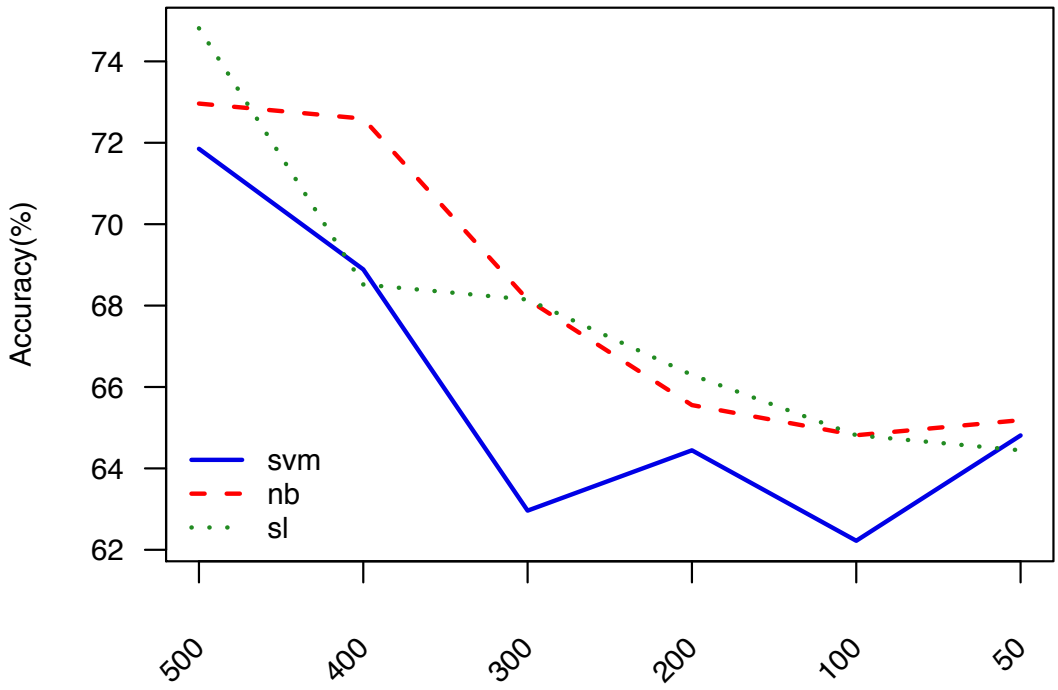

Figure 5.5: POS bigram results : 3 languages

Metric	English	French	German	Russian
prepratio	0.008145514	0.005600991	0.008586654	0.008219819
nounratio	0.02009025	0.02466903	0.0219087	0.02335919
lexrich	0.02786631	0.02720127	0.02478349	0.02632408
numratio	0.003038002	0.004918889	0.002365936	0.00304264
fverbratio	0.0149036	0.01085644	0.01399259	0.01291329
ari	3.668141	3.194295	2.443307	2.072083
conjratio	0.01258417	0.01093823	0.01071646	0.01311662

Table 5.10: Standard deviations for document-level features: 4 source languages

5.5.2 Single-word features

Viewing Table 5.12, the German translations have a much higher frequency of the word *toward* as opposed to the other texts. The most likely explanation for this is due to the nationality of the translators of the German texts, two were American[5], while the other texts were published in the US. The two contractions *it's* and *that's* are examined in Table 5.13. Olohan (2001) has shown that these forms tend to be less prevalent in translated English as a whole, however in this case they may be found to be less/more prevalent in translations from different languages.

Table 5.13 displays the frequencies of both *that's* and *it's* and the expanded versions of the same, *it is* and *that is*. As evidenced in the table, Russian has a much larger proportion of *that's* and *it's*, although the proportion of *it is* in the Russian corpus is also relatively high.

[5] Amory Coffin and William Cooper

L1	No. of tokens
German	185413
French	180813
English	148565
Russian	183448

Table 5.11: Number of tokens in each L1 sub-corpus

Text	**toward**	**towards**
English	0.000000	0.000441
Dorian Gray	0.000000	0.000188
Great Expectations	0.000000	0.000320
Jude The Obscure	0.000000	0.000466
Middlemarch	0.000000	0.000640
Treasure Island	0.000000	0.000596
French	0.000028	**0.000454**
Count Monte Cristo	0.000028	0.000865
Fr Goriot	0.000000	0.000160
Hunchback Notre Dame	0.000028	0.000385
Madame Bovary	0.000028	0.000469
Round World 80 Days	0.000057	0.000400
German	**0.000744**	0.000022
Debit and Credit	0.000513	0.000000
Effi Briest	0.000508	0.000000
Merchant of Berlin	0.000983	0.000000
Rider White Horse	0.001228	0.000000
Venus Furs	0.000485	0.000108
Russian	0.000185	0.000376
Fathers and Children	0.000000	0.000194
The Idiot	0.000000	0.000322
The Man Who Was Afraid	0.000938	0.000055
A Man of our Time	0.000000	0.000489
The Cossacks	0.000000	0.000810

Table 5.12: Frequency of toward/towards relative to total words

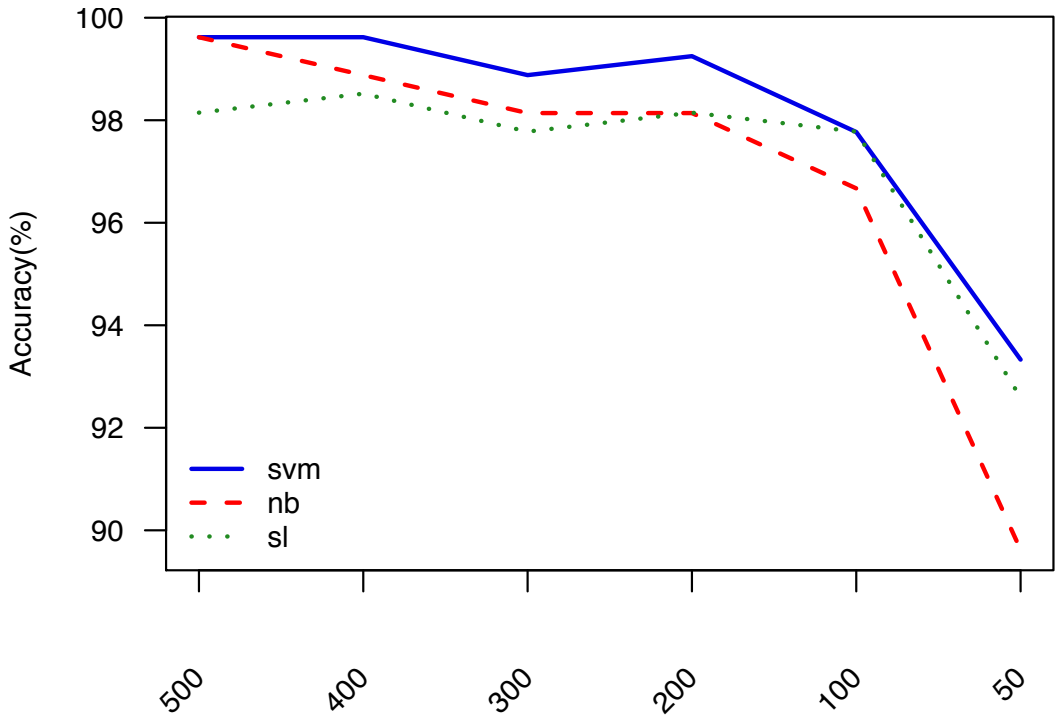

Figure 5.6: Mixed feature results : 3 languages

One possible explanation for this is that in French and German, *that is* and *it is* are two words [6], whereas in the Russian language, one word zto serves both purposes.

Table 5.14 displays the frequencies for the contractions *I'm* and *I'll* in the three corpora. Russian contains the highest frequency for the two contractions of the languages, in this case higher than in the original English corpus. This behaviour may also be due to an artifact from the source language: In German there is no equivalent contraction, *Ich bin* is I am, and in French the same phrase is *je suis*, both of these constructions contain two words.

In Russian *I am* corresponds to ya[7], with *I will* also being one word, budu[8].

The same behaviour can be seen in Table 5.15, with this being a possible explanation for the abundance of contracted forms in the translations with Russian as L1, however it is also the case that the expanded versions are highly frequent in the translations from Russian.

Table 5.16 displays the frequencies for the next four words in the list. It is less straightforward to ascertain whether these are true source language artifacts, although one might suggest that the frequency of *drink* in the translations from Russian may reflect a rather unsavoury national stereotype. It is interesting that the characters in the German translations tend to agree with an affirmative head movement more often than French or Russian. The high frequency of *thousand* in the French corpus is likely as a result of references to the French *franc* which at the time appears to have been referred to in large denominations.

The RB-CC bigram is featured in Table 5.5, which is the most discriminatory POS bigram

[6] Ger. *es ist* or *das ist* and Fre. *il est* or *qui est*.
[7] pronounced *ya* with a short a sound.
[8] pronounced *buuduu*

Text	it is	it's	that is	that's
English	0.002358	0.000361	0.000754	0.000538
Dorian Gray	0.004681	0.000000	0.002152	0.000000
Great Expectations	0.001225	0.000426	0.000746	0.000293
Jude The Obscure	0.002850	0.000000	0.000110	0.000685
Middlemarch	0.002171	0.000390	0.000724	0.000223
Treasure Island	0.000933	0.000959	0.000052	0.001451
German	0.002931	0.000194	0.001106	0.000116
Debit and Credit	0.003347	0.000027	0.000702	0.000108
Effi Briest	0.005668	0.000053	0.003850	0.000000
Merchant of Berlin	0.001572	0.000000	0.000618	0.000084
Rider White Horse	0.001411	0.000366	0.000575	0.000601
Venus in Furs	0.004152	0.000000	0.001079	0.000027
French	**0.003236**	0.000092	**0.001370**	0.000167
Count Monte Cristo	0.003013	0.000028	0.001228	0.000056
Fr Goriot	0.004440	0.000080	0.001872	0.000053
Hunchback Notre Dame	0.002035	0.000000	0.000880	0.000000
Madame Bovary	0.002761	0.000552	0.001215	0.000221
Round World 80 Days	0.002343	0.000314	0.000286	0.000257
Russian	0.003216	**0.001058**	0.001112	**0.001052**
Fathers and Children	0.001659	0.002074	0.000774	0.002074
The Idiot	0.005158	0.000887	0.001827	0.000484
The Man Who Was Afraid	0.003864	0.000883	0.001270	0.001684
Man of our Time	0.004347	0.000109	0.001358	0.000054
The Cossacks	0.001026	0.001350	0.000324	0.000999

Table 5.13: Frequency of that's/it's

in the feature set. This corresponds to an adverb-coordinating conjunction pair.

As an example, the text string "*ly and*" was chosen to investigate, basing assumptions on the fact that *and* is the most frequent English coordinating conjunction and a considerable percentage of adverbs in English end with *ly*. Querying the translation corpora for this string provides some interesting observations. The string occurs 110 times in the Russian corpus, 120 times in the German corpus, 49 times in the original English corpus but only 18 times in the French corpus. Indeed, after inspecting the results, it appears to be the case that the POS trigram RB-CC-RB is also more common in the translations from Russian and German than the translations from French.

5.6 Testing on unseen data

In order to further investigate whether the features observed in these studies are robust features for the classification of source language or to some extent biased towards this particular training set, a further test corpus of translated and original text from the same era has been compiled to validate the features. The same criteria as in Section 5.1 above apply, and the

Text	I am	I will	I'm	I'll
English	0.003112	0.000452	0.000318	0.000555
Dorian Gray	0.005327	0.000861	0.000000	0.000027
Great Expectations	0.003461	0.000213	0.000160	0.000506
Jude The Obscure	0.003946	0.000164	0.000164	0.000603
Middlemarch	0.002115	0.001002	0.000306	0.000111
Treasure Island	0.000778	0.000052	0.000933	0.001477
French	0.002500	**0.001416**	0.000061	0.000088
Count Monte Cristo	0.003571	0.001785	0.000000	0.000000
Fr Goriot	0.004226	0.002674	0.000053	0.000000
Hunchback of Notre Dame	0.001760	0.001045	0.000000	0.000000
Madame Bovary	0.001767	0.000497	0.000193	0.000276
Round World 80 Days	0.001086	0.001029	0.000057	0.000171
German	0.003463	0.001219	0.000092	0.000205
Debit and Credit	0.002646	0.002160	0.000000	0.000135
Effi Briest	0.004385	0.001016	0.000027	0.000214
Merchant of Berlin	0.001965	0.002022	0.000028	0.000000
Rider White Horse	0.000732	0.000209	0.000392	0.000418
Venus in Furs	0.007604	0.000755	0.000000	0.000243
Russian	**0.003598**	0.000883	**0.000627**	**0.000725**
Fathers and Children	0.003596	0.001106	0.001577	0.000332
The Idiot	0.004675	0.000537	0.000860	0.000457
The Man Who Was Afraid	0.004416	0.000386	0.000166	0.001242
Man of our Time	0.003043	0.001250	0.000136	0.000163
The Cossacks	0.002268	0.001134	0.000405	0.001431

Table 5.14: Frequency of I'll/I'm

same amount of text has been selected from each work. The test set is larger this time, comprised of 96 segments drawn across the works.

As can be seen from Table 5.19, the accuracy is lower here than on the test set which was drawn from the same corpus as the training set, with the SVM classifier managing only 43% accuracy using the 18 document level features. Table 5.20, displays the results from experimentation without the English original data included, the highest accuracy is again provided by the SVM classifier with 62% accuracy using the document level features. Although these results are significantly lower than the classification results on the original test set drawn from the same corpus, they still remain higher than the baseline in each case.

5.7 Conclusion

A hybrid approach towards detection of source language from literary text has resulted in high classification accuracies using ten-fold cross validation on the original translation corpus. Large sets of word n-gram features result in almost perfect classification accuracy(ca. 99%) using SVM and Simple Logistic classifiers, while a mixed set of fifty document-level,

Text	he is	he's	you are	you're
English	0.000355	0.000242	0.001927	0.000269
Dorian Gray	0.000484	0.000000	0.002529	0.000000
Great Expectations	0.000426	0.000186	0.001704	0.000266
Jude the Obscure	0.000384	0.000438	0.003233	0.000000
Middlemarch	0.000501	0.000111	0.001726	0.000056
Treasure Island	0.000000	0.000467	0.000518	0.000985
French	0.000752	0.000094	0.002091	0.000022
Count Monte Cristo	0.000558	0.000028	0.002678	0.000000
Fr Goriot	0.002247	0.000027	0.003209	0.000000
Hunchback Notre Dame	0.000385	0.000055	0.001595	0.000000
Madame Bovary	0.000166	0.000304	0.001878	0.000083
Around World 80 Days	0.000343	0.000057	0.001029	0.000029
German	**0.000766**	0.000011	0.002449	0.000038
Debit and Credit	0.000270	0.000000	0.001782	0.000000
Effi Briest	0.001604	0.000000	0.003048	0.000000
Merchant Berlin	0.000562	0.000028	0.001404	0.000000
Rider White Horse	0.000470	0.000000	0.001463	0.000183
Venus in Furs	0.000917	0.000027	0.004530	0.000000
Russian	0.000665	**0.000594**	**0.002497**	**0.000376**
Fathers and Children	0.000498	0.001189	0.002710	0.000968
The Idiot	0.000967	0.000296	0.003009	0.000081
The Man Who Was Afraid	0.000883	0.000442	0.003809	0.000304
Man of our Time	0.000652	0.000054	0.002119	0.000081
The Cossacks	0.000324	0.000999	0.000864	0.000459

Table 5.15: Frequency of he's/you're

POS bigram and word unigram features without content words can obtain 85.55% using the Simple Logistic classifier on the four language set.

These results show that although proper noun spotting can aid classification of the source language of a translation to a high extent, it is possible to create a robust feature set without these features that still obtains high classification accuracy.

A number of features have been attributed to effects other than source language influence, including whether the translator used US or British English in their translations.

A number of trends have been identified in the corpus of translations, such as the frequency of certain English contractions (*I'm*, *it's* etc) and the frequency of certain POS bigrams (adverb + coordinating conjunction in particular) which may be attributable to source language influence, however more research is needed to determine the origins of these effects.

