NATIVE LANGUAGE IDENTIFICATION INCORPORATING SYNTACTIC KNOWLEDGE

Sze-Meng Jojo Wong
Master of Information Technology, Monash University
Bachelor of Computer Science, Monash University

This dissertation is presented for the degree of

Doctor of Philosophy

at

October 2013

Contents

List of Tables

List of Figures

Abstract

Inferring characteristics of authors from their textual data, often termed authorship profiling, is typically treated as a classification task, where an author is classified with respect to characteristics including gender, age, native language, and so on. This profile information is often of interest to marketing organisations for product promotional reasons as well as governments for crime investigation purposes.

The thesis focuses on the specific task of inferring the native language of an author based on texts written in a second language, typically English; this is referred as *native language identification* (NLI). Since the seminal work of Koppel et al. in 2005, this task has been primarily tackled as a text classification task using supervised machine learning techniques. Lexical features, such as function words, character n-grams, and part-of-speech (PoS) n-grams, have been proven to be useful in NLI. Syntactic features, on the other hand, in particular those that capture grammatical errors, which might potentially be useful for this task, have received little attention. The thesis explores the relevance of concepts from the field of second language acquisition, with a focus on those which postulate that constructions of the native language lead to some form of characteristic errors or patterns in a second language.

In the first part of the thesis, an experimental study is conducted to determine the native language of seven different groups of authors in a specially constructed corpus of non-native English learners (International Corpus of Learner English). Three commonly observed syntactic errors that might be attributed to the transfer effects from the native language are examined — namely, subject-verb disagreement, noun-number disagreement, and misuse of determiners. Based on the results of a statistical analysis, it is demonstrated that these features generally have some predictive power, but that they do not improve the level of best performance of the supervised classification, in comparison with a baseline using lexical features.

In the second part, a second experimental study aims to learn syntax-based errors from syntactic parsing, with the purpose of uncovering more useful error patterns in the form of parse structures which might characterise language-specific ungrammaticality. The study demonstrates that parse structures, represented by context-free grammar (CFG) production rules and parse reranking features, are useful in general sentence grammaticality judgement. Consequently, adapting these syntactic features to NLI, with the use of parse production rules in particular, a statistically significant improvement over the lexical features is observed in the overall classification performance.

The final part of the thesis takes a Bayesian approach to NLI through topic modeling in two ways. Topic modeling, using a probabilistic CFG formulation, is first taken as a feature clustering technique to discover coherent latent factors (known as 'topics') that might capture predictive features for individual native languages. The topics, rather than the word n-grams that are typical of topic modeling, consist of bi-grams over part of speech. While there is some evidence of topic cluster coherence, this does not improve the classification performance. The second approach explores adaptor grammars, a hierarchical non-parametric extension of probabilistic CFGs (and also interpretable as an extension of topic modeling), for feature selection of useful collocations. Adaptor grammars are extended to identify n-gram collocations of arbitrary length over mixtures of PoS and function words, using both maxent and induced syntactic language model approaches to NLI classification. It is demonstrated that the learned collocations used as features can also improve over the baseline (lexical) performance, although success varies with the approach taken.

Declaration

The research presented in this thesis is the original work of the author except where otherwise indicated. This work has not been submitted for a degree or any other qualification to any other university or institution. All verbatim extracts have been distinguished by quotations, and all sources of information have been specifically acknowledged.

Some parts of this thesis include the revised versions of the published papers:

- Sze-Meng Jojo Wong and Mark Dras. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61, Sydney, Australia, December 2009
- Sze-Meng Jojo Wong and Mark Dras. Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1600–1610, Edinburgh, Scotland, July 2011
- Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. Topic modeling for native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 115–124, Canberra, Australia, December 2011
- Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. Exploring adaptor grammars for native language identification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 699–709, Jeju Island, Korea, July 2012

Additional papers co-authored by the author of this thesis that are referred to but not incorporated into the text of this thesis:

- Sze-Meng Jojo Wong and Mark Dras. Parser features for sentence grammaticality classification. In *Proceedings of the Australasian Language Technology Association Workshop 2010*, pages 67–75, Melbourne, Australia, December 2010
- Shervin Malmasi, Sze-Meng Jojo Wong, and Mark Dras. NLI Shared Task 2013: MQ Submission. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133, Atlanta, Georgia, June 2013

Signed: Sze-Meng Jojo Wong

Date:

Acknowledgements

First of all, I would like to express my deepest and abiding appreciation to my principal supervisor, Mark Dras, for his invaluable supervision and his unyielding support both intellectually and emotionally throughout my PhD candidature. I am grateful and indebted to Mark, who was willing to guide through a person, without any prior knowledge in linguistics, to produce a natural language processing thesis. I would also like to thank Josef Pieprzyk for being my co-supervisor and the financial support for this thesis. I would like to thank all the three reviewers as well, Ted Briscoe, Moshe Koppel and Timothy Baldwin, for their insightful and constructive comments that had transformed this thesis into a much polished piece of work.