Examining the frequencies of distinguishing features within the individual works, it appears that the frequencies can vary to quite some extent between the works that make up the individual L2 corpora. This must be taken into consideration when training a classifier, and indeed a larger training corpus may result in more robust features. In testing the classifiers

Text	drink	nodded	resumed	thousand
English	0.000194	0.000075	0.000048	0.000075
Dorian Gray	0.000027	0.000000	0.000000	0.000054
Great Expectations	0.000186	0.000213	0.000080	0.000213
Jude the Obscure	0.000329	0.000082	0.000000	0.000027
Middlemarch	0.000111	0.000056	0.000028	0.000000
Treasure Island	0.000311	0.000026	0.000130	0.000078
French	0.000083	0.000011	0.000227	**0.000785**
Count Monte Cristo	0.000028	0.000028	0.000056	0.000251
Fr Goriot	0.000027	0.000000	0.000080	0.001391
Hunchback Notre Dame	0.000055	0.000000	0.000495	0.000440
Madame Bovary	0.000166	0.000028	0.000000	0.000690
Round World 80 Days	0.000143	0.000000	0.000514	0.001143
German	0.000129	**0.000248**	0.000027	0.000167
Debit Credit	0.000243	0.000135	0.000027	0.000162
Effi Briest	0.000214	0.000294	0.000027	0.000053
Merchant Berlin	0.000056	0.000056	0.000084	0.000533
Rider White Horse	0.000052	0.000523	0.000000	0.000105
Venus in Furs	0.000081	0.000216	0.000000	0.000000
Russian	**0.000627**	0.000033	0.000016	0.000076
Fathers and Children	0.000166	0.000055	0.000000	0.000000
The Idiot	0.000296	0.000000	0.000000	0.000054
The Man Who Was Afraid	0.001132	0.000055	0.000055	0.000166
Man Time	0.000299	0.000054	0.000000	0.000054
The Cossacks	0.001242	0.000000	0.000027	0.000108

Table 5.16: Common word frequencies

on a test set drawn from unseen data from the same genre and time period, an average drop in classification accuracy of approx. 20% was observed, however the results were still almost double the baseline result on the three languages set and on the set containing English original text also. It may be of interest for future work to compile a larger corpus and examine whether a more robust feature set can be learned from a larger amount of data. Any future experiments could investigate a corpus containing a variety of textual genres, as well as a larger set of source languages. It may also be of interest to examine longer ngram sequences such as bigrams and trigrams of words and parts-of-speech, with the possibility of supporting non-contiguous sequences as used in the work by van Halteren (2008).

Text	anyone	presently	sense	suddenly	though
English	0.000022	**0.000113**	**0.000522**	0.000302	0.002385
Dorian Gray	0.000000	0.000000	0.000942	0.000511	0.001883
Great Expectations	0.000000	0.000160	0.000346	0.000160	0.002370
Jude The Obscure	0.000000	0.000247	0.000329	0.000356	0.003398
Middlemarch	0.000000	0.000111	0.000835	0.000083	0.002032
Treasure Island	0.000104	0.000052	0.000181	0.000389	0.002255
French	0.000061	0.000017	0.000188	0.000326	0.001869
Count of Monte Cristo	0.000000	0.000000	0.000195	0.000223	0.001674
Fr Goriot	0.000000	0.000000	0.000348	0.000134	0.001578
Hunchback of Notre Dame	0.000000	0.000000	0.000165	0.000385	0.001980
Madame Bovary	0.000276	0.000083	0.000083	0.000746	0.002292
Round The World 80 Days	0.000029	0.000000	0.000143	0.000143	0.001829
German	0.000043	0.000005	0.000232	0.000582	0.001375
Debit and Credit	0.000000	0.000027	0.000189	0.000243	0.001080
Effi Briest	0.000000	0.000000	0.000214	0.000374	0.002219
Merchant of Berlin	0.000000	0.000000	0.000084	0.000702	0.001039
Rider White Horse	0.000131	0.000000	0.000209	0.000679	0.001463
Venus in Furs	0.000081	0.000000	0.000458	0.000917	0.001052
Russian	**0.000213**	0.000016	0.000485	**0.001058**	**0.003505**
Fathers and Children	0.000000	0.000000	0.000498	0.000802	0.003485
The Idiot	0.000564	0.000000	0.000699	0.001585	0.004728
The Man Who Was Afraid	0.000055	0.000000	0.000580	0.000966	0.004195
A Man Of Our Time	0.000190	0.000027	0.000326	0.000706	0.002689
The Cossacks	0.000243	0.000054	0.000324	0.001215	0.002431

Table 5.17: More common word frequencies

.....No head was raised more **proudly and** more radiantly..... .....an offer which she **eagerly and** gratefully accepted..... .....**unceremoniously and** with no notice at all..... .....But after this I mean to live **simply and** to spend nothing..... .....I placed myself **blindly and** devotedly at your service..... .....**Outwardly and** in the eyes of the world .....They had parted **early and** she was returning home..... .....as the English law protects **equally and** sternly the religions of the Indian people..... .....vain attempts of dress to augment it, was **peculiarly and** purely Grecian.....

Figure 5.7: examples of RB-CC from the French corpus

Title	Author	Source	Date pub.	Translator	Translation pub.	Person
Jane Eyre	Charlotte Brontë	English	1847	n/a	n/a	1st
Vanity Fair	George Makepeace Thackeray	English	1847	n/a	n/a	3rd
A Study In Scarlet	Arthur Conan Doyle	English	1883	n/a	n/a	3rd
Dead Souls	Nikolai Gogol	Russian	1842	D. J Hogarth	1846	3rd
The Precipice	Ivan Goncharov	Russian	1869	Anon	n/a	3rd
Yama(The Pit)	Aleksandr Kuprin	Russian	1909	Guerney	1922	3rd
The Dream	Emile Zola	French	1888	Elizabeth Chase	1893	3rd
The Red And The Black	Stendahl	French	1831	C K Moncrieff	1925	3rd
Bel Ami	Guy de Maupassant	French	1885	Anon	1901	3rd
Michael Kohlhaas	Heinrich Von Kleist	German	1811	Frances H King	1914	3rd
Undine	Friedrich La Motte Fouqué	German	1811	Thomas Tracy	1897	3rd
Little Barefoot	Berthold Auerbach	German	1856	HWDulcken	1914	3rd

Table 5.18: Texts in reference corpus

Run	Training	Test	Classifier	Feature Set	Accuracy
1	Full	10-f cv	Baseline	n/a	25%
2	Full	Test	NB	18 doc-level	38%
3	Full	Test	SVM	18 doc-level	43%
4	Full	Test	SimpLog	18 doc-level	43.75%

Table 5.19: Summary of classification accuracy: 4 languages reference set

Run	Training	Test	Classifier	Feature Set	Accuracy
1	Full	10-f cv	Baseline	n/a	33%
2	Full	Test	NB	18 doc-level	56%
3	Full	Test	SVM	18 doc-level	**62%**
4	Full	Test	SimpLog	18 doc-level	54%

Table 5.20: Summary of classification accuracy: 3 languages reference set

Chapter 6

Stylistic markers of a literary translator

6.1 Introduction

This chapter contains the results of experiments towards a more more fine-grained stylometric analysis, the identification of distinguishing features between different styles of a literary translator. This study examines the writing of William Archer, which include his translations of plays by Henrik Ibsen either completed by him alone, or in collaboration with others, and his own original writings, consisting of a single dramatic work, *The Green Goddess* and several other non-fiction works. The initial point of focus is the Ibsen drama *Ghosts*, for which there exists a comparable contemporaneous translation by R. Farqhuarson Sharp. By comparing these two texts, a list of features which distinguish the two translations from one another are obtained, and then further examination of a corpora of translated and original text by Archer in comparison with more translations by Sharp is carried out in order to establish which features can be attributed to stylistic choices by the translator himself and which features may be due to influence from the source language or the topic or genre of a text.

6.2 Corpus

Texts were downloaded from *Project Gutenberg*,[1] with the exception of *The Green Goddess* which was downloaded from The Library of Congress online archive.[2] All of the front matter was removed from the Project Gutenberg versions of the texts, and some manual editing was performed on the online copy of *The Green Goddess* which contained some OCR errors and hyphenation of words at the end of lines which could confound any automated analysis of the text which treats these hyphenated compounds as separate words. For the initial experiment involving the parallel translations of Ibsen's *Ghosts*, 5K chunks of text were used to ensure a large number of segments for classification. For the experiments involving the larger corpus of six translations, 10K segments were used as in this case more text was available for examination.

6.3 Experiments on parallel translations of *Ghosts*

6.3.1 Word unigram results

In the first experiment, the two translations of *Ghosts* are compared. The text has been divided up into sections, 46 in total, 22 from Archer's translation and 24 from Sharp's translation (See Table 6.1 for total size per translator). First, the top 10 single-word features are selected using the chi-squared metric in the Weka toolkit. This provides a closer view of the words which prove to be discriminatory between the two translations. Using these 10 features for the classifier, 91% classification accuracy is obtained between the two translators

[1] `www.gutenberg.org`, last verified May 7, 2013

[2] `http://www.archive.org/details/greengoddessplay01arch`last verified May 7, 2013

Translator	Total Words
Archer	21412
Sharp	22482

Table 6.1: Number of words per translation

Token	Sharp	Archer
pastor	0.00004	0.0018
because	0.0008	0.0001
mr	0.0226	0.0033
i've	0	0.0009
i'm	0.00008	0.00116
back	0.0014	0.0011
standing	0.0007	0.0001
nearer	0.0002	0.00004
recollect	0.00004	0.00042
h'm	0	0.0005

Table 6.2: 10 most distinguishing words with frequencies relative to total words in each translation: *Ghosts*

using ten-fold cross validation.

Table 6.2 displays the 10 features. In his translation, Archer translates the name of one of the main characters as *Pastor Manders*, which is more or less identical to the original Norwegian name for the character, also refered to as Presten Manders[3]. Sharp however introduces him as *Parson Manders*, and also refers to him as *Mr Manders*, which explains the prevalence of Mr in Sharp's translation and Pastor in Archer's. Archer also prefers to use abbreviated forms in his translation, as seen by the discrepancy of the frequencies of *i've* and *i'm* in the two translations. It is not clear however, whether this is a stylistic choice on the part of the translator or whether an editor been involved in the standardisation process. However, is is the case that work by Olohan (2008) which examines a comparable corpus of translation and text from the British National Corpus has shown that optional and contracted forms tend to occur more often in translations than in non-translations.

The word *because* is an interesting distinguishing feature to examine further, as it may represent a more unconscious tendency towards the translation of a common closed-class word, rather than the preference of one lexical item over another[4] or an artifact of the editing process. A further analysis is presented in Section 6.6.1.

6.3.2 Word bigram results

Table 6.3 displays the most distinguishing bigrams for the two versions of *Ghosts*. Comparing these with the word unigrams in Table 6.2 above, abbreviated forms or the lack thereof

[3]similar to Father or Parson in English, used to denote a clergyman.

[4]Lexical choice in translation is indeed a topic of some interest in translation studies however this study will focus more on common word frequencies and document-level trends as features of a translator's style.

Token	Sharp	Archer
do not	0.0005	0.00009
i don't	0.0006	0.0017
very well	0.0001	0.0007
was the	0.0004	0.0001
it is	0.0034	0.0019
with all	0	0.0002
of the	0.0023	0.0028
don't want	0.0001	0
be very	0.0002	0
getting up	0.0002	0

Table 6.3: 10 most distinguishing bigrams with relative frequencies: *Ghosts*

Chi value	Rank	Bigram
15.372 +- 1.846	1.3 +- 0.46	SYM-UH
12.184 +- 1.55	2.1 +- 0.7	VBD-VBG
9.674 +- 1.877	3.3 +- 1.42	SYM-VBG
8.081 +- 1.005	4.6 +- 1.11	SYM-RB
8.272 +- 1.675	5.1 +- 1.76	VBG-NN
6.779 +- 0.659	7+- 1.41	VBN-NNS
6.821 +- 0.916	7.1 +- 1.45	NNP-NNP
4.587 +- 3.156	7.8 +- 1.66	VBP-VBG
3.538 +- 3.6	8.1 +- 2.43	NNP-RB
5.718 +- 0.747	8.6 +- 1.11	RP-CD

Table 6.4: 10 most distinguishing POS bigrams : *Ghosts*

prove discriminatory between the two translations. *I don't* vs *do not* and *it is* are part of the set of features, which results in 93% classification accuracy using ten-fold cross validation on the full training set, although the same caveat regarding editorial intervention should be mentioned here.

6.3.3 POS bigram results

Table 6.4 displays a ranked list of the most distinguishing part-of-speech bigrams in the two translations of *Ghosts*. The bigram *SYM-VBG* refers to a grammatical structure which consists of a symbol followed by the present participle. This construction manifests itself in the stage directions in the excerpt from *Ghosts* below.[5] Sharp's translation is first:

> Oswald (**going** into the hall). You shan't go out. And no one shall come in. (Turns the key in the lock.)
>
> Mrs. Alving (**coming** in again). Oswald! Oswald!–my child!
>
> Oswald (**following** her). Have you a mother's heart–and can bear to see me suffering this unspeakable terror?

[5]The opening parenthesis followed by the present participle.

```
simplecomplex <= 3.689655: sharp (9.0)
simplecomplex > 3.689655
|   avgsent <= 5
|   |   prepratio <= 0.024793: sharp (3.0/1.0)
|   |   prepratio > 0.024793: archer (12.0)
|   avgsent > 5
|   |   prepratio <= 0.033755
|   |   |   numratio <= 0.006369: sharp (13.0/1.0)
|   |   |   numratio > 0.006369: archer (2.0)
|   |   prepratio > 0.033755: archer (9.0/1.0)
```

Figure 6.1: J48 decision tree trace using document-level features for two translations of *Ghosts*

> Mrs. Alving (**controlling** herself, after a moment's silence). There is my hand on it.

followed by Archer's

> Oswald. [Also outside.] You shall not go out. And no one shall come in. [The locking of a door is heard.]
>
> Mrs. Alving. [**Comes** in again.] Oswald! Oswald–my child!
>
> Oswald. [**Follows** her.] Have you a mother's heart for me–and yet can see me suffer from this unutterable dread?
>
> Mrs Alving. [After a moment's silence, **commands** herself, and says:] Here is my hand upon it.

Sharp favours the use of the present participle in this excerpt from the parallel translations of stage instructions, although interestingly in these extracts, Archer prefer's *the locking of a door* vs. Sharp's *turns the key in the lock*.

6.3.4 Document-level results

Experiments using document-level metrics comparing the style of the two translations of *Ghosts* were also carried out. The Support Vector Machine classifier obtained 75% accuracy using ten-fold classification with the eighteen document-level metrics as features, the Simple Logistic Regression classifier performed slightly better with 77% accuracy using the same feature set. Figure 6.1 displays output of the J48 decision tree classifier which obtained a lower accuracy of 66% for the task of distinguishing between the translators, however the output obtained from this classifier is relatively easy to interpret by hand, with each level in the tree representing a decision point based on a particular value of one of the above metrics.

6.4 Comparing Archer's and Sharp's translations of different works

In order to investigate whether the distinguishing features between Archer's and Sharp's translations of *Ghosts* are indicative of a more general translation style or confined to the translation of that particular drama, three of Archer's other translations of Ibsen were compared with three of Sharp's. The plays chosen are listed in Table 6.5

Translator	Play
Archer	Little Eyolf
Archer	When We Dead Awaken
Archer	John Gabriel Borkmann
Sharp	An Enemy Of The People
Sharp	Rosmersholm
Sharp	Pillars Of Society

Table 6.5: Works in Ibsen translation corpus

The same experimental setup is used as in previous experiments, the SVM classifier and ten-fold cross validation. One issue with this particular experiment is that it is no longer a case of comparing parallel translations of the same plays, so any stylistic differences obtained may be due to other factors and not necessarily translator style. This is taken into consideration in the analysis and thus all features which contain any proper nouns which would naturally distinguish between different plays are removed. One could also argue for removal of all features containing common nouns, as content words can also vary in frequency based on the topic of a drama, however this is not carried out in this study.

Dividing the plays up into 10 kilobyte chunks, there are 83 files in total and the SVM classifier obtains 95% classification accuracy using the bigram features in Table 6.7 and 97.5% using the unigram features in Table 6.6. The relative frequencies displayed here are based on treating the translations of Sharp and Archer from Table 6.5 as separate corpora.

6.4.1 Unigram and bigram results

Examining the features in Table 6.7, the first six features represent pairs of functional words containing prepositions and common verbs. The bigram *comes in*, occurs generally in the translation of stage directions in the works in question by Sharp, as does *in from*.[6] A number of the highly-ranked bigrams contain nouns, when features six to ten in Table 6.7 are removed, classification accuracy drops to 84%.

Token	Sharp	Archer
community	0.0009	0
eyes	0.0002	0.0014
outburst	0	0.0003
public	0.0007	0
vehemently	0.00001	0.0004
smiling	0.00008	0.0006
hm	0.0006	0.0002
rises	0	0.0002
nodding	0.00004	0.0003
whispers	0.0001	0.00001

Table 6.6: 10 most distinguishing unigrams with relative frequencies: *Archer's translations vs Sharp's translations*

Token	Sharp	Archer
comes in	0.0006	0.00006
at him	0.0002	0.0015
beside the	0	0.0003
at her	0.0002	0.0015
in from	0.0005	0.00001
from me	0.00003	0.0005
the town	0.0009	0.00004
an outburst	0	0.0002
a man	0.0007	0.0001
his eyes	0.00001	0.0002

Table 6.7: 10 most distinguishing bigrams with relative frequencies: *Archer's translations vs Sharp's translations*

Chi value	Rank	Bigram
58.519 +- 2.407	1 +- 0	SYM-VBG
39.643 +- 4.073	2.3 +- 0.46	SYM-VBP
36.406 +- 1.964	3.1 +- 0.83	SYM-WP
34.237 +- 1.681	5 +- 0.77	SYM-IN
32.576 +- 2.106	5.7 +- 1.35	SYM-RB
31.992 +- 1.973	5.9 +- 1.45	VBD-VBG
32.563 +- 1.986	6 +- 1.48	SYM-CC
29.314 +- 4.205	7.2 +- 1.94	SYM-WRB
22.797 +- 1.757	9.2 +- 0.6	SYM-VB
22.624 +- 2.239	9.6 +- 0.66	SYM-PRP$

Table 6.8: 10 most distinguishing POS bigrams : *Archer's translations vs Sharp's translations*

6.4.2 POS bigram results

Table 6.8 displays the top-ten ranked part-of-speech bigrams from the corpus of Archer and Sharp translations. The top-ranked item is the SYM-VBG bigram which is examined using selections from the two translations of *Ghosts* in Section 6.3.3. This again amounts to a difference in the translation of stage instructions. 10-fold cross validation using this feature set results in accuracy of 95% for classification of translator.

6.4.3 Document-level results

The next set of experiments used the document-level features which were detailed in Table 3.2 above. The same plays are used as in Table 6.5 above, and the values for the eighteen features are calculated for each translator. Running a cross-validation experiment on the whole corpus, 97% classification accuracy is obtained for translator. This result is promising as these metrics should not be grossly affected by the occurrence of proper nouns or other features whose frequency may not be related to translatorial style. [7]

Viewing Table 6.9 which is obtained by ranking the eighteen document-level features by classification merit on the corpus of translations, average sentence length proves to be most discriminatory, followed by simple-complex ratio, complex-total ratio, type-token ratio and the ARI readability metric detailed in Chapter 3.

Tables 6.10 and 6.12 display average values for the document-level features which are most distinguishing between the two translators. With the exception of the ARI metric and the average word length value, all of the other features display similar relationships in both of the tables, indicating that the stylistic differences between the two translations of *Ghosts* are related to the stylistic differences between other Ibsen translations by the two translators. This is further explored with a number of cross-corpus experiments in Section 6.5.

6.5 Training on translator set and testing on parallel translations of Ghosts

The next experiment seeks to investigate the robustness of document-level features on unseen texts. Returning to the parallel translations of *Ghosts* once more, these are used as the test set for the next experiment. The training set is the document-level feature-set for the corpus of plays translated by Archer and Sharp, which does not include the translations of *Ghosts*.

This experiment seeks to identify whether it is possible to learn a particular translator's style from a number of different translations of texts by the same author and to apply the

[6] Conversely, the word *enters* occurs 13 times in the corpus of Archer's translations and does not occur at all in Sharp's.

[7] Of course, one cannot be completely certain that this is the case, subject matter and other factors may also prove discriminatory, however the robustness of the chosen features can be defended based on the fact that source language, original author and genre are held constant in this particular experiment

Chi Value	Rank	Feature
58.741	1	avgsent
54.1799	2	simplecomplex
54.1799	3	complextotal
54.1799	4	simpletotal
40.7983	5	ari
36.7533	6	avgwordlength
19.0563	7	typetoken
14.7002	8	lexrich
0	9	nounratio
0	10	fverbratio

Table 6.9: Average rank values of document-level features: *Archer's translations vs Sharp's translations*

Feature	Archer	Sharp
avgsent	4.852941	8.326531
simplecomplex	7.766311	4.126253
complextotal	8.766311	5.126253
simpletotal	1.138220	1.278736
ari	0.7945123	2.5428933
avgwordlength	3.18267911	3.4000256734
typetoken	0.2306864	0.2477705
lexrich	0.2511613	0.265975816

Table 6.10: Mean values of document-level features: *Archer's translations vs Sharp's translations*

Feature	Archer	Sharp
avgsent	0.7020469	2.401282
simplecomplex	2.228499	1.451242
complextotal	2.228499	1.451242
simpletotal	0.03637468	0.1122542
ari	0.6761878	1.236725
avgwordlength	0.12702167	0.131816
typetoken	0.01967373	0.01814374
lexrich	0.0199105	0.019283

Table 6.11: Standard deviations of document-level features: *Archer's translations vs Sharp's translations*

Feature	Archer	Sharp
avgsent	6.625	6.79166
simplecomplex	6.9167953	4.9812946
complextotal	7.91679533	5.98129462
simpletotal	1.161924746	1.23228306
ari	2.1516959623	1.3919062405
avgwordlength	3.3517745552	3.2858195566
typetoken	0.2845114552	0.2917867919
lexrich	0.322213	0.33101948971

Table 6.12: Mean values of document-level features: *Archer's Ghosts vs Sharp's Ghosts*

Feature	Archer	Sharp
avgsent	4.105272	1.2503622663
simplecomplex	2.577942912	2.1248945401
complextotal	2.577942912	2.12489454017
simpletotal	0.052524323	0.085150774742
ari	3.6485425260	0.85378511457059
avgwordlength	0.4468444854	0.15467892032
typetoken	0.02882899228	0.0194888386387
lexrich	0.031788511868	0.0179573998350

Table 6.13: Standard deviations of document-level features: *Archer's Ghosts vs Sharp's Ghosts*

learned classifier to classify which translator translated a parallel translation of the same text.

As the training sets for each translator contain different texts and the document-level metrics used do not take the content of words into account, topic-based side-effects should not be an issue in these experiments.

Running a cross-validation experiment using the SVM classifier in Weka, 79.167% accuracy is obtained for the classification of individual translator of *Ghosts*.

All eighteen of the document-level features are used in this experiment. Using the same experimental setup but with the J48 decision tree classifier instead of the SVM, an improved accuracy of 83.33% for the classification of the translator of each text is obtained.

Examining the decision tree trace output in Figure 6.2, one can see that the simple-complex sentence ratio is a discriminatory feature, along with average sentence length and preposition ratio, the first two features also ranked highly in Table 6.9. When the two translations of *Ghosts* are used as the training set and the corpus of Archer and Sharp is used as the test set, an even higher classification accuracy of 87% is obtained. The J48 decision tree gives an even better accuracy of 90% and the trace is provided in Figure 6.3.

The fact that classification accuracy is higher when trained on the parallel translations suggests that it may be easier for the machine to learn robustly distinguishing features when the training set is comprised of texts which are more similar to each other, in this case parallel translations of the same source text.

```
avgsent <= 5: archer (30.0/2.0)
avgsent > 5
|   avgsent <= 6
|   |   prepratio <= 0.031754: sharp (13.0/1.0)
|   |   prepratio > 0.031754: archer (5.0)
|   avgsent > 6: sharp (35.0)
```

Figure 6.2: J48 decision tree trace trained on Archer and Sharp corpus and tested on Ghosts

```
simplecomplex <= 3.689655: sharp (9.0)
simplecomplex > 3.689655
|   avgsent <= 5
|   |   prepratio <= 0.024793: sharp (3.0/1.0)
|   |   prepratio > 0.024793: archer (12.0)
|   avgsent > 5
|   |   prepratio <= 0.033755
|   |   |   numratio <= 0.006369: sharp (13.0/1.0)
|   |   |   numratio > 0.006369: archer (2.0)
|   |   prepratio > 0.033755: archer (9.0/1.0)
```

Figure 6.3: J48 decision tree trace trained on Ghosts and tested on Archer and Sharp corpus

Examining Figure 6.2 and 6.3, the ratio of prepositions to total words and the average sentence length are features which are shared by both decision trees.

6.6 Analysis of frequent discriminatory word forms in Ghosts

This section displays a closer analysis of a number of discriminatory words in the two translations of Ghosts. These words are obtained from Table 6.2 which lists a number of highly distinguishing unigrams from the two parallel translations.

6.6.1 Frequency of *because* in Archer and Sharp translations

This example from Sharp's translation displays the first usage of *because*:

> Engstrand. Yes, **because there** will be a lot of fine folk here tomorrow. Parson Manders is expected from town, too.

contrast with Archer's version:

> Engstrand. You see, **there's to** be heaps of grand folks here to-morrow. Pastor Manders is expected from town, too.

and the original:

> Engstrand. Ja, **for her** møder jo så mange fine folk imorgen. Presten Manders er jo også ventendes fra byen.

The next usage of *because* by Sharp is as a translation of a different phrase:

> Engstrand. But we must have some women in the house; that is as clear as daylight. **Because** in the evening we must make the place a little attractive

contrasting with Archer:

> Engstrand. But there must be a petticoat in the house; that's as clear as daylight. **For** I want to have it a bit lively like in the evenings, with singing and dancing, and so on.

and the original:

> Engstrand. Men fruentimmer må der være i huset, det er grejt som dagen, det. **For** om kvellerne skal vi jo ha' det lidt morosomt med sang og dans og sligt noget.

It is interesting how Archer uses the cognate in English whereas Sharp tries to use *because* in a sentence-initial position which does not sit as well from a stylistic point of view. Sharp and Archer's usage of *because* does coincide however as is demonstrated in the below passage:

> Mrs. Alving. I will tell you what I mean by that. I am frightened and timid, **because** I am obsessed by the presence of ghosts that I never can get rid of.

compared with Archer's translation

> Mrs. Alving. Let me tell you what I mean. I am timid and faint-hearted **because of** the ghosts that hang about me, and that I can never quite shake off.

and the original

> Fru Alving. Nu skal De høre, hvorledes jeg mener det. Jeg er ræd og sky, **fordi** der sidder i mig noget af dette gengangeragtige, som jeg aldrig rigtig kan bli' kvit.

Archer translates *fordi* in Norwegian as *because*, but in all other cases where Sharp uses *because* in the English translation, Archer prefers an alternative construction. Further investigation will determine whether this usage of *because* is reflected across other translations of Ibsen by Archer.

6.6.2 *Nearer* in both translations

Another distinguishing feature in the two translations of *Ghosts* is the word *nearer* which is used by Sharp more than Archer, as is shown in Table 6.16.

Archer's translation tends towards the use of the cognate *first* in this example:

> (Mrs. Alving enters by the door on the left; she is followed by Regina, who immediately goes out by the **first** door on the right.)

where Sharp uses nearer:

> (Mrs. Alving comes in by the door on the left. She is followed by Regina, who goes out again at once through the **nearer** door on the right.)

Another possible translation for the highlighted source is *foremost*.

> (Fru Alving kommer ind gennem dren p venstre side. Hun er fulgt af Regine, som straks gr ud gennem den **forreste** dr til hjre.)

Sharp prefers *nearer* in the next example, translating the Norwegian *nærmere*:

> Engstrand (going **nearer** to him). Yes, indeed one can; because here stand I, Jacob Engstrand.

With Archer prefering close:

> Engstrand. [Comes **close** to him.] Ay, but it can though. For here stands old Jacob Engstrand.

and the original:

> Engstrand (**nærmere**). Å jo såmæn gør det så. For her står Jakob Engstrand og jeg.

In the next example we see a similar pattern, with Sharp using *nearer* once more:

> Engstrand (coming a few steps **nearer**). Not a bit of it! Not before we have had a little chat.

and Archer preferring an alternative construction:

> Engstrand. [**Advances** a step or two.] Blest if I go before I've had a talk with you.
>
> Engstrand (et par skridt **nærmere**). Nej Gu' om jeg går, før jeg får snakket med dig.

In another example, Sharp prefers *nearer*:

> Mrs. Alving (coming cautiously **nearer**). Do you feel calmer now?

with Archer preferring *near*:

> Mrs. Alving. (Drawing **near** cautiously.)Do you feel calm now?

And the original:

> Fru Alving (**nærmer** sig varsomt), Føler du dig nu rolig?

In this next example however, Sharp eschews *nearer* for *in*:

> Engstrand is standing close to the garden door. His left leg is slightly deformed, and he wears a boot with a clump of wood under the sole. Regina, with an empty garden-syringe in her hand, is trying to prevent his coming **in.**)

whereas Archer again chooses *advancing*:

> (Engstrand, the carpenter, stands by the garden door. His left leg is somewhat bent; he has a clump of wood under the sole of his boot. Regina, with an empty garden syringe in her hand, hinders him from **advancing.**)

and the original:

> Snedker Engstrand står oppe ved havedøren. Hans venstre ben er noget krumt; under støvlesålen har han en træklods. Regine, med en tom blomstersprøjte i hånden, hindrer ham fra at komme **nærmere.**)

6.6.3 *Recollect* in both translations

Archer displays a tendency towards using the verb *recollect* when translating the Norwegian *husker*, whereas Sharp tends towards *remember* and other forms.

Archer uses *recollect* three times here:

> Oswald. Yes. I was quite small at the time. I **recollect** I came up to father's room one evening when he was in great spirits.
>
> Mrs Alving. Oh, you can't **recollect** anything of those times.
>
> Oswald. Yes, I **recollect** it distinctly. He took me on his knee, and gave me the pipe.

while Sharp prefers *remember*:

> Oswald. Yes; it was when I was quite a little chap. And I can **remember** going upstairs to father's room one evening when he was in very good spirits.
>
> Mrs. Alving. Oh, you can't **remember** anything about those days.
>
> Oswald. Yes, I **remember** plainly that he took me on his knee and let me smoke his pipe.

and the original *husker*:

> Osvald. Ja. Jeg var ganske liden dengang. Og så **husker** jeg, jeg kom op på kammeret til far en aften, han var så glad og lystig
>
> Fru Alving. Å, du **husker** ingenting fra de år.
>
> Osvald. Jo, jeg **husker** tydeligt, han tog og satte mig på knæet og lod mig røge af piben.

Archer's preference for *recollect* continues, also with *distinctly*:

Drama	TotalWords	ActualFreq	Relative Freq
John Gabriel Borkman	24239	7	0.0002
When We Dead Awaken	18070	11	0.0006
Ghosts	21412	3	0.0001
Little Eyolf	19078	11	0.0005
Hedda Gabler	29495	12	0.0004
The Master Builder	24810	11	0.0004

Table 6.14: Archer Translations: relative frequencies of *because*

> Manders. But then how to account for? I **recollect** distinctly Engstrand coming to give notice of the marriage. He was quite overwhelmed with contrition, and bitterly reproached himself for the misbehaviour he and his sweetheart had been guilty of.

with Sharp's translation using *remember*:

> Manders. I can't understand it, I **remember** clearly Engstrand's coming to arrange about the marriage. He was full of contrition, and accused himself bitterly for the light conduct he and his fiancee had been guilty of.

and the original:

> Pastor Manders. Men hvorledes skal jeg da forklare mig ? Jeg **husker** tydeligt, da Engstrand kom for at bestille vielsen. Han var så rent sønderknust, og anklaged sig så bitterligt for den letsindighed, han og hans forlovede havde gjort sig skyldig i.