This thesis was shaped in many ways through intellectual interaction with many people that I had meet during my PhD candidature. I would like to express a special thanks to Mark Johnson for his guidance on machine learning related topics, topic modeling in particular, that formed the final part of this thesis. Much gratitude is due to Dominique Estival, who allowed me to sit in her classes to have my first formal lecture in syntax. I would also like to extend my honest appreciation to Jennifer Foster, for her invaluable advice and the hospitality during my visiting at Dublin City University. I thank Joachim Wagner and Josef van Genabith as well for taking time off for some great discussion. A sincere thanks to Joel Teatrault, who I met in Edinburgh and exchanged much knowledge with.

Not forgetting to mention that, much intellectual and emotional support was received from both the current and past members of the Centre of Language Technology at Macquarie University. I extend my heartfelt thanks to, not in a particular order, Teresa Lynn, Yasaman Motazedi, Jette Viethen, Mary Gardiner, Susan Howlett, Marta Vila, Benjamin Börschinger, Simon Zwarts, François Lareau, Diego Mollá, Rolf Schwitter, Robert Dale, Matthew Honnibal, Bevan Jones, Shervin Malmasi, Andrew Lampert, Stephen Wan, Abeed Sarker, Pawel Mazur, Chris Rauchle, Steve Cassidy, Marc Tilbrook, Ben Hachey, Mehdi Parviz, Jason Naradowsky, Tony Zhao, Lan Du, Mac Kim, Ilya Anisimoff, Nataliya Sokolovska, Anders Johannsen, and Luiz Pizzato. I appreciated all the great moments that we shared at lunch, dinner, reading groups, meetings, and conferences.

My endless gratitude goes to my husband, Victor Choi, for his love, patience, and support throughout this long journey. I thank him for being understanding and not making much complaints that I had to work through days and nights. And without his loving care and wonderful meals, I would have not survived for the last couple of months towards the completion of this thesis.

Finally, I dedicate this thesis to my lovely family members: to the memory of my father, to whom that I owe so much that I cannot express in words; to the memory of my grandmother, who always wanted to see me with a graduation gown; to my beloved mother, who allowed me to be far away from home to pursue my dream; and my two dearest sisters, Jamie and Jacqueline, who agreed to take good care of both grandmother and mother while I was away. And a special thanks must go to my cousin, Yin Mei, and my cousin-in-law, Long Vu, who supported me with a warm shelter and a plenty of good food throughout my PhD journey.

Chapter 1

Introduction

Back in the 19th century, the work of Mendenhall (1887) marked the beginning of the field of *stylometry* (in which statistical approaches are used for *stylistic analysis*), attempting to discover the true author of Shakespearean works through frequency distribution of varying word lengths. Following the authorship study on the twelve disputed *Federalist Papers* by Mosteller and Wallace (1964), a substantive number of pieces of research have emerged to explore various statistical and computational approaches to stylistic analysis, also commonly known as *authorship analysis* (Stamatatos, 2009; Koppel et al., 2009). A particular subfield that emerged from this broader research area is the attempt to profile authors with respect to their characteristics or personal traits (such as gender, age, native language, and so forth), typically under circumstances where no known candidates are available. Profile information as such is often of interest to, for example, marketing organisations for product promotional reasons as well as governments for crime investigation purposes.

Stylistic behaviours and characteristics of authors can be gauged from their compositions (both texts and speeches). To give some insight, we present a few notable examples. Revisiting the classic disputed *Federalist Papers*, it was discovered that words such as *enough*, *while*, and *upon* appear to be the unique markers for Alexander Hamilton, while the word *whilst* seems to be exclusively used by James Madison (Mosteller and Wallace, 1964), as seen in Table 1.1. These are evident in the two excerpts (Figures 1.1 and 1.2)[1] illustrating how often *upon* and *whilst* occur in the corresponding papers produced by Hamilton and Madison, respectively.

Author gender, for instance, is predictable through a selection of keywords as investigated in Argamon et al. (2003), in which the extent of use of pronouns and certain types of noun modifiers (such as determiners and quantifiers) have been shown to differ between male- and female-

[1] The *Federalist Papers* are available at `http://thomas.loc.gov/home/histdox/fedpapers.html`.