6.6.4 Comparing Archer's and Sharp's translations of Ibsen

Table 6.14 compares relative frequencies of *because* in the translations of Ibsen by William Archer. The texts in italics are collaborative translation efforts, translations undertaken by Archer together with at least one co-translator. Observing the texts, *Ghosts* has the lowest relative frequency for the word *because* of the translations examined. Sharp on the other hand uses *because* more frequently than Archer in his translations, as evidenced by the figures in Table 6.15.

Table 6.16 displays the relative frequencies for a number of words in the works translated by Sharp and Archer. Frequencies of *recollect* and *nearer* differ highly in the translations of Ghosts by each translator, however the absolute frequencies of *and*, *or* and *but* do not differ to such a high extent in the parallel translations.

6.6.5 Frequency of *because* in Archer's original works

A number of original language works by Archer are examined here, his self-penned melodrama, *The Green Goddess*, and two prose works, one a manual on the art of writing drama, and the other a collection of letters and essays about his travels in the United States.

Drama	TotalWords	ActualFreq	Relative Freq
An Enemy of the People	31137	27	0.0008
Pillars of Society	27374	36	0.0013
Rosmersholm	31962	43	0.0013

Table 6.15: Sharp Translations: relative frequencies of *because*

Drama	and	but	nearer	or	recollect
Ghosts(Archer)	0.051980	0.008266	**0.000047**	0.033813	**0.000420**
John Gabriel Borkmann	0.029127	0.006188	0.000206	0.057016	0.000041
Little Eyolf	0.028148	0.007338	0.000262	0.035643	0.000052
When We Dead Awaken	0.034311	0.005036	0.000277	0.055008	0.000221
Ghosts(Sharp)	0.049595	0.007828	**0.000267**	0.029179	**0.000044**
Enemy of the People	0.026078	0.007708	0.000064	0.036323	0.000000
Pillars of Society	0.029566	0.008385	0.000188	0.032163	0.000000
Rosmersholm	0.022722	0.006831	0.000110	0.032111	0.000037

Table 6.16: Common word frequencies, Archer vs. Sharp translations

Table 6.17 shows relative and actual frequencies of *because* in original works authored by Archer. *The Green Goddess* contains a comparable proportion of the word with *Ghosts*, however the other works contain a more frequent usage, this may be due to the differing genres of the works in question. At this point, it may be of interest to consider a temporal effect in the difference in frequency for *because* and other terms in the translations.

6.6.6 Historical frequencies of *because*, *recollect* and *nearer*

Archer's translation of Ghosts was the first English version, and although information on an approximate publication date for the translation has proven difficult to obtain, the drama was first published in the original language in 1881 and the first performance in the English language occurred on 13th March 1891. Bibliographic information for the Sharp translation states that the first date of publication was in 1911[8].

In the interim period between the two translations, it may be interesting to note how the frequencies of certain constructions in English have changed. For a chronological overview of change in English literary text, the Google Books Corpus, (Michel, Shen, Aiden, Veres,

[8] `http://openlibrary.org/books/OL6371103M/Ghosts`

Work	TotalWords	ActualFreq	Relative Freq
The Green Goddess	24928	4	0.0001
Play-Making	100045	70	0.0006
America Today	51556	18	0.0003

Table 6.17: Archer Originals: relative frequencies of *because*

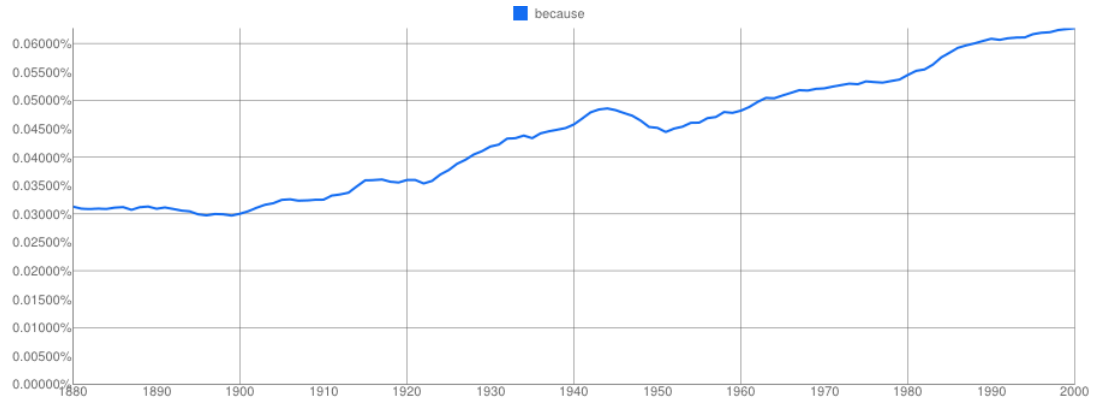

Figure 6.4: Relative frequency of because, British English subsection of Google Books Corpus: 1880-2000

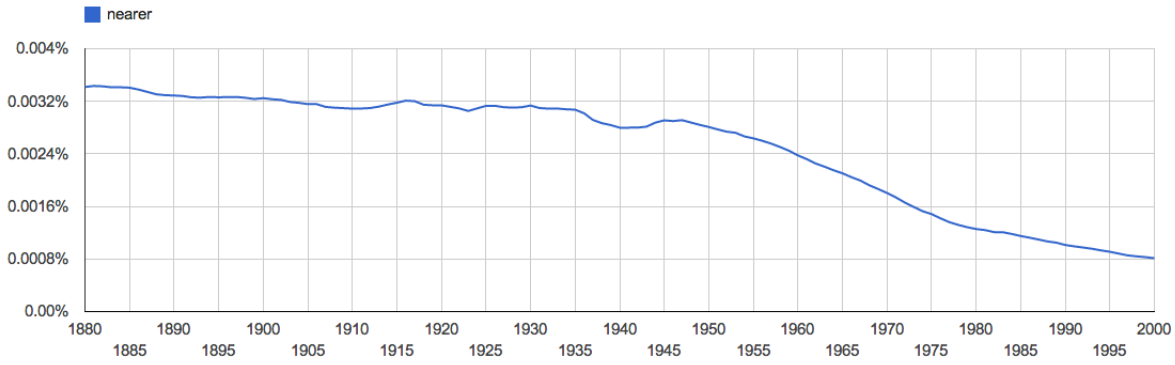

Figure 6.5: Relative frequency of nearer, British English subsection of Google Books Corpus: 1880-2000

Gray, Pickett, Hoiberg, Clancy, Norvig, Orwant, et al., 2011) and associated n-gram viewer[9] are used.

The Google Books corpus contains 5 millions books in a number of languages, which the authors claim represent 4% of all books ever printed, and is temporally tagged by date of publication, with texts ranging in time period from 1500 AD to the end of the 20th century.

A subsection of the corpus is taken as a reference, consisting of British English from 1880 to 2000 and the frequency of the discriminatory words are examined in this corpus.

Figure 6.5 displays the frequency of *nearer* in the Google Books Corpus, a slight decrease from 0.0328% in 1891 vs. 0.0309% in 1911.

Figure 6.6 shows a slightly steeper decline in the frequency of *recollect* in the Google Books Corpus, from 0.001% in 1891 to 0.0007% in 1911 however this is still proportionally less than the difference in relative frequencies for this term in the two translations of *Ghosts*.

From Figure 6.4, it can been observed how the frequency of *because* in British English has doubled from the beginning to the end of the 20th century, however the increase in frequency between 1891 and 1911 is relatively small, 0.035% in 1891 vs. 0.030% in 1911

[9] `http://books.google.com/ngrams`, last accessed May 7, 2013

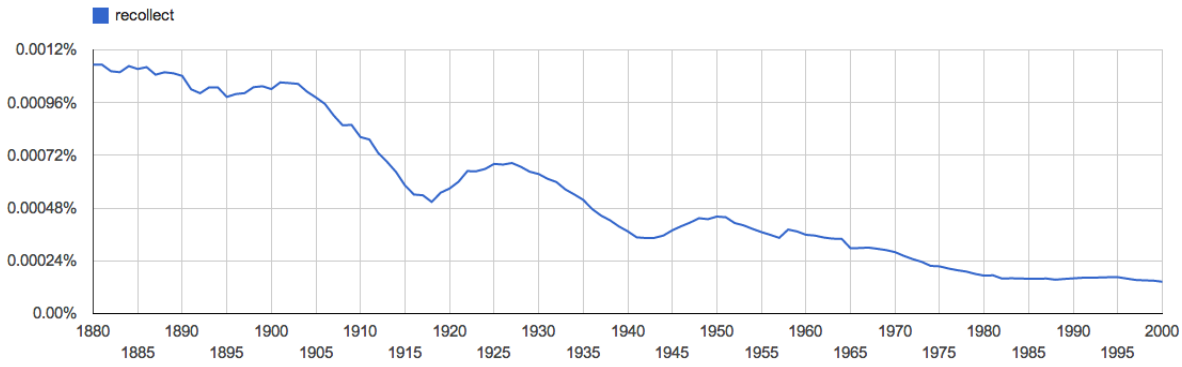

Figure 6.6: Relative frequency of recollect, British English subsection of Google Books Corpus: 1880-2000

Training	Test	Feature Set	Classifier	Accuracy
Ghosts vs. *Ghosts*	10fold CV	10 word unigrams	SVM	91%
Ghosts vs. *Ghosts*	10fold CV	10 word bigrams	SVM	93%
Ghosts vs. *Ghosts*	10fold CV	19 doclevel	SVM	75%
Ghosts vs. *Ghosts*	10fold CV	19 doclevel	SimpLog	77%
Ghosts vs. *Ghosts*	10fold CV	10 POS	SVM	95%
Archer vs. *Sharp*	10fold CV	10 word unigrams	SVM	97.5%
Archer vs. *Sharp*	10fold CV	10 word unigrams	SVM	95%
Archer vs. *Sharp*	10fold CV	19 doclevel	SVM	97%
Archer vs. *Sharp*	10fold CV	10 POS	SVM	97.5%
Archer vs. *Sharp*	*Ghosts* vs. *Ghosts*	17 doclevel	J48	83%
Ghosts vs. *Ghosts*	*Archer* vs. *Sharp*	17 doclevel	J48	90%

Table 6.18: Summary of classification accuracy over all experiments

in the larger corpus compared with 0.0008% in Sharp's translation vs. 0.0001% for Archer's in the two versions of *Ghosts*[10], which may suggest that the discrepancy in frequency of *because*, *nearer* and *recollect* in the translations of *Ghosts* by Archer and Sharp is related to translator style or some yet indeterminable factor rather than temporal variation in the target language.

6.7 Conclusion

In the experiments in this chapter, a number of stylistic traits have been established which distinguish the translations of Henrik Ibsen by William Archer from those translated by R. Farquharson Sharp using machine-learning classifiers and features from the field of text classification and computational stylometry, as employed in previous studies (Baroni & Bernardini, 2006; Ilisei et al., 2010) on comparable monolingual corpora of translations.

Cross-validation experiments have resulted in high classification accuracy between the two translators using both document-level features and ngrams, which suggests either set of

[10] 8 times as many in Sharp's version of *Ghosts* vs. 1.16 times as many in the Google Books corpus.

features or indeed ultimately a combination of both feature types are most useful for the task of distinguishing between the two translators examined here.

Archer appears to tend towards usage of contracted forms more than Sharp, with the lists of distinguishing features for both bigrams and unigrams containing these forms, *do not* vs. *don't* and *it is* vs. *it's* are two particular examples, although as previously stated this could indeed be an artifact that is introduced at the editing stage. Regarding the frequency of certain part-of-speech bigrams, Sharp appears to favour the present participle over the past tense in the translation of stage directions, *coming in* vs *comes in* in Archer's translation.

The ratio of simple to complex sentences and the average sentence length are two document-level features which distinguish Archer's translations from Sharp's, this has been further verified by training on a non-parallel set of different Ibsen plays translated by both playwrights and testing on the parallel translations of *Ghosts*, resulting in ca. 80% accuracy for detecting the translator of a particular text chunk, based on stylistic fingerprints obtained from the larger corpus.

Section 6.6.6 describes a basic chronological word frequency analysis by comparing the increase of frequency of the words *because*, *nearer* and *recollect* in a general temporally-tagged corpus of British English with the relative frequency of these words in the two translations of *Ghosts* by Sharp and Archer and concludes that the discrepancy in frequency is more likely to be as a result of translator style or some yet-unknown factor other than temporal variation in the source language.

The decision tree classifier trace in Figure 6.2 indicates that Archer may have a tendency towards shorter sentences, with the ratio of simple to complex sentences also playing a role in the distinction, this again is likely affected by Sharp's preference for using present participles in stage direction translation as the ratio is calculated based on the number of finite verbs in a sentence. This phenomenon is also captured in the average frequencies of the distinguishing document-level features which are presented in Table 6.10. Stage directions prove to be of further interest regarding the relative frequency of word bigrams *comes in* and *in from* in Sharp's translation when compared with the frequency of *enters* in Archer's.

It is interesting to note that many of the document-level features did not prove discriminatory in the experiments, one possible reason for this could be the genre of the texts examined here, a number of these ratios were obtained from work by Ilisei et al. (2010) who examined technical and medical translation in Spanish, one can imagine that certain discourse markers, for instance, may not occur as frequently in dramatic text as in flat prose or technical writing. However, the identification of certain document-level and part-of-speech trends as discriminatory is highly promising as these features could be deployed in studies which seek to attribute the provenance of a translation of unknown origin to a known translator, although it is yet unclear whether they can robustly identify a translator's style across different genres and source languages. The features established as distinguishing in these experiments may vary as a function of original authorship and translator individual choice; however the document-level features may indeed be more robustly discriminating between translators, and perhaps

may generalise to different translators, authors, genres or source and target languages, but further research on various comparable corpora of translations is required using the methods employed in the current study, in order to investigate these claims more thoroughly.

Chapter 7

Conclusions and future directions

7.1 Introduction

This chapter concludes the thesis and summarises the results from Chapters 4, 5 and 6 and describes a plan for future work on detecting markers of translation style in English textual corpora. Section 7.2 sums up the individual experiments from each chapter with Section 7.3 comparing common traits in the three experiments. Section 7.5 details a number of future experimental directions and Section 7.6 provides some concluding remarks.

7.2 Overview of results

This thesis presents the results of experiments on a number of different comparable corpora, these experiments seek to answer questions of a coarse-grained nature, such as detecting whether a text is an original text or a translation to progressively more fine grained analyses such as the detection of the source language of a literary translation right down to the detection of the translator of a particular parallel translation.

In general the classifiers and feature sets which were used in the experiments in this thesis performed comparably to the current state-of-the-art in each of the research subquestions, with accuracies of ca. 80% and higher reported across the experiments.

7.2.1 Chapter 4

Chapter 4 examined two different comparable corpora using the feature set described in Chapter 3. Classification results on the Europarl corpus were highest for a mixed feature set of 500 features, containing word unigrams, word bigrams POS bigrams and document level statistics. This gave an accuracy of 88% using the Simple Logistic Regression classifier. The best result on the NYT corpus was 69% accuracy using six of the document-level metrics.

Examining the combined feature sets, both corpora had document-level ratios in the top 10 features, readability metrics such as the Automated Readability Index and Coleman-Liau index, average word length, ratio of nouns to total words and the ratio of closed-class to open-class words were all features which distinguished the translated sections of both corpora from the original sections. The word bigram *believe that* was a feature of interest in both corpora, occurring almost twice as often in the translated side of each corpus.

7.2.2 Chapter 5

Chapter 5 focused on detecting the source language of a literary translation. Classification results were high in general for cross-validation and test set experiments, but perhaps not as high as those for the experiments in Chapters 4 and 6, indicating that automated detection of the source language of a literary text may be a more challenging task than classifying the translator of a literary text or indeed separating translated parliamentary proceedings from those whose original language was English.

Results on this task varied from almost full (ca. 99%) classification accuracy using 500 word feature sets and SVM and Simple Logistic classifiers, to 85.5% using a mixed feature set containing words, part-of-speech bigrams and document level statistics from which all content words had been removed.