Table 1.1: Occurrence rates per 1,000 words in the Federalist Papers (Mosteller and Wallace, 1964)

	enough	*while*	*whilst*	*upon*	**Total words in 1,000's**
Hamilton	0.59	0.26	0	2.93	45.7
Madison	0	0	0.47	0.16	51.0
Disputed	0	0	0.34	0.08	23.9
Joint	0.18	0	0.36	0.36	5.5

Suppose the difficulties of agreeing **upon** a rule surmounted, and the apportionment made. Still there is great room to suppose that the rule agreed **upon** would, **upon** experiment, be found to bear harder **upon** some States than **upon** others. Those which were sufferers by it would naturally seek for a mitigation of the burden. The others would as naturally be disinclined to a revision, which was likely to end in an increase of their own incumbrances.

Figure 1.1: An excerpt from the *Federalist Paper No.7* which was known to be written by Hamilton

Whilst the objection itself is levelled against a pretended oligarchy, the principle of it strikes at the very root of republican government. The aim of every political constitution is, or ought to be, first to obtain for rulers men who possess most wisdom to discern, and most virtue to pursue, the common good of the society; and in the next place, to take the most effectual precautions for keeping them virtuous **whilst** they continue to hold their public trust.

Figure 1.2: An excerpt from the disputed *Federalist Paper No.57* which was shown to be written by Madison

Table 1.2: Scores of feminine and masculine keywords by *The Gender Genie*

Austen's				**Milton's**			
Feminine		**Masculine**		**Feminine**		**Masculine**	
with	208	around	0	with	0	around	0
if	94	what	0	if	47	what	70
not	81	more	0	not	135	more	0
where	0	are	56	where	0	are	84
be	102	as	46	be	102	as	299
when	34	who	0	when	0	who	38
your	0	below	0	your	0	below	0
her	135	is	8	her	0	is	24
we	16	these	0	we	32	these	0
should	0	the	168	should	14	the	126
she	78	a	72	she	0	a	114
and	60	at	30	and	60	at	6
me	0	it	24	me	0	it	18
myself	0	many	0	myself	0	many	6
hers	0	said	0	hers	0	said	0
was	7	above	0	was	2	above	0
		to	28			to	18
	815		432		392		**803**

authored texts. Using an online gender prediction system, *The Gender Genie*[2], with a short excerpt from the two classical texts — Jane Austen's *Northanger Abber* (Figure 1.3) and John Milton's *Aeropagitica* (Figure 1.4)[3] — it is evident that Austen and Milton used certain keywords that reflect their gender. Table 1.2 presents the scores assigned to the feminine and masculine keywords used by each author; it is apparent that Austen has a higher score ratio of feminine over masculine while the opposite holds for Milton.

The type of authorship analysis that most closely relates to the studies set out in this thesis is to uncover the native background of an author. Considering the following two sentences:

(1.1) a. This is a very difficult *to solve* problem. (Swan and Smith, 2001)

b. This is a very difficult problem *to solve*.

It is obvious that the first sentence as opposed to the second is not native-like and is ungrammatical with respect to standard English grammar. The noun modifier in 1.1a appears at the position before the head noun which violates the language system of English which typically uses post-modifiers. The ungrammatical feature indicates that the first sentence is most likely to be produced

[2]The Gender Genie was implemented using the algorithm developed by Argamon et al. (2003) and is available at `http://bookblog.net/gender/genie.php`.

[3]Both Milton's and Austen's texts were obtained from `http://www.teachit.co.uk/armoore/lang/gender.htm#spotting`.

Figure 1.3: An excerpt from Jane Austen's Northanger Abber (1818) analysed by *The Gender Genie*

by one whose native language uses pre-modifiers; in this regard, Chinese and Japanese are possible candidates.

Of many of the author characteristics that are of interest, the task of inferring authors' native language has been one gaining much attention in recent years. *Native language identification* (NLI) is the task of determining the native language (often referred to as the first language) of an author (or a writer) based on texts written in a second language, most commonly English.[4] The task is to discover different features through comparisons of texts written by authors of different native language backgrounds. Since the seminal work of Koppel et al. (2005), the approaches taken to address this have been primarily as a text classification task using various supervised machine learning techniques. *Lexical features*, such as function words, character n-grams, and part-of-speech (PoS) n-grams, have been proven to be useful in the NLI paradigm by Koppel et al. (2005) and other subsequent works that followed (Estival et al., 2007; Tsur and Rappoport, 2007). *Syntactic features*, on the other hand, in particular those that capture grammatical errors, which

[4]We note that one's native language can be inferred from spoken utterances in a second language (in the form of speech transcripts) as demonstrated in an early work by Tomokiyo and Jones (2001). However, only written texts in a second language — English to be precise — is the central focus of this thesis.

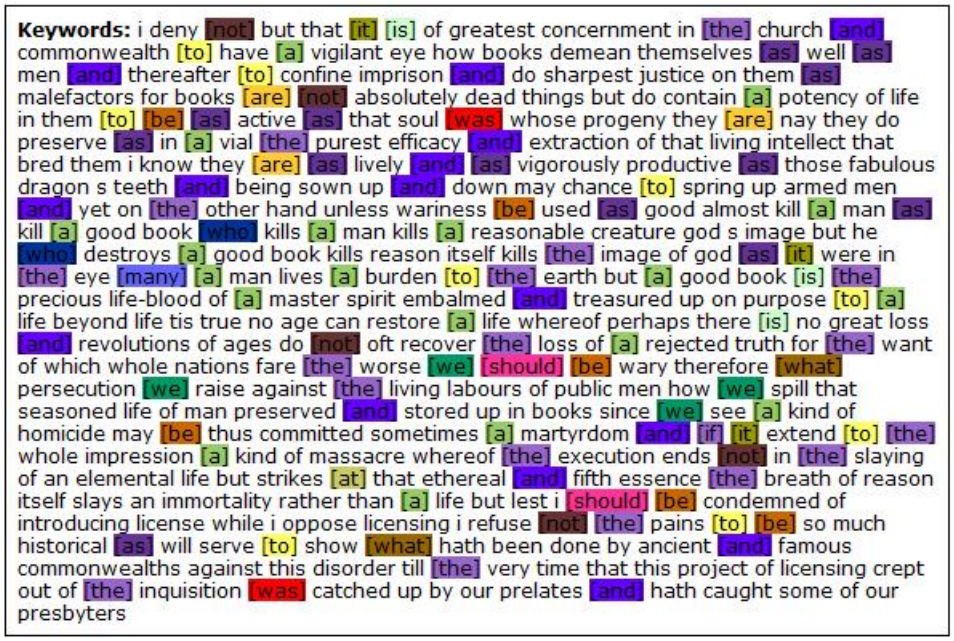


Figure 1.4: An excerpt from John Milton's Aeropagitica (1644) analysed by *The Gender Genie*

might potentially be useful for this task, have not received similar degree of attention; Koppel et al. (2005) suggested that they could potentially be useful but left their exploration for future work.

The underlying idea that syntactic features — and in particular syntactic errors — might capture characteristics of a particular native language is drawn from the fundamental assumption of one of the early theoretical frameworks of second language acquisition (SLA) put forward by Lado (1957) more than half a century ago. Taking the weaker form of Lado's *Contrastive Analysis Hypothesis* (CAH), it is postulated that constructions of the native language, to some extent, may be reflected in some form of characteristic errors or patterns in second language constructions. The influence of the native language on second language productions can also manifest in other ways, such as overuse (overproduction) or avoidance (underproduction) of particular constructions in the second language (Odlin, 1989; Ellis, 2008). Non-native speakers of a particular second language tend to resort to more simple sentence constructs (and hence overuse) to avoid using those that they perceive as complicated and less confident with. It is these sorts of syntactic characteristics that this thesis intends to explore, and the extent to which they serve as helpful clues for identifying the native language of the writers based on their written compositions in a second language.

General problem statement. The broad intention of this thesis is to investigate whether *syntax*, in the form of syntactic errors and other manifestations, as inspired by second language acquisition (SLA) theories of cross-linguistic influence, is at all useful in the task of native language

identification (NLI).

1.1 Use Cases

As well as the innate interest of investigating the nature of cross-linguistic influence, there are potential applications. We briefly discuss two of them.

Second language learning. Within the paradigm of second language acquisition (SLA) studies, understanding the native language of second or foreign language learners is undoubtedly crucial in identifying the challenges faced by learners of diverse native language backgrounds (Ellis, 2008; Gass and Selinker, 2008). With better knowledge about which of the errors are indeed the effects of a specific native language and which are consequences of the more general process of learning a language, more informed pedagogy that better suits the learners can be applied accordingly.