When experiments were carried out on a new unseen set of texts from the same era using document-level features, the classification accuracy dropped, to 43% and 62% using the SVM classifier on a four language set and a three language set respectively.

Contractions were found to be efficient in distinguishing between source languages in the experiments, this can be a reflection of whether an English contraction represents two words or one word in the individual source languages, this theory is discussed to a greater extent in Chapter 5, Section 5.5.

Perhaps the higher number of categories involved[1] played a role in this, although it must be reiterated that all classification results were a significant improvement on the baseline in all cases.

Two items of information must be taken into consideration when examining these results. The first is that the classifiers were trained on only twenty different literary works, each from different authors and translators and spanning a range of topics from the four source language sections, although there were four hundred files in the experiment. Indeed, a larger and more diverse corpus might result in a more robust classifier, however 85% accuracy for source language detection over a baseline of 25-33% is a reasonable result, as are the results on the unseen texts.

The second item to note is that although the Europarl source language detection experiments described in van Halteren (2008) obtained 96.7% accuracy, this is a compound result as five translations of the same text were available for analysis. Results using one target language only obtained between 81% and 87% over the different target languages examined, and the corpus size in this case was much larger[2] and more consistent in style, consisting entirely of European parliamentary proceedings.

In this context the results of the experiments in Chapter 5 can be seen as promising, although it will be of interest to examine a larger corpus in future experiments.

7.2.3 Chapter 6

Chapter 6 describes experiments towards detecting the translator of a parallel translation of a work by Henrik Ibsen. Feature sets containing words, word bigram, part of speech bigrams and document-level features all performed well in cross validation experiments on the parallel translations of the play *Ghosts* by Ibsen, giving accuracies over 90% for single-feature sets. Further investigation into the distinguishing features is conducted, identifying trends in differing usage of the words *because*, *nearer* and *recollect* between the translators, cou-

[1] Four source languages to choose from in this case, compared with the binary question of translated vs. original or translator A vs. translator B.

[2] 1000 texts vs 400.

pled with the usage of differing verb forms in the translation of stage instructions. A cross validation experiment is designed where a classifier is trained on the larger set of six dramas and tested on the parallel translations of *Ghosts*, which resulted in 83% accuracy for classification of translator using document-level features alone using the J48 decision tree classifier, identifying stylistic traits between the two translators over the corpus which consists of three translations of different Ibsen plays, which are also distinguishing factors in the parallel translations. An even more interesting result is the fact that a classifier trained on the two parallel translations using document-level features and tested on the larger corpus performed even better, with 90% accuracy using the decision tree classifier. Perhaps training on the parallel translations resulted in a more robust model of translator style, as both translations were from the same source, as opposed to the larger more diverse corpus where different texts made up the two sections of the training set and confounding factors could emerge due to differences in the content of the texts. Archer's translations tended to have a lower average sentence length and lower average word length and also a lower ARI score and lexical richness measure. It may be interesting in future work to examine the values for these metrics coupled with a human judgment of translation quality, although of course this could be highly subjective.

It can not be confirmed based on the results of a single study of two translators only that these features will be discriminatory for the work of other translators, although the literature provides examples of average sentence length discriminating between translators as well as being a discriminating feature between translated and original text, as touched upon in Section 3.4.1.

It is of interest to examine other parallel translations from other authors using the same methodology employed here in order to compare discriminating features across a range of translators, authors and textual genres. For example, it may be the case that it is easier to distinguish parallel translations of a drama from one another than it is to distinguish parallel translations of a novel, based on the fact that a drama may have a more rigid stylistic structure.

7.3 Trends across experiments

7.3.1 Features

Table 7.1 contains the top-ten mixed features from the two corpora in Chapter 4, the literary text corpus from Chapter 5 and the top 8 features from the experiments in Chapter 6 which used separate translations of Ibsen by the two translators examined as the training set and the two parallel translations of the same Ibsen play as the test set, although in this particular experimental case only document-level features are used. Those features in italics occur in at least three of the corpora examined.

Europarl		NYT		SL		TS	
1	avgsent	1	*avgwordlength*	1	toward	1	avgsent
2	*nounratio*	2	grammlex	2	prepratio	2	simplecomplex
3	*avgwordlength*	3	people who	3	nounratio	3	complextotal
4	grammlex	4	community	4	thousand	4	simpletotal
5	though	5	JJS NNS	5	it's	5	*ARI*
6	infoload	6	state of	6	towards	6	*avgwordlength*
7	must	7	decade	7	numratio	7	typetoken
8	IN-WDT	8	RP IN	8	*ARI*	8	lexrich
9	*ARI*	9	an american	9	fverbratio	9	*nounratio*
10	conjratio	10	the u	10	lexrich	10	

Table 7.1: Overview of features: Translationese vs. source language experiments vs. translator style

7.3.2 *though* in literary and parliamentary *translationese*

The word *though* occurred 174 times in the translated side of the Europarl corpus and 34 times in the original side, this is a relative comparison of 0.000196 to 0.000040. In the source language corpus, this word occurred more frequently in the section of the corpus which had been translated from Russian, although not to such a drastic extent[3] when compared with the original English.

7.3.3 *believe that*: Frequency of *complementizer that* constructions in translated text

As detailed in Chapter 4, Section 4.7, the frequency of the word bigram *believe that* was considerably higher in the translated sections of both the NYT comparable corpus and the Europarl comparable corpus. This is an example of an optional construction in English, as described in work by Olohan (2001). This bigram occurred over twice as frequently in the translated section of each of the comparable corpora,

7.3.4 The efficacy of contractions for source language detection and markers of a translator's style

Contractions have been identified in Chapters 5 and 6 as distinguishing features in the experiments on translator style and on source language detection. Investigating further the frequency of these features in Chapter 5, it is found that although general trends exist in the source language subcorpora when the frequencies of these items are measured in the subcorpus as a monolithic whole, there is a large degree of variation between the usage of contractions in the individual works in each sub-corpus. Trends suggest that the translations from Russian, a language which incorporates single-word items to represent many of the ex-

[3] See Table 5.17 in Chapter 5 for details.

panded forms of the contraction in English, contain a higher frequency of these contractions, which could be interpreted as source-language influence on the translations.

However, the work in Chapter 6 illustrates a number of contractions[4] which are found to be distinguishing between the two translators in question, a finding which concurs with the results in Chapter 5 which show variation in frequency for these features amongst the different literary works in the corpus. Indeed, contractions have been found to be distinguishing between translation and original texts also by Olohan (2008) who found that contracted forms were more frequent in original text from the British National Corpus than a corpus of comparable translations, lending support to the *explicitation* universal of translation. Further study on corpora of translated text from different source languages and a variety of parallel translations is required to determine how contractions and indeed other optional items in English vary across text types, source languages and translators.

7.4 Experimental results in the context of translation universal theory

It is of interest to examine the experimental results in the thesis with respect to the notion of translation *universals* proposed by Baker (1996) and also examined by Pastor et al. (2008) in their work which combines statistical analysis and translation studies methodology.

In Chapter 4, metrics calculated on comparable corpora displayed similar behaviour for some features such as higher average word length for original text and higher ratio of nouns, but did not display similar behaviour with regard to others such as certain readability scores, with Europarl originals having a lower ARI score than comparable translations and the opposite being observed on the NYT corpus. Both translated sections contained higher proportions of closed-class words than their original counterparts, which alludes to the universal of *simplification*. As mentioned in Section 7.3.3, certain bigrams did display similar behaviour across both corpora, indicating that two comparable corpora in different genres can still share universal behaviours of this nature.

However, experiments carried out in Chapter 4 which were themselves inspired by work by Koppel and Ordan (2011) show that training a classifier using document-level statistics as features on one corpus and using it to try and classify translated and original sections of another parallel corpus in the same language produces poor results, which provides some evidence against the existence of any universals of translation.

Indeed, the existence of these *dialects* of *translationese* is given some weight in Chapter 5, where classifiers are trained to detect the source language of contemporaneous literary translations. The fact that this is possible to a statistically significant degree in itself indicates that translations are often quite distinguishable from one another and not as homogeneous as Baker and proponents of her theories may suggest, although the question still remains as to whether they are more internally homogenous than a corpus of original text in the same

[4] don't and it's are two examples which are expanded upon in this chapter.

genre. Experiments using four language categories, one of which was original English, and those examining three categories of translated text only reported comparable classification accuracy results to the baseline in each case.

Section 7.3.4 discusses the topic of contractions in translated text in English. These are a form of optional item in English and as discussed have been found to be more frequent in translated text in past studies, lending weight to the universal of *explicitation*. In Chapter 5 and Chapter 6, these items were found to be discriminatory features of source language and translator's style respectively across several different textual genres and works. Whether this can be accepted as a *universal* of translation will require further study, however it is an interesting result in the context of translation studies in general.

Finally, with regard to the study by Pastor et al. (2008) which found no evidence for *convergence*, the experimental results detailed in this thesis concur with the lack of evidence for this particular universal. Although no detailed comparison was carried out on a comparable set of original texts, save for the inclusion of English originals in the experiments to detect source language, the very fact that stylistic differences between a translation of the same source text and works by the same author could be learnt to such an accurate degree suggests that asserting that translations as a class of text are somehow more integrally homogeneous in general than original text is a weak argument, although it will be of interest to examine more parallel translations by different authors and translators to ascertain whether this is a trait which can be observed in several cases.

7.5 Areas for future exploration

This section describes areas which might be explored in future work on the topic of translation markers in English text.

7.5.1 Classifiers

The experiments carried out here used single classifiers in general. Future experiments could benefit from using ensembles of classifiers as is done in the work by Baroni and Bernardini (2006), using voting schemes such as *majority voting* or *recall maximisation*. The Support Vector Machine classifier generally performed well across the different experiments, although there were some cases where the Naive Bayes classifier actually outperformed the SVM classifier, such as in the experiments on the NYT corpus in Section 4.5.2. The J48 decision tree performed well in the experiments on translator style in Chapter 6. Future experiments could benefit from combinations of these classifiers.

7.5.2 Experimental design

The experimental design was motivated in the case of each sub-question by how an individual corpus would facilitate the answering of that particular sub-question. However, with some

perspective on the sub-questions in general, a future experiment might attempt to investigate the three questions dealt with in this thesis using the same or subsections of the same corpus of texts for each question, in order to enable a more direct comparison of features and results obtained. A corpus of literary texts would be a likely candidate for investigation in this case as the identity of translator and author should be clearly indicated, indeed a larger version of the corpus examined in Section 5 would be an ideal starting point for such an experiment. In an ideal case, this comparable corpus would have the following properties:

- The corpus should be contemporary in the sense that all translations and originals should be drawn from a limited time period in order to avoid temporal issues.
- The corpus should contain translations from several source languages.
- The corpus should contain translations from several translators and authors.
- The corpus should contain a number of parallel contemporaneous translations of texts.

There are several other criteria which would enable further issues of translation style between authors and translators to be examined, namely:

- The corpus should contain translations and original texts by the same author/translator and ideally in the same genre.
- The corpus should contain translations by the same translator from different authors and/or source languages.

These two criteria can be difficult to fulfill in reality, as not all translators are also published authors in their native languages, and from experience during the compilation of the source language corpus for the work in Chapter 5, it appeared common practice for one translator to translate the entire *oeuvre* of a particular author, or a number of authors from the same source language, although of course this is not exclusively the case. However, with access to a database of modern translations in a digital format, as many researchers in translation studies may have, some of these issues should become less of a concern than in the case of trying to assemble such a corpus solely based on works in the public domain.

7.5.3 Industrial applications

As evidenced by the literature in the machine translation community, automatically detecting translated text from original text has become an important research question, as institutions look to the web to obtain parallel corpora and language models for training large-scale statistical machine translation systems. It is of the utmost importance that machine-translated text does not find its way back into the training corpora for these systems, as this would be detrimental to the training of any future models, thus the need for systems to detect different textual qualities. Another application of these classifiers could be in estimations of translation quality, however this would require human annotated judgments of translation quality

attached to a corpus of translations, which could prove to be rather subjective. The methodology developed in this thesis could be applied in other domains where questions of textual style are important, such as controlled language verification software for large multinational corporations.

7.5.4 Metrics used

The metrics employed in this thesis were taken from the literature on translation stylometry, text analytics and corpus linguistics. However, it is the case that the document-level features were based to a greater extent on the work by Ilisei et al. (2010) which initially focused on translationese in Spanish and then moved on to Romanian text. The work on Romanian contained extra features pertaining to the Romanian language itself, and this may also be a direction which can be explored in future work, using features which pertain more to linguistic phenomena in the English language. Future experiments could employ parse trees as features, as in the work on L1 detection by Wong and Dras (2009), or information theoretic measures such as *perplexity* as used by Koppel and Ordan (2011). Perhaps most interesting would be features which capture elements of English which are not yet captured by the document-level metrics used in this thesis, such as the frequency of passive voice vs. active voice in a text. With the speed of development in natural language processing systems in recent years, it may not seem so outlandish to suggest features which attempt to capture more complex phenomena such as the level of *metaphoricity* of a text in future experimentation on the stylistics of translations.

7.6 Final remarks

The experiments in this thesis have applied computational linguistics methods to answer questions which relate to corpus-based translation studies. The field of traditional translation studies has been rapidly adjusting to changes in their landscape for the past twenty years and it may be the case that there is some resentment from the traditionalists who favour qualitative approaches to translation studies rather than corpus-based studies. The intention with the work carried out here is not to displace the qualitative studies, but in fact to augment these studies with tools which can identify patterns in text which are more difficult to spot in qualitative work, which in turn may lead to the development of new theories of a qualitative nature. It is hoped that more collaboration between researchers in computational linguistics and translation studies with a similar focus to the seminal work of Baroni and Bernardini (2006) will be fostered in the coming years and this researcher in particular would welcome a spirit of collaboration between the disciplines.

Bibliography

Aizawa, A. (2003). An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, *39*(1), 45–65.

Altintas, K., Can, F., & Patton, J. (2007). Language change quantification using time-separated parallel translations. *Literary and Linguistic Computing*, *22*(4), 375–393.

Argamon, S., Koppel, M., Pennebaker, J., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, *52*(2), 119–123.

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. *Benjamins Translation Library*, *18*, 175–186.

Baker, M. (2000). Towards a methodology for investigating the style of a literary translator. *Target*, *12*(2), 241–266.

Baker, M., et al. (1993). Corpus linguistics and translation studies: implications and applications. *Text and technology: in honour of John Sinclair*, *233*, 250.

Baroni, M., & Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, *21*(3), 259.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge Univ Pr.

Biber, D. (2003). Variation among university spoken and written registers: A new multi-dimensional analysis. *Language and Computers*, *46*(1), 47–70.

Biber, D. (1988). Variation across speech and writing. *Cambridge: CUP*.

Borin, L., & Pruetz, K. (2001). Through a glass darkly: Part-of-speech distribution in original and translated text. *Language and Computers*, *37*(1), 30–44.

Brooke, J., & Hirst, G. (2012). Measuring Interlanguage: Native Language Identification with L1-influence Metrics. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* Istanbul, Turkey. European Language Resources Association (ELRA).

Burrows, J. (2002a). The Englishing of Juvenal: computational stylistics and translated texts. *Style*, *36*(4), 677–699.

Burrows, J. (2002b). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, *17*(3), 267–287.

Carpuat, M. (2009). One translation per discourse. In *Proc. of the NAACL HLT workshop on Semantic Evaluation*, pp. 19–26.

Carter, D., & Inkpen, D. (2012). Searching for poor quality machine translated text: learning the difference between human writing and machine translations. *Advances in Artificial Intelligence*, 49–60.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2003). TiMBL: Tilburg memory based learner, version 5.0, reference guide. *ILK Research Group Technical Report Series*.

Dale, E., & Chall, J. (1948). A formula for predicting readability: Instructions. *Educational Research Bulletin*, 37–54.

El-Fiqi, H., Petraki, E., & Abbass, H. (2011). A computational linguistic approach for the identification of translator stylometry using Arabic-English text. In *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, pp. 2039–2045. IEEE.

Flesch, R. (1948). A new readability yardstick.. *Journal of applied psychology*, *32*(3), 221.