Since the introduction of computer-assisted language learning (CALL) (Levy, 1997), there has been a growing interest in automated error detection and correction research aiming to assist second or foreign language teaching and learning. An abundance of studies exploring various approaches, ranging from rule-based to statistical approaches, to detecting as well as correcting learner errors have constantly emerged at least for the last two decades. Leacock et al. (2010) provides an extensive survey on many of the related works, in particular those that target aspects that are amongst the most challenging of the English language system for non-native learners — namely, articles, prepositions, collocations, and orthography (spelling). In relation to this, there has been a recurring workshop, Building Educational Applications (BEA) (Tetreault et al., 2012) and a recent shared task, Helping Our Own (HOO) (Dale et al., 2012) devoted to this field of research. Development of end-to-end CALL systems, such as *Criterion*[5] by Educational Testing Service (Burstein et al., 2004) (Figure 1.5) and *ESL Assistant* by Microsoft Research (Gamon et al., 2008), as well as writing assistance systems (Hermet and Désilets, 2009; Chen et al., 2012) has also been underway.

Many of the previous studies have not been widely taking the learners' native language into account for the tasks of detecting and correcting the learner errors. Only in recent years, the knowledge of the native language has started gaining in importance and being incorporated into detection and correction mechanisms (Hermet and Désilets, 2009; Tetreault and Chodorow, 2009; Gamon, 2010; Han et al., 2010; Dahlmeier and Ng, 2011; Rozovskaya and Roth, 2011). In this regard, our work presented in this thesis can be complementary. Through native language identification informed by cross-linguistic effects, different types of ungrammatical error and unidiomatic

[5]The demonstration of *Criterion* is available at `http://www.ets.org/criterion/higher_ed/demo/`.

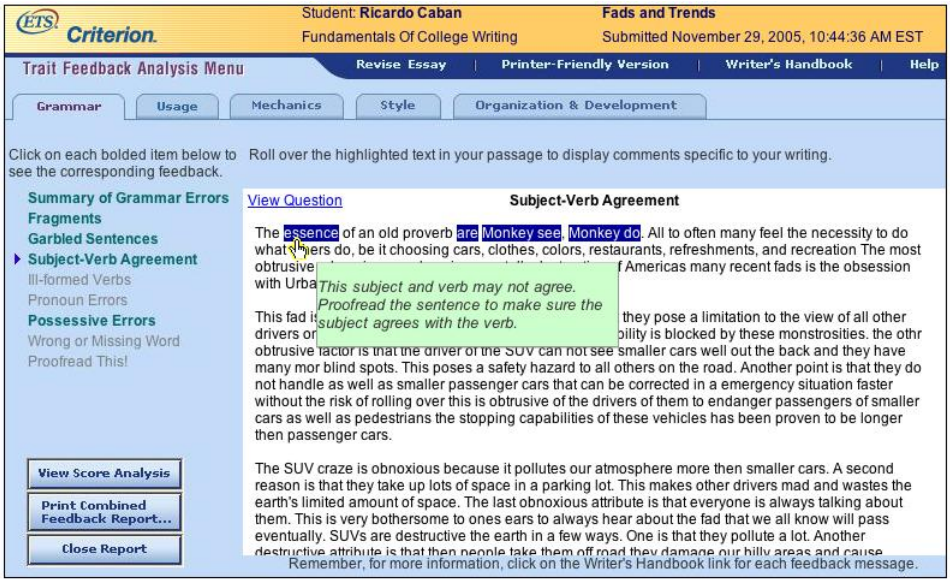


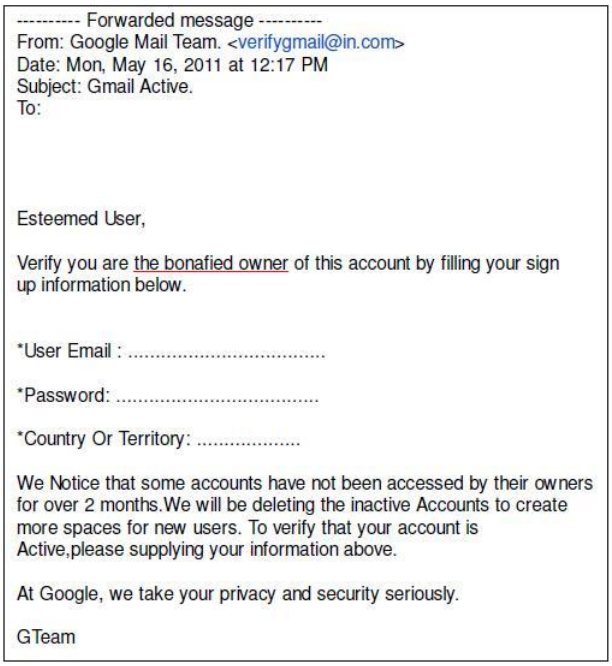
Figure 1.5: A screen shot of the interface of *Criterion* by Educational Testing Service

construction in a second language produced by learners of diverse native language backgrounds can be gauged with a fair degree of confidence. It is likely that learners whose first langauge is similar to the second language (e.g. French versus English) produce certain sorts of errors that are different from those whose the first language is unrelated to the second language (e.g. Chinese versus English). As for instance, native English speakers normally refer to ***take*** *medicine* in contrast to Chinese learners of English who tend to say ***eat*** *medicine*. Likewise, Japanese speakers would naturally use ***drink*** *medicine* with the influence of their cultural background. These expressions are regarded as unidiomatic given that they do not exhibit native-like constructions, although they are syntactically acceptable.[6]

Phishing profiling. A less direct possible application is that of phishing profiling. A phishing medium (such as e-mail or a website) is designed to deceive victims (primarily Internet users) into giving away personal confidential details through social engineering techniques (Myers, 2007). An example of a phishing e-mail is given in Figure 1.6.

There have been some research efforts focusing on authorship profiling in the forensic context specifically by utilising various forms of stylometric features, ranging from lexical to structural and content-specific, serving as individual stylistic discriminators (de Vel et al., 2001; Zheng et al., 2003; Fette et al., 2007; Abbasi and Chen, 2008). These research efforts have demonstrated the

[6]From the perspective of SLA, idiomaticity is in connection with the ability to express fluently and accurately at the same time Robinson (2013).



---------- Forwarded message ----------
From: Google Mail Team. <verifygmail@in.com>
Date: Mon, May 16, 2011 at 12:17 PM
Subject: Gmail Active.
To:

Esteemed User,

Verify you are the bonafied owner of this account by filling your sign up information below.

*User Email : .....

*Password: .....

*Country Or Territory: .....

We Notice that some accounts have not been accessed by their owners for over 2 months.We will be deleting the inactive Accounts to create more spaces for new users. To verify that your account is Active,please supplying your information above.

At Google, we take your privacy and security seriously.

GTeam

Figure 1.6: A example of a phishing e-mail

potential of authorship analysis in identifying criminal suspects where convincing results are not implausible, in which high level of accuracies up to 95% are achievable under the supervised learning setting with the availability of sufficient amount of training data and potential suspects (de Vel et al., 2001; Abbasi and Chen, 2008). However, under the worst case scenario when neither training data nor potential suspects are available such as the case of phishing attacks, unsupervised learning is more appropriate. There have been a number of recent studies which attempt to profile phishing texts using various types of clustering methods (Ma et al., 2009; Layton and Watters, 2009; Layton et al., 2010; Iqbal et al., 2010).

The sorts of syntactic characteristics particular to a native language, as investigated in this thesis, could potentially contribute useful features to this task. Although phishing e-mails generally aim to be formal, they are often flawed with grammatical and orthographic errors as well as unidiomatic (non-native like) expressions and *malapropisms* (such as the phrase of *the bonafied owner* in the given figure) (Schaffer, 2012). By only identifying the native language, this might not suffice in pinning down the real attackers (i.e. the phishers). Nonetheless, such information could potentially be useful to law enforcement agencies and could be used in conjunction with other clues to assist investigation by narrowing down potential suspects.

1.2 The Sub-Problems and Approaches

As stated above, the broad intention of this thesis is to investigate whether *syntax*, in the form of syntactic errors and other manifestations, as inspired by second language acquisition (SLA) theories of cross-linguistic influence, is at all useful in the task of native language identification (NLI). The overall framework for identifying the writers' native language rests on the typical *supervised machine learning* techniques to text classification. We specifically investigate the classification of written texts with respect to the writers' native language through exploiting different syntactic patterns manifested in second language productions, which include grammatical errors as well as overuse and avoidance of certain syntactic constructions.

The first of the three approaches taken is an exploratory study to investigate whether syntactic features — *syntactic errors* to be precise — are at all potentially useful for NLI. Three commonly observed syntactic errors that may be attributed to the transfer effects from the native language are investigated — namely, subject-verb disagreement, noun-number disagreement, and misuse of determiners. A descriptive statistical analysis is performed to gauge the extent to which the number of errors manifested by writers of different native language backgrounds differ. Subsequently, the distributions of the three selected errors are used as features in a supervised machine learning-based classification to identify the writers' native language. The contributions of these syntactic features are then compared with those lexical features (i.e. function words, character n-grams, and PoS n-grams) that have been proven useful in previous work (Koppel et al., 2005; Tsur and Rappoport, 2007; Estival et al., 2007). As a preliminary study, this approach is limited to only three possible types of syntactic error that are manually identified.