Garside, R., Leech, G., McEnery, T., et al. (1997). *Corpus annotation: linguistic information from computer text corpora*. Longman.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. *Translation studies in Scandinavia*, 88–95.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Ilisei, I., & Inkpen, D. (2011). Translationese Traits in Romanian Newspapers: A Machine Learning Approach. *International Journal of Computational Linguistics and Applications*.

Ilisei, I., Inkpen, D., Corpas Pastor, G., & Mitkov, R. (2010). Identification of Translationese: A Machine Learning Approach. *Computational Linguistics and Intelligent Text Processing*, 503–511.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, 137–142.

Kochmar, E. (2011). Identification of a writers native language by error analysis. Master's thesis, University of Cambridge.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, Vol. 5.

Koppel, M., Argamon, S., & Shimoni, A. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, *17*(4), 401–412.

Koppel, M., & Ordan, N. (2011). Translationese and its dialects. *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA*.

Kurokawa, D., Goutte, C., & Isabelle, P. (2009). Automatic Detection of Translated Text and its Impact on Machine Translation. In *Proceedings of the XII MT Summit,Ottawa, Ontario, Canada*. AMTA.

Lauttamus, T., Nerbonne, J., & Wiersma, W. (2007). Detecting Syntactic Contamination in Emigrants: The English of Finnish Australians. *SKY Journal of Linguistics*, *20*, 273–307.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, *43*, 557–570.

Laviosa-Braithwaite, S. (1997). Investigating simplification in an English comparable corpus of newspaper articles. *Klaudy and Kohn*, *1997*, 531–540.

Lembersky, G., Ordan, N., & Wintner, S. (2011). Language Models for Machine Translation: Original vs. Translated Texts. *Empirical Methods in Natural Language Processing, Edinburgh, Scotland, 2011*.

Lembersky, G., Ordan, N., & Wintner, S. (2012). Adapting translation models to translationese improves SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France*, p. 255.

Li, D., Zhang, C., & Liu, K. (2011). Translation Style and Ideology: a Corpus-assisted Analysis of two English Translations of Hongloumeng. *Literary and Linguistic Computing*, *26*(2), 153.

Luyckx, K., & Daelemans, W. (2008). Personae: a corpus for author and personality prediction from text. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*.

Lynch, G., & Vogel, C. (2009). Chasing the Ghosts of Ibsen: A Computational Stylistic Analysis Of Drama in Translation. In *Digital Humanities 2009: University of Maryland, College Park, MD, USA*, p. 192. ALLC/ACH.

Mairesse, F., & Walker, M. (2008). Trainable generation of big-five personality styles through data-driven parameter estimation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 165–173.

Michel, J., Shen, Y., Aiden, A., Veres, A., Gray, M., Pickett, J., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176.

Mikhailov, M., & Villikka, M. (2001). Is there such a thing as a translators style?. In *Proceedings of Corpus Linguistics 2001, Lancaster, UK*, pp. 378–385.

Murphy, B., & Vogel, C. (2007). The syntax of concealment: reliable methods for plain text information hiding. *Security, Steganography and Watermarking of Multimedia Contents IX. Proceedings of SPIE-IS&T Electronic Imaging SPIE*, *6505*.

Nerbonne, J., & Wiersma, W. (2006). A measure of aggregate syntactic distance. In *Proceedings of the Workshop on linguistic Distances*, pp. 82–90. Association for Computational Linguistics.

Olohan, M. (2001). Spelling out the optionals in translation: a corpus study. *UCREL technical papers*, *13*, 423–432.

Olohan, M. (2008). Leave it out! Using a Comparable Corpus to Investigate Aspects of Explicitation in Translation.. *Cadernos de Tradução*, *1*(9), 153–169.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics.

Pastor, G., Mitkov, R., Afzal, N., & Pekar, V. (2008). Translation universals: do they exist? A corpus-based NLP study of convergence and simplification. In *8th AMTA conference*, pp. 75–81.

Popescu, M. (2011). Studying Translationese at the Character Level. In *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP'2011). Hissar, Bulgaria.*

Puurtinen, T. (2003). Genre-specific features of translationese? Linguistic differences between translated and non-translated Finnish children's literature. *Literary and Linguistic Computing*, *18*(4), 389–406.

Pym, A., Shlesinger, M., & Simeoni, D. (2008). Beyond Descriptive Translation Studies. *Amsterdam and Philadelphia: John Benjamins*.

Quinlan, J. (1993). *C4. 5: programs for machine learning*. Morgan Kaufmann.

Quinlan, J. (1996). Improved use of continuous attributes in C4. 5. *Arxiv preprint cs/9603103*.

Ramakrishnan, N. (2009). C4. 5. *The top ten algorithms in data mining. Chapman & Hall/CRC, Boca Raton (FL)*, 1–19.

Rybicki, J. (2006). Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations. *Literary and Linguistic Computing*, *21*(1), 91–103.

Rybicki, J. (2012). The great mystery of the (almost) invisible translator. *Quantitative Methods in Corpus-Based Translation Studies: A Practical Guide to Descriptive Translation Research*, 231.

Santos, D. (1995). On grammatical translationese. In *Short papers presented at the Tenth Scandinavian Conference on Computational Linguistics*, pp. 59–66.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of international conference on new methods in language processing*, Vol. 12, pp. 44–49. Manchester, UK.

Smith, E., & Senter, R. (1967). Automated readability index.. *AMRL-TR. Aerospace Medical Research Laboratories (6570th)*, 1.

Stein, S., & Argamon, S. (2006). A mathematical explanation of Burrows Delta. In *Proceedings of Digital Humanities 2006*.

Tan, P., Steinbach, M., Kumar, V., et al. (2006). *Introduction to Data Mining*. Pearson Addison Wesley Boston.

Thomas, M., Pang, B., & Lee, L. (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 327–335. Association for Computational Linguistics.

van Halteren, H. (2008). Source Language Markers in EUROPARL Translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 937–944. Coling 2008 Organizing Committee.

Venuti, L. (1995). *The translator's invisibility: A history of translation*. Routledge.

Wang, Q., & Li, D. (2012). Looking for translator's fingerprints: a corpus-based study on Chinese translations of Ulysses. *Literary and Linguistic Computing*.

Witten, I., Frank, E., & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Wong, S., & Dras, M. (2009). Contrastive Analysis and Native Language Identification. In *Australasian Language Technology Association Workshop 2009*, p. 53. ALTA.

Wong, S., & Dras, M. (2011). Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1600–1610. Association for Computational Linguistics.

Wu, X., & Kumar, V. (2009). *The Top Ten Algorithms in Data Mining*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Taylor & Francis.

Zhang, H., & Su, J. (2004). Naive Bayesian classifiers for ranking. *Machine Learning: ECML 2004*, 501–512.

Appendix A

First Appendix

A.1 Corpora

Title	SL	Author	Translator	Date	File
Israel Without Clichés	EN	Tony Judt	Tony Judt	10/06/10	orig0001
But Deng Is The Leader To Celebrate	EN	Ezra F. Vogel	Ezra F. Vogel	03/10/09	orig0002
The Ice Storm	EN	Gauti Kristmannsson	Gauti Kristmannsson	16/10/08	orig0003
Erin Go Bust	EN	John Banville	John Banville	16/10/08	orig0004
Back To The Blitz	EN	Andrew O'Hagan	Andrew O'Hagan	16/10/08	orig0005
The Okinawa Question	EN	TZE.M.Loo	TZE.M.Loo	10/06/10	orig0006
One Myth:Many Pakistans	EN	Ali Sethi	Ali Sethi	11/06/10	orig0007
Europe's Banks	EN	Guy Verhofstadt	Guy Verhofstadt	01/06/10	orig0008
Pumpkin Eaters	EN	Peter Mayle	Peter Mayle	24/10/09	orig0009
South Korea Rising	EN	Philip Bowring	Philip Bowring	23/10/09	orig0010
Cyprus and 'Chosen Trauma	EN	H.D.S. Greenway	H.D.S. Greenway	20/10/09	orig0011
The Myth Of The New India	EN	Pankaj Mishra	Pankaj Mishra	06/07/06	orig0012
Defenders of the Faith	EN	Slavoj Zizek	Slavoj Zizek	12/03/06	orig0013
Pirates of The Mediterranean	EN	Robert Harris	Robert Harris	30/09/06	orig0014
Reasonable Doubt	EN	Rebecca N. Goldstein	Rebecca N. Goldstein	29/7/06	orig0015
Democracy's Double Standard	EN	Hossein Derakhshan	Hossein Derakhshan	28/1/06	orig0016
Castro At The Bat	EN	Roberto G. Echevarria	Roberto G. Echevarria	11/01/06	orig0017
Mexico's Fast Diagnosis	EN	Julio Frenk	Julio Frenk	30/4/06	orig0018
A Past That Makes Us Squirm	EN	Craig Childs	Craig Childs	02/01/07	orig0019
A Way To Peace In Mexico	EN	Jorge G. Castaneda	Jorge G. Casaneda	06/09/06	orig0020
Why Israel Feels Threatened	EN	Benny Morris	Benny Morris	30/12/08	orig0021
Silence = Despotism	EN	Alejandro Toledo	Alejandro Toledo	06/06/07	orig0022
The Winner In Honduras:Chavez	EN	Alvaro Vargas Llosa	Alvaro Vargas Llosa	30/6/09	orig0023
Who Cares About Zelaya?	EN	Roger Marin Neda	Roger Marin Neda	07/07/09	orig0024
A Holiday To End All Wars	EN	Alexander Watson	Alexander Watson	11/11/08	orig0025
The Fictions Of Günter Grass	EN	Peter Gay	Peter Gay	20/8/06	orig0026
Back When Spies Played By The Rules	EN	David Kahn	David Kahn	13/1/06	orig0027
The Memory Hole	EN	David Shenk	David Shenk	03/11/06	orig0028
Guiding Germany's Unification	EN	Robert B. Zoellick	Robert B. Zoellick	06/11/09	orig0029
It takes A Crisis To Make A Continent	EN	Gabor Steingart	Gabor Steingart	21/5/10	orig0030
Change Germans Can't Believe In	EN	Susan Neiman	Susan Neiman	26/07/08	orig0031
Save The Dresden Elbe Valley	EN	Guenter Blobel	Guenter Blobel	04/06/09	orig0032
To Resist Hitler and Survive	EN	Susan Nieman	Susan Nieman	03/02/08	orig0033
North Korea Will Never Disarm	EN	B.R Myers	B.R Myers	28/5/09	orig0034
Leave Swiss Banks Alone	EN	Pierre Bessard	Pierre Bessard	02/04/09	orig0035
Departure	EN	Kumiko Makihara	Kumiko Makihara	19/6/09	orig0036
20 Years Of Collapse	EN	Slavoj Zizek	Slavoj Zizek	09/11/09	orig0037
To Russia With Tough Love	EN	Strobe Talbot	Strobe Talbot	26/2/05	orig0038
Road Maps and Dead Ends	EN	Yossi Beilin	Yossi Beilin	20/10/05	orig0039
Happy Birthday Nikita Kruschev	EN	Nina L. Khrushcheva	Nina L. Khrushcheva	16/4/05	orig0040
The Great Unifier	EN	Jaroslav Pelikan	Jaroslav Pelikan	04/04/05	orig0041
Jihad's Fresh Face	EN	Waleed Ziad	Waleed Ziad	16/9/05	orig0042
Stop Blaming Putin and Start Helping Him	EN	Fiona Hill	Fiona Hill	10/09/04	orig0043
Living in the Dead Zone	EN	Martin Cruz Smith	Martin Cruz Smith	22/12/04	orig0044
New Kids on The Bloc	EN	Veronica Khokhlova	Veronica Khoklova	26/11/04	orig0045
Arise Ye Prisoners Of Starvation	EN	Bill Keller	Bill Keller	23/2/02	orig0046
China's Workers Are Stirring	EN	Han Dongfang	Han Dongfang	17/6/10	orig0047
A Warning On Iraq From a Friend	EN	Jean-David Levitte	Jean-David-Levitte	14/2/03	orig0048
Workers Of The World Relax	EN	Alain De Botton	Alain De Botton	06/09/04	orig0049
Give The Chechens A Land Of Their Own	EN	Richard Pipes	Richard Pipes	9/904	orig0050
Why Chile Is Hopeful	EN	Ariel Dorfman	Ariel Dorfman	11/09/04	orig0051
The Citizen Stranger	EN	Jonathan Rosen	Jonathan Rosen	09/01/04	orig0052
The Siren Call Of Africa	EN	Ken Wiwa	Ken Wiwa	18/9/04	orig0053
Picking A Fight With Venezuela	EN	Michael Shifter	Michael Shifter	20/9/04	orig0054
Poison Politics In Ukraine	EN	Jason T. Shaplen	Jason T. Shaplen	25/9/04	orig0055
The International Pastime	EN	Robert Whiting	Robert Whiting	02/10/04	orig0056

Table A.1: NYT corpus: part 1

Title	SL	Author	Translator	Date	File
Two Peoples:One State	EN	Michael Tarazi	Michael Tarazi	04/10/04	orig0057
Africa Earned Its Debt	EN	Robert Guest	Robert Guest	06/10/04	orig0058
Spray Now Or Pay Later	EN	Jan Egeland	Jan Egeland	06/10/04	orig0059
The Next Green Revolution	EN	Pedro Sanchez	Pedro Sanchez	06/10/04	orig0060
Saving Central Asia	EN	Paul Quinn-Judge	Paul Quinn-Judge	20/6/10	orig0061
A New Path For Japan	EN	Yukio Hatoyama	Yukio Hatoyama	26/809	orig0062
The Call From The Swiss Minaret	EN	Claudio Cordone	Claudio Cordone	01/12/09	orig0063
A Slaughter Waiting To Happen	EN	Lakhdar Brahmi	Lakhdar Brahmi	20/3/09	orig0064
Pakistan's Slow Motion Emergency	EN	Ali Sethi	Ali Sethi	02/12/07	orig0065
Island Of Lost Girls	EN	Dea Birkett	Dea Birkett	29/10/04	orig0066
Under The Cover Of Islam	EN	Irshad Manji	Irshad Manji	18/11/04	orig0067
Behind Enemy Lines	EN	Antoine Audouard	Antonine Audouard	03/01/05	orig0068
The Red White And Blue Guide	EN	Francois Simon	Francois Simon	04/03/05	orig0069
Can Hezbollah Go Straight?	EN	Michael Young	Michael Young	09/04/05	orig0070
Guilty Of Popularity	EN	Carmen Boullosa	Carmen Boullosa	19/4/05	orig0071
Woe Canada	EN	David Frum	David Frum	19/4/05	orig0072
Just Say Non	EN	Stephen Clarke	Stephen Clarke	27/5/05	orig0073
Our Ally Our Problem	EN	Peter Bergen	Peter Bergen	08/07/05	orig0074
The Wages Of Denial	EN	Courtney Angela Brkic	Courtney Angela Brkic	11/07/05	orig0075
The Danger Next Door	EN	Seth G. Jones	Seth G. Jones	23/9/05	orig0076
The Revolt Of Ennui	EN	Antoine Audouard	Antoine Audouard	09/11/05	orig0077
Agent Provocateur	EN	Kamila Shamsie	Kamila Shamsie	15/2/06	orig0078
Mind Over Splatter	EN	Don Foster	Don Foster	19/2/06	orig0079
Israel's Tragedy Foretold	EN	Gershom Gorenberg	Gershom Gorenberg	10/03/06	orig0080
Italy's Natural Selection	EN	Gianni Riotta	Gianni Riotta	13/4/06	orig0081
A Lobby Not A Conspiracy	EN	Tony Judt	Tony Judt	19/4/06	orig0082
Israel's Invasion:Syria's War	EN	Michael Young	Michael Young	14/7/06	orig0083
No Sex Please:We're French	EN	Stephen Clarke	Stephen Clarke	23/3/07	orig0084
Latin Lovers	EN	Pamela Druckerman	Pamela Druckerman	06/04/07	orig0085
Friend Or Faux	EN	Oliver Roy	Oliver Roy	15/5/07	orig0086
Not Much Kinder And Gentler	EN	Stephen Sestanovich	Stephen Sestanovich	03/02/05	orig0087
The War We Haven't Finished	EN	Frank C. Carlucci	Frank C. Carlucci	22/2/05	orig0088
A Wall Of Faith And History	EN	David Fromkin	David Fromkin	24/3/05	orig0089
The Vatican's Sin Of Omission	EN	Arthur Hertzberg	Arthur Hertzberg	14/5/05	orig0090
From the Ashes	EN	Daniel Libeskind	Daniel Libeskind	23/6/05	orig0091
The Russian Card	EN	Rose Gottemoeller	Rose Gottemoeller	03/05/05	orig0092
The Persian Complex	EN	Abbas Amanat	Abbas Amanat	25/5/06	orig0093
Bar None	EN	Jack Turner	Jack Turner	28/8/06	orig0094
Letter From Europe	EN	Sam Ryan	Sam Ryan	06/12/06	orig0095
The Politics Of Eurovision	EN	Duncan J. Watts	Duncan J. Watts	22/5/07	orig0096
Losing Count	EN	Thane Rosenbaum	Thane Rosenbaum	14/6/07	orig0097
Plunder Goes On Tour	EN	Allan Gerson	Allan Gerson	23/2/08	orig0098
China's Inside Game	EN	April Rabkin	April Rabkin	02/07/08	orig0099
Summer's Last Call	EN	Fiona Maazel	Fiona Maazel	06/07/08	orig0100
Subprime Europe	EN	Liaquat Ahamed	Liaquat Ahamed	08/03/09	orig0101

Table A.2: NYT corpus: part 2

Title	SL	Author	Translator	Date	File
A Table For Tyrants	EN	Vaclav Havel	Vaclav Havel	11/05/09	unknown0001
Where History's March Is a Funeral Procession	PL	Olga Tokarzuk	Antonia Lloyd-Jones	16/04/10	trans0001
Euro Trashed	DE	Joachim Starbatty	John Cullen	28/3/10	trans0002
Perestroika Lost	RU	Gorbachev	Pavel Palazhchenko	14/03/10	trans0003
Russia Never Wanted A War	RU	Gorbachev	Pavel Palazchenko	19/8/08/	trans0004
Two First Steps on Nuclear Weapons	RU	Gorbachev	Pavel Palazchenko	24/9/09	trans0005
For Every Iraqi Party:an Army of Its Own	Ar	Najim Abed Al-Jabouri	Sterling Jensen	29/10/09	trans0006
A Flash Of Memory	JP	Issey Miyake	Staff	14/7/10	trans0007
In China the Red Flags Still Fly for Mao	CN	Kang Zhengguo	Xiaoxuan Li	04/10/09	trans0008
In Gold We Trust	DE	Christoph Peters	John Cullen	16/10/08	trans0009
The Mexican Evolution	ES	Enrique Krauze	Hank Heifetz	24/3/09	trans0010
Obama at the Gate	DE	Christoph Peters	John Cullen	17/9/08	trans0011
Fight Fire With a Cease-Fire	HB	David Grossman	Haim Watzman	30/12/08	trans0012
Why The Muslim World Cannot Hear Obama	AR	Alaa Al Aswany	Geoff D. Porter	08/02/09	trans0013
Time Out of Mind	DE	Stefan Klein	Shelley Frisch	07/03/08	trans0014
Paris Isn't Burning	FR	Corinne Maier	The Times	30/12/07	trans0015
Russia's Last Hope	RU	Victor Erofeyev	IHT	29/2/08	trans0016
My Views Of Israel	FR	Bernard Henri Levy	Charlotte Mandell	06/08/06	trans0017
Italy's American Baggage	IT	Andrea Camilleri	Stephen Sartarelli	23/8/07	trans0018
A Prisoner of the Nobel	DE	Daniel Kehlmann	Ross Benjamin	20/8/06	trans0019
Recounting Our Way to Democracy	ES	Andres Obrador	Rogelio Ramirez	11/08/06	trans0020
There's A Word For People Like You	FR	Martine Rousseau	The Times	06/09/07	trans0021
What We See In Hugo Chavez	ES	Luisa Valenzuela	Esther Allen	17/3/07	trans0022
Waiting For Freedom Messing It Up	PL	Adam Michnik	Irena Gross	25/3/07	trans0023
The Way We War	HB	Etgar Keret	Sondra Silverstone	18/7/06	trans0024
Man in the Middle	FR	Tahar Ben Jelloun	The Times	03/09/06	trans0025
Bringing Mexico Closer To God	ES	Enrique Krauze	Natasha Wimmer	28/6/06	trans0026
Our Fetid City	IT	Elena Ferrante	Ann Goldstein	15/1/08	trans0027
Why I Parted Ways With Chavez	ES	Raul Isaias Baduel	Kristina Cordero	01/12/07	trans0028
Chile's Rising Waters and Frozen Avocados	ES	Antonio Skarmeta	Kristina Cordero	23/12/07	trans0029
Our Moral Footprint	CZ	Vaclav Havel	Gerald Turner	27/9/07	trans0030
The Great Swiss Meltdown	DE	Peter Stamm	Philip Boehm	29/7/07	trans0031
The View From Guantanamo	UI	Abu Bakker Qassim	Nury Turael	17/9/06	trans0032
Money Can't Buy Us Democracy	FA	Akbar Ganji	Unknown	01/08/06	trans0033
Swiss Miss	DE	Peter Stamm	Philip Boehm	10/17/07	trans0034
Cloudy With A Chance of Climate Change	IC	Kristin Steinsdottir	Gauti Kristmannsson	04/03/07	trans0035
On The Road With Bush And Chavez	ES	Fernando Baez	Kristina Cordero	11/03/07	trans0036
Another Last Chance To Change Your Life	FR	Pascal Bruckner	The Times	01/01/07	trans0037
What They Are Reading About in Moscow	RU	Solomon Volkov	Antonina W. Bois	19/7/02	trans0038
Germans Are From Mars Italians Are From Venus	IT	Roberto Pazzi	Ann McGarrell	13/7/03	trans0039
China's Selective Memory	CN	Pu Zhiqiang	Perry Link	28/4/05	trans0040
The Gravest Generation	DE	Guenter Grass	UPS Translations	07/05/05	trans0041
How Russia Lost World War II	RU	Victor Erofeyev	Andrew Bromfield	10/05/05	trans0042
Working Hard at Nothing All Day	FR	Corinne Maier	The Times	05/09/05	trans0043
The New Berlin Wall	DE	Peter Schneider	Philip Boehm	04/12/05	trans0044
Scarves and Symbols	FR	Guy Coq	The Times	30/1/04	trans0045
The Basque Spring	ES	Bernardo Atxaga	Esther Allen	29/3/06	trans0046
French Twist	FR	Corinne Maier	The Times	31/3/06	trans0047
Dj Vu All Over Again	FR	Abdellah Taia	The Times	13/4/06	trans0048
What Russia Knows Now	RU	Victor Erofeyev	Andrew Bromfield	11/09/04	trans0049
My Tortured Inheritance	ES	Rafael Gumucio	Kristina Cordero	13/12/04	trans0050
Castro's Latest Victim	ES	Vladimiro Roca	Joseph McSpedon	22/3/04	trans0051
Fictions Embraced by an Israel at War	HB	David Grossman	Haim Watzman	01/10/02	trans0052
Smoking And Fuming	ES	Javier Marias	Kristina Cordero	22/1/06	trans0053
Measuring the Distance Across The Sea Of Japan	JP	Yomota Inuhiko	Ioannis Mentzas	10/10/02	trans0054
Free Trade Won't Free Cuba	ES	Claudia Marquez Linares	The Times	06/11/03	trans0055
Even in a New Russia:Stalin Shadows Putin	RU	Victor Erofeyev	Andrew Bromfield	08/03/03	trans0056

Table A.3: NYT corpus: part 3

Title	SL	Author	Translator	Date	File
Rusty And Radioactive	RU	Ashot Sarkissov	Ilya Feliciano	30/9/03	trans0057
The Country America Cannot See	KO	Mun Yol Yi	Bruce Fulton	27/7/03	trans005
Where's The Boeuf?	FR	Vincent Tournier	The Times	27/5/05	trans0059
The Dispossessed	FR	Elie Wiesel	The Times	21/8/05	trans0060
Past Wrongs:Future Rights	ES	Enrique Krauze	Natasha Wimmer	10/08/04	trans0061
Harry Potter:Market Whiz	FR	Ilias Yocaris	The Times	18/7/04	trans0062
Why The Next Pope Needs To Be Italian	IT	Roberto Pazzi	Ann Goldstein	11/01/04	trans0063
The French Disconnection	FR	Corinne Maier	The Times	08/01/06	trans0064
Stupor in Our Time	HB	Etgar Keret	Sondra Silverston	27/3/06	trans0065
When A Godfather Becomes Expendable	IT	Andrea Camilleri	Stephen Sartarelli	21/4/06	trans0066
Putin's Baby Love	RU	Viktor Erofeyev	IHT	20/5/06	trans0067
Praise the Lord and Pass a Budget	ES	Mayra Montero	Edith Grossman	20/5/06	trans0068
How To Remember:How To Forget	ES	Javier Marias	Esther Allen	11/09/04	trans0069
Ordinary Men	RU	Ludmila Ulitskaya	Peter Evgenev	08/11/04	trans0070
A President Who Listened	RU	Michael Gorbachev	Pavel Palazhchenko	07/06/04	trans0071
We Don't Want To Be Alone	ES	Antonio Munoz Molina	Catherine Rendon	20/3/04	trans0072
Feeling London's Bombs in Madrid	ES	Javier Marias	Kristina Cordero	10/07/05	trans0073
All Rock:No Action	FR	J.Claude Shanda Tonme	The Times	15/7/05	trans0074
Illusions of a Separate Peace	HB	David Grossman	Haim Watzman	12/07/02	trans0075
Russia and the Wages of Terror	RU	Anna Politkovskaya	Robert Coalson	08/11/02	trans0076
Always Darkness Visible	HB	Aharon Appelfeld	Barbara Harshav	27/1/05	trans0077
Winning Back Europe's Heart	DE	Elfriede Jelinek	Martin Chalmers	20/2/05	trans0078
Poland's Holy Father	PL	Stefan Chwin	Phillip Boehm	05/04/05	trans0079
The Pope Without a Country	DE	Martin Mosebach	Phillip Boehm	30/4/05	trans0080
The Emptiest Cradle	NL	P.F.Thomese	Sam Garrett	19/6/05	trans0081
Country Girl	DE	Jana Hensel	Kurt Beals	22/11/05	trans0082
Senora Presidente?	ES	Rafael Gumucio	Kristina Cordero	09/12/05	trans0083
No Soul on Ice	DE	Katarina Witt	Christina Knight	22/2/06	trans0084
Riding My Father's Motorcycle	ES	Aleida Guevara	Pilar Aguilera	09/10/04	trans0085
How the World Watched the Returns;Oil And Politics	ES	Ana Teresa Torres	Esther Allen	08/11/04	trans0086
Magic And Realism	ES	Mayra Montero	Edith Grossman	30/11/04	trans0087
Putin's Pursuit of the National Idea	RU	Solomon Volkov	Antonina W. Bouis	14/2/02	trans0088
A Trap Israel Sets for Itself	HB	Meir Shalev	Barbara Harshav	28/5/01	trans0089
Conquering Europe Word For Word	DE	Peter Schneider	Phillip Boehm	01/05/01	trans0090
Germany's Newfound Peace	DE	Peter Schneider	Phillip Boehm	04/08/97	trans0091
Trapped in a Body at War With Itself	HB	David Grossman	Haim Watzman	25/8/01	trans0092
For Germans:Guilt Isn't Enough	DE	Peter Schneider	Leigh Hafrey	05/12/96	trans0093
A City Indebted To Its migrs	RU	Solomon Volkov	Antonina W. Bouis	07/09/01	trans0094
New Democracies for Old Europe	CZ	Vaclav Havel	Paul Wilson	17/10/1993	trans0095
Human Currency in Mexico's Drug Trade	ES	Mario Bellantin	Kurt Hollander	28/3/10	trans0096
Best Invention:How The Bean Saved Civilization	IT	Umberto Eco	William Weaver	18/4/1999	trans0097
No Hurt Feelings In Germany	DE	Christoph Peters	John Cullen	04/04/09	trans0098
Denying History Disables Japan	JP	Kenzaburo Oe	Hiroaki Sato	02/09/95	trans0099
Switzerlands Invisible Minarets	DE	Peter Stamm	Philip Boehm	05/12/09	trans0100
Ordinary Men	RU	Ludmila Ulitskaya	Peter Evgenev	08/11/04	trans0101

Table A.4: NYT corpus: part 4

A.2 Code

```
WDT 10 39,246,297,668,829,878,1259,1772,1876,2264
POS 9 343,407,773,995,1217,1346,1714,1983,2084
UH 7 222,1739,1741,1786,2038,2197,2199
JJS 7 433,513,818,1240,1427,1566,1575
EX 6 155,778,1264,1788,1811,2378
RBR 4 799,1629,1839,2008
JJR 3 74,365,2355
PDT 2 538,2050
```

Figure A.1: Sample .t1 file

```
went VBD go
from IN from
one CD one
to TO to
another DT another
,,,
keeping VBG keep
up RP up
our PP$ our
spirits NNS spirit
and CC and
lending VBG lend
a DT a
hand NN hand
wherever WRB wherever
```

Figure A.2: sample .tagged file

This section contains the Java code which generates the document-level features used in the experiments. Before the metrics can be calculated for each file, word frequency and postag frequency files are required. This program expects these files in a particular format:

A *.t1* file in Figure A.1 contains a sorted list of POS tags, in this case in the first column, followed by the frequency in the second column and the next column containing a list of the position of these tags in the file. The *.w1* file has the same format as the *.t1* file, with the exception that this particular file contains the frequency and position of single words instead of POS tags.