To further explore the usefulness of syntactic features, the second approach taken uses *statistical parsers*, with the purpose of uncovering more useful error patterns in the form of parse structures and distributions over parse structures which might characterise language-specific ungrammaticality. Parse structures, represented by context-free grammar (CFG) production rules and features used for parse reranking (Charniak and Johnson, 2005), are first tested on a more general task for judging the grammatical status of sentences to see whether they can detect grammatical errors at all. Subsequently, these syntactic features are adapted to the task of NLI. As compared to the first approach addressing only three syntactic error types, such an approach may capture other syntactic productions in addition to errors — overproduction and underproduction of certain syntactic constructions, in particular.

The third and last approach to be investigated in this thesis explores unsupervised learning. It takes a Bayesian approach to NLI through *topic modeling* in two ways. Adopting the probabilistic context-free grammar (PCFG) formulation of Johnson (2010), topic modeling is first taken as

a *feature clustering* technique to discover coherent latent factors (often known as 'topics' under a typical topic modeling setting) that might capture predictive features for individual native languages by grouping distributions of parse structure components into possibly interpretable topics. The second method explores an extension of topic modeling — *adaptor grammars* (a hierarchical non-parametric extension of PCFGs) — for *feature selection* of characteristic collocations. For the purpose of NLI, adaptor grammars are extended in a number of ways to discover n-gram collocations of arbitrary length over the purely PoS type as well as mixtures of PoS and function words. In a sense, these are approximations to the syntactic structures used in the previous approach in the thesis. The NLI classification task with n-gram collocations as features is then tackled through conventional machine learning-based classification techniques and also syntactic languange modeling ideas.

1.3 The Corpus

All the native language identification tasks in this thesis are performed on written texts extracted from the *International Corpus of Learner English* (ICLE) compiled by Granger et al. (2009), the purpose of which is to examine the English writings of non-native English learners of diverse backgrounds. All the contributors to the corpus were university students in their third- and fourth-year of undergraduate studies and were claimed to possess similar levels of English proficiency, ranging from intermediate to advanced. It is worth mentioning that English to the contributors is a foreign rather than a second language given that English is learned in a non-English speaking country and primarily in a classroom setting.

The first initial version of the corpus consists of 11 sub-corpora of English essays contributed by the university students from various countries whose native languages are mostly from the European and Slavic groups — namely, Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish and Swedish. The second extended version includes five other native languages, including two Asian languages — Chinese and Japanese in addition to Norwegian, Turkish and Tswana. In this extended version, there are 6,085 essays with a total number of 3.7 million words and each essay has an average length of 617 words. The majority of the essays are argumentative (up to 91%) and the remaining is literary.

Taking Koppel et al. (2005) as a point of reference, the native languages to be investigated throughout this thesis include the five Indo-European languages as examined by them — Bulgarian, Czech, Russian, French, and Spanish, along with the only two Asian languages from the newer version of the corpus (the second version) — Chinese and Japanese. The latter two languages are of particular interest given that they are of different language family groups as compared to the

others, and they also appear to be somewhat understudied in the context of native language identification.

1.4 Outline of the Thesis

The remainder of the thesis is organised into five chapters as follows.

- **Chapter 2:** This chapter introduces the key concepts and reviews the related work of the main topics of interest in this thesis: firstly on *authorship analysis*; and followed by the central focus of the thesis, *native language identification*. The last section of this chapter pays its attention to the concepts of *second language acquisition* that attempt to account for cross-linguistic effects in particular.

- **Chapter 3:** The first of the three approaches investigated in this thesis is presented in this chapter, after replicating the work of Koppel et al. (2005) for use as a baseline. This exploratory chapter first examines the potential usefulness of three manually identified syntactic errors through statistical analysis and subsequently use them as features for the classification task of NLI.

- **Chapter 4:** This chapter covers the approach of exploiting *parse structures* as classification features through *syntactic parsing*. This chapter is structured into two sections. In the first part, it presents the more general task of examining the degree of usefulness of parse structures for sentence grammaticality judgement. The second part incorporates parse structures into the NLI task.

- **Chapter 5:** This chapter discusses the adaptation of *topic modeling* into NLI classification: first as a form of feature dimensionality reduction technique (clustering) and then as a form of feature selection to discover n-gram collocations of arbitrary length. The latter idea is realised through the extension of *adaptor grammars*.