The *.tagged* file in Figure A.2 is a pre-processing step before the *.t1* file, which consists of the raw output from the TreeTagger POS tagger, a list of tokens in order with their assigned POS and lemma in adjacent columns.

```
1  import java.io.* ;
2  import java.util.*;
3  /* This Java program generates a list of document-level statistics
4   * when given a directory of text files.
5   * Text files should be in UTF8 format and plain-text only,
6   * free from any XML markup.
7   * Files containing word frequencies and POS unigram frequencies are
8   * required to run this program, these
9   * are generated externally and placed in the same directory.
10  *
```

```
11 * Author: Gerard Lynch
12 * Date: January 2012
13 */
14     public class GenerateARFFDir {
15    // Still to implement: Friday November 19th 2010
16    // contractions such are there's. it's etc
17     public static Hashtable wfpair;
18     // storage for word frequency pair items
19     public static Hashtable tfpair;
20     // storage for tag frequency pair items
21     public static Hashtable ttpair;
22     // storage for lemma frequency triples
23     public static  Vector <String> taggedlist;
24     // for an in-order list of POS tags for the file.
25    // Should probably read these in from a file
26      public static String [] adjs =  {"JJ","JJR","JJS"};
27      public static String [] nouns = {"NN","NNS","NPS","NP"};
28      public static String [] dets ={"WDT","DT"};
29      public static String [] conj ={"CC","XX"};
30      public static String [] preps ={"IN","XX"};
31      // Three types of finite verbs in English
32      // ,those which are inflected for person and tense
33      // VBD, VBZ, VBP in Penn Tagset
34      public static String [] fverbs = {"VBD","VBZ","VBP"};
35      public static String [] numerals = {"CD","XX"};
36      public static String [] pronouns = {"PP","PP$","WP","WP$"};
37      // Lexical words are classed by Ilisei et al (2010)
38      //as verbs,nouns,adjectives,adverbs and numerals
39      public static String [] lex = {"NN","NNS","VB","VBD","VBG","VBN","VBP","VBZ","NPS",
40      "NP","JJR","JJS","JJ","RB","RBR","RBS","CD","XX"};
41      public static double dmarkers ;
42      public static String [] arffheader  ;
43      public static String [] discoursemarkers = {"therefore","as a result","consequently",
44      "moreover","furthermore","in addition",
45      "however","nonetheless","nevertheless","on the other hand","while","whereas"
46      ,"with regard to","regarding","as regards","as for"};
47      // Grammatical words in Ilisei et al (2010) are classed as
48      // determiners, prepositions, auxiliary verbs, pronouns and interjections
49      // *What about TO in English?
50      public static String [] gramm = {"WDT","DT","PDT","IN","UH","MD","PP","PP$","WP$","WP"};
51      public static void main(String [] args){
52
53     try{
54     // Before running this script you need the word unigram frequency and tag unigram frequency for
55     // all of the files you wish to convert
56     // This represents what directory you wish to convert
57     String dir = args[0];
58     // Filename for output files
59     String out = args[1];
60     // Set a minimum length in bytes for files
61     //int min = Integer.parseInt(args[2]);
62     File current = new File(dir);
63     File temp ;
64     File output = new File(out);
65     // Get a list of the files in the directory, this sorts the list alphabetically
66     File [] contents = current.listFiles() ;
67     File list = new File(out + ".list");
68     // Store the file split on spaces
69     String [] farray;
70     String [] posarray;
71     // Store the tagged files, word frequency and tag frequency files
72     Vector <POSPair> vppair = new Vector();
```

```
73      Vector <POSPair> pnpair = new Vector();
74      taggedlist = new Vector();
75      wfpair = new Hashtable();
76      tfpair = new Hashtable();
77      ttpair = new Hashtable();
78      // Store the number of unique words or tags
79      int wfpairsize = 0 ;
80      int tfpairsize = 0 ;
81      int ttpairsize = 0 ;
82      POSPair [] pparray ;
83      String pathname ;
84      String stripped ;
85      String soutput = "";
86      String posoutput = "";
87      String token = "";
88      // File I/O
89      BufferedReader br;
90      FileReader fr;
91      POSPair ppair ;
92      FileOutputStream fout ;
93      PrintStream p ;
94      FileOutputStream arffout;
95      PrintStream ap ;
96      FileOutputStream lout;
97      PrintStream pl;
98      Integer a ;
99      FileOutputStream arout = new FileOutputStream(new File(args[1] + ".arff"));
100     String line = "";
101     ap = new PrintStream(arout);
102     lout = new FileOutputStream(list);
103     pl = new PrintStream(lout);
104    for(int i = 0;i < contents.length;i++){
105      if(!(isInvalid(contents[i]))){
106     System.out.println(contents[i].toString());
107     pl.println(contents[i].toString());
108     }
109
110
111     }
112    lout.close();
113   for (int i = 0; i < contents.length;i++) {
114
115         temp = contents[i] ;
116        // if the file is valid
117         dmarkers = 0.0;
118         if(temp.isFile() && !(isInvalid(temp))){
119         // read in the text of the file
120          fr = new FileReader(temp);
121          br = new BufferedReader(fr);
122         // while there is text in the file
123          while(br.ready()){
124           line = br.readLine();
125           dmarkers = dmarkers + countDiscourseMarkers(line);
126           soutput += line + "\n";
127           }
128         // close BufferedReader
129          br.close();
130          fr.close();
131       // Convert soutput to String array
132      farray = soutput.split(" ");
133     //  Read in tagged file
134      int incr = 0;
```

```
135     Integer tempinteger ;
136     fr = new FileReader(temp.toString() + ".tagged");
137     br = new BufferedReader(fr);
138         while(br.ready()){
139           posoutput = br.readLine();
140          // System.out.println(posoutput);
141           posarray = posoutput.split("\t");
142           taggedlist.add(posarray[1]);
143           if(ttpair.containsKey(posarray[2])){
144
145            tempinteger = (Integer)ttpair.get(posarray[2]);
146            incr = tempinteger.intValue();
147            incr++;
148            ttpair.put(posarray[2],new Integer(incr));
149           }
150           else
151           {
152
153           ttpair.put(posarray[2],new Integer(1));
154
155           }
156
157           }
158    br.close();
159    fr.close();
160    ttpairsize = ttpair.keySet().size(); // get the number of unique lemmas in the file.
161    System.out.println("Number of lemmas: " + ttpairsize);
162     // Read in tag frequency file
163       fr = new FileReader(temp.toString() + ".t1");
164     br = new BufferedReader(fr);
165     tfpairsize = Integer.parseInt(br.readLine());
166    while(br.ready()){
167           posoutput = br.readLine();
168         //  System.out.println(posoutput);
169           posarray = posoutput.split("\t");
170        //  ppair = new POSPair(posarray[1],posarray[0]);
171        //   System.out.println(ppair);
172           wfpair.put(posarray[0],new Integer(Integer.parseInt(posarray[1])));
173           }
174    br.close();
175    fr.close();
176       //read in word frequency file.
177   fr = new FileReader(temp.toString() + ".w1");
178     br = new BufferedReader(fr);
179     wfpairsize = Integer.parseInt(br.readLine());
180    while(br.ready()){
181           posoutput = br.readLine();
182         //  System.out.println(posoutput);
183           posarray = posoutput.split("\t");
184        //  ppair = new POSPair(posarray[1],posarray[0]);
185         //  System.out.println(ppair);
186           tfpair.put(posarray[0],new Integer(Integer.parseInt(posarray[1])));
187           }
188    br.close();
189    fr.close();
190
191
192
193
194
195       //for each item in the Vector, add it to the vector if it is a proper noun.
196    System.out.println(getARFFLine(out,tfpairsize,wfpairsize,ttpairsize));
```

```
197     ap.println(getARFFLine(out,tfpairsize,wfpairsize,ttpairsize));
198
199
200
201 //}
202
203     } // if (temp.isFile()
204
205 //soutput ="";
206 //vppair.clear();
207 //pnpair.clear();
208 wfpair.clear();
209 tfpair.clear();
210 ttpair.clear();
211 taggedlist.clear();
212 dmarkers = 0.0;
213 }//for each file in directory
214 ap.close();
215 }//try
216
217   catch(IOException e){
218
219   e.printStackTrace();
220
221
222   }//catch
223
224  }//main method
225
226
227
228 public static boolean isPunctuation(String s){
229 boolean punc = false;
230 char c ;
231 if(s.length() > 1){
232
233   if(s.equals("SEN")){
234
235
236   punc = true;
237
238   }
239
240
241 }
242 else{
243
244
245 c = s.charAt(0);
246 if(!(Character.isLetterOrDigit(c) || Character.isWhitespace(c))){
247
248 punc = true;
249 }
250
251
252 }
253 return punc;
254 }
255
256 public static String stripPunc(String s){
257
258 String nopunc = "";
```

```
259 char [] carray = s.toCharArray();
260
261 for(int i =0;i < carray.length;i++){
262
263 if(!(Character.isLetterOrDigit(carray[i]) || Character.isWhitespace(carray[i]))){
264
265 }
266 else{
267
268 nopunc += (new Character(carray[i]).toString());
269
270 }
271
272 }
273 return nopunc;
274
275 }
276 // Simple method to disregard any invalid text files, due for updating
277 public static boolean isInvalid(File f){
278
279     String fs = f.toString();
280
281     boolean b = false ;
282
283     System.out.println(fs) ;
284
285     if (fs.indexOf(".java") > -1 ){
286
287         b = true ;
288     }
289
290     else if (fs.indexOf(".class") > -1 ){
291
292         b = true ;
293     }
294
295      else if (fs.indexOf(".sh") > -1 ){
296
297             b = true ;
298     }
299
300       else if (fs.indexOf(".exe") > -1 ){
301
302              b = true ;
303     }
304     else if (fs.indexOf(".tagged") > -1 ){
305
306              b = true ;
307     }
308       else if (fs.indexOf(".t1") > -1 ){
309
310              b = true ;
311     }
312      else if (fs.indexOf(".w1") > -1 ){
313
314              b = true ;
315     }
316     else if (fs.indexOf(".w!2") > -1 ){
317
318              b = true ;
319     }
320      else if (fs.indexOf(".t2") > -1 ){
```

```
321
322                 b = true ;
323       }
324         else if(fs.indexOf(".pn") > -1){
325
326
327               b = true;
328
329       }
330       else if(fs.indexOf(".arff") > -1){
331
332
333               b = true;
334
335       }
336   else if(fs.indexOf("files1") > -1){
337
338
339               b = true;
340
341       }
342   else if(fs.indexOf("ranklist") > -1){
343
344
345               b = true;
346
347       }
348       return b ;
349
350
351   }
352
353
354   public static double countDiscourseMarkers(String line){
355
356   double disc = 0.0;
357   line = line.toLowerCase();
358   for(int i = 0;i < discoursemarkers.length;i++){
359
360   if(line.indexOf(discoursemarkers[i]) > -1){
361   System.out.println("Found discourse marker in: " + line);
362   disc++;
363   }
364
365
366   }
367
368   return disc;
369
370
371   }
372   public static double getComplexSentenceCount(){
373
374   // scroll through sentences, counting finite verbs
375   int fverbcount = 0;
376   int complexsentcount = 0;
377   for(int i = 0; i < taggedlist.size();i++){
378
379   if(inStringArray(fverbs,taggedlist.get(i))){
380
381   fverbcount++;
382
```

```
383
384 }
385
386 if(taggedlist.get(i).equals("SENT")){
387
388   if(fverbcount > 1){
389
390   complexsentcount++;
391
392
393 }
394 fverbcount = 0;
395 }
396
397
398
399 }
400
401 return complexsentcount;
402 }
403
404 public static double getSimpleComplexRatio(){
405
406 int sentences = getNumberOfSentences();
407
408 double complex = getComplexSentenceCount();
409
410 double simple = sentences - complex ;
411
412 double ratio = simple / complex ;
413
414 return ratio;
415
416 }
417 //Returns percentage of sentences with more than one verb
418 public static double getComplexTotalRatio(){
419
420 int sentences = getNumberOfSentences();
421
422 double complex = getComplexSentenceCount();
423
424 double ratio = sentences / complex;
425
426 return ratio;
427
428 }
429
430 public static double getSimpleTotalRatio(){
431
432 int sentences = getNumberOfSentences();
433
434 double complex = getComplexSentenceCount();
435
436 double simple = sentences - complex ;
437
438 double ratio = sentences / simple;
439
440 return ratio;
441
442 }
443
444
```

```
445 // Searches a string array for a String value
446 public static boolean inStringArray(String [] sa,String s){
447 boolean b = false;
448 for(int i = 0;i < sa.length ;i++){
449
450 if(sa[i].equals(s)){
451
452 b = true;
453
454 }
455
456 }
457
458 return b;
459
460 }
461 /********************
462 * Get the average     *
463 * sentence length     *
464 * for a file          *
465 *                     *
466 ********************/
467
468 public static double getAverageSentenceLength(int total){
469 // declare variable for result
470 double avg = 0.0;
471 int sentences = getNumberOfSentences();
472 if(sentences > 0){
473 avg = total / sentences;
474 }
475 else{
476 avg = 1;
477 }
478 // return result
479 return avg;
480
481 }
482
483
484 // Get the number of sentences in the file from the hashtable
485 public static int getNumberOfSentences(){
486 int sent = 0;
487
488 Integer sentences = (Integer)wfpair.get("SENT");
489 if(!(sentences == null)){
490 sent = sentences.intValue();
491 }
492 else{
493 sent = 1;
494 }
495 return sent;
496
497
498
499 }
500 // Get the average word length from a document
501 public static double getAverageWordLength(int total){
502 Object [] arraystring;
503 String s = "";
504 double d = 0;
505 double accum =0;
506 Integer temp ;
```

```
507 Set keys = tfpair.keySet();
508 arraystring =  keys.toArray();
509
510 for(int i =0; i < arraystring.length;i++){
511
512  s = (String) arraystring[i];
513  d = s.length();
514  temp = (Integer)tfpair.get(s);
515  accum += (d * temp.intValue());
516
517
518 }
519 System.out.println("Total Word Length: " + accum);
520 System.out.println("Total Words: " + total);
521 return accum / total;
522
523 }
524 public static double getTypeTokenRatio(int total){
525 Object [] arraystring;
526 Set keys = tfpair.keySet();
527 arraystring =  keys.toArray();
528 double size = arraystring.length;
529 return size / total ;
530
531
532
533 }
534
535 public static double getFreqWordType(String [] list){
536 // Pass in a String array with the word types to count
537 Integer intholder;
538 int value = 0;
539 double end = 0.0;
540
541
542     for(int i = 0;i < list.length;i++){
543
544      if(wfpair.containsKey(list[i])){
545
546        intholder = (Integer)wfpair.get(list[i]);
547        value += intholder.intValue();
548        System.out.println(list[i] + ":" + value);
549
550      }
551 }
552
553 return end + value;
554
555 }
556
557
558 // Information load is given in Ilisei et al (2010) as
559 // the proportion of lexical words to overall tokens
560
561 public static double getInformationLoad(int t){
562
563 double d = 0.0;
564
565 d = getFreqWordType(lex) / t ;
566
567 return d;
568
```

```
569
570 }
571 // Get the ARI(Automated Readability Index) for a text
572 //  ARI = 4.71(total characters/total words) + 0.5(total words/(total sentences)) -21.43
573
574 public static double getARI(int total){
575
576 double ari = 0.0;
577 double firstterm = (4.71 * getAverageWordLength(total));
578 double secondterm = (0.5 * (total / getNumberOfSentences()));
579 ari = (firstterm + secondterm - 21.43);
580
581 return ari;
582 }
583 // Get the Coleman-Liau Readability Index for a text
584 // CLI = 5.89(total characters/total words) - 29.5((total sentences)/total words) - 15.8
585 public static double getCLI(int total){
586 double cli ;
587 double firstterm = (5.89 * getAverageWordLength(total));
588 double secondterm = (29.5 * (getNumberOfSentences() / total ));
589 cli = (firstterm - secondterm - 15.8);
590
591 return cli ;
592 }
593
594 // This method generates a line in the ARFF
595 // file which corresponds to a document in the directory
596
597 public static String getARFFLine(String value,int unique, int total, int lemmas){
598
599 String comma = ",";
600 double lem = lemmas;
601 double grammlexratio = (getFreqWordType(gramm) / getFreqWordType(lex));
602
603 double infoload = getInformationLoad(total);
604 double avgsent = getAverageSentenceLength(total);
605 double nounratio = getFreqWordType(nouns) / total;
606 double fverbratio = getFreqWordType(fverbs) /total;
607 double pnounratio = getFreqWordType(pronouns) /total;
608 double prepratio = getFreqWordType(preps) /total;
609 double conjratio = getFreqWordType(conj) /total;
610 double numratio = getFreqWordType(numerals) /total;
611 double typetoken = getTypeTokenRatio(unique);
612 double avgwordlength = getAverageWordLength(unique);
613 double cli = getCLI(total);
614 double ari = getARI(total);
615 double lexrichness = lem / total ;
616 double simplecomplex = getSimpleComplexRatio();
617 double dmark = dmarkers / total ;
618 double complextotal = getComplexTotalRatio();
619 double simpletotal = getSimpleTotalRatio();
620 return (grammlexratio + comma + infoload + comma + avgsent +
621 comma + nounratio + comma + fverbratio +
622 comma + pnounratio + comma + conjratio + comma + prepratio +
623 comma + numratio + comma + typetoken +
624 comma + avgwordlength + comma + ari + comma + cli + comma +
625 lexrichness + comma + simplecomplex + comma + dmark + comma
626 + complextotal + comma + simpletotal + comma + value);
627
628 }
629 // Divide the number of grammatical words by the number of lexical words
630 // Larger ratio = less lexical words
```

```
631 public static double getGrammLexRatio(){
632
633 double grammlex = 0.0;
634 double gramms = 0.0;
635 double lexes = 0.0;
636
637 lexes = getFreqWordType(lex);
638 gramms = getFreqWordType(gramm);
639
640 grammlex = gramms / lexes ;
641
642 return grammlex ;
643
644 }
645
646
647 }
```

Auxiliary classes are required, the following code describes a matched POS-word pair:

```
1  import java.io.*;
2  import java.util.*;
3
4  public class POSPair{
5
6     public String pos ;
7     public String token ;
8
9   public POSPair(String p, String t){
10
11  pos = p;
12  token = t;
13  }
14  public void setPOS(String p){
15  pos = p;
16  }
17  public void setToken(String t){
18  token = t;
19  }
20  public String getPOS(){
21  return pos;
22  }
23  public String getToken(){
24  return token;
25  }
26  public String toString(){
27  return pos + " " + token ;
28  }
29
30 }
```