- **Chapter 6:** This concluding chapter brings together the results of the three main components of the thesis, and reflects on various limitations and possible extensions.

Chapter 2

Literature Review

In this chapter, we will first discuss the broader problem of authorship analysis and its subfields which include the widely studied tasks of authorship attribution and authorship profiling. We will then dive into the main focus of the thesis — native language identification (NLI) — which can be seen as a specific subtask of authorship profiling. Finally, we will review both some foundational and more recent work in second language acquisition (SLA), which this thesis draws on in developing the approaches to NLI of the subsequent chapters.

2.1 Authorship Analysis

Authorship analysis, in a broader sense, is a process of investigating the unique characteristics of a created work in order to derive conclusions on its authorship. It is rooted in a field of linguistic research — *stylometry* — in which statistical methods are used for literary style analysis (Holmes, 1998). In a more specific sense, the study of authorship analysis can be categorised into three major subfields according to the past literature: authorship identification, authorship characterisation, and similarity detection (Gray et al., 1997; Zheng et al., 2006; Stamatatos, 2009).

Authorship identification, also known as *authorship attribution* within the linguistic community, determines the author of a given anonymous work (such as a written text), by comparing other work produced by the same author. A predefined set of candidate authors is usually assumed. This subfield of authorship analysis originated from the attribution work of Mendenhall dating back to the mid 19th century, which studied the authorship of Shakespearean works attributed to Bacon, Marlowe, and Shakespeare based on the hypothesis of the British logician, Augustus de Morgan, suggesting that authorship problems can be settled through comparison of authors' average word length in composition (Mendenhall, 1887). A more recent prominent example of authorship iden-

tification is that of the analysis of the twelve disputed *Federalist Papers* claimed variously to be written by Madison and Hamilton (Mosteller and Wallace, 1964). Through several statistical methods including Bayesian inference (essentially the Naïve Bayes classification at present), Mosteller and Wallace (1964) concluded that the most likely author of the disputed papers was Madison, which agreed with the findings of historians. Since then, this seminal work has served as the benchmark for all the subsequent works investigating the *Federalist Papers*. While these historically important documents became a common test bed for authorship analysis in the early years, various other domains have been explored ranging from classic literature (Ledger and Merriam, 1994; Gamon, 2004) and biblical studies (García and Martín, 2007; Koppel et al., 2011a) to scientifiic works (Bergsma et al., 2012; Feng et al., 2012) and social media (Schler et al., 2006; Rao and Yarowsky, 2010). On the whole, authorship identification remains the most active research area compared to the other subfields of authorship analysis, as presented in a comprehensive summary compiled by Zheng et al. (2006).

In contrast to authorship identification, *authorship characterisation* is when there is no known candidate set available. This challenging subfield of authorship analysis exploits the sociolinguistic observation that predicts the characteristics or traits of different classes of authors in order to generate author profiles, which can then be used to predict an author's identity. It is therefore also called *authorship profiling*. Estival et al. (2007) categorise author traits (attributes) as either demographic or psychometric. Demographic traits include gender, age, country of origin, level of education, as well as native language; psychometric traits for Estival et al. (2007) cover characteristics such as the commonly known 'Big Five' personality traits: openness, conscientiousness, extraversion, agreeableness, as well as neuroticism. Most research into authorship profiling mainly focuses on the prediction of a small subset of these traits with gender and age being the two widely studied ones (de Vel et al., 2001; Koppel et al., 2002; Burger and Henderson, 2006; Schler et al., 2006); the study by Estival et al. (2007) predicted the largest number of traits: up to ten in total. Another similar piece of research by Argamon et al. (2006), looked at fewer traits (these include age, gender, and native language, as well as neuroticism); their work identified the most discriminating features for each predicted trait based on the measure of information gain. In recent years, there has been growing interest in other author attributes: for instance, to investigate whether an author is a native or non-native speaker of English (Garera and Yarowsky, 2009; Bergsma et al., 2012) and also the ethnicity of the author (Eisenstein et al., 2011; Rao et al., 2011).

A third subfield of authorship analysis — *similarity detection* — compares multiple pieces of anonymous work and evaluates the degree of similarity between them. This task of authorship analysis does not necessarily conclude by identifying the author(s) in question. An instance of this has been investigated by Koppel et al. (2011b) where they found that similarity-based methods

are helpful for situations in which the set of known candidates is extremely large and also when the actual author might not be covered. Similarity detection has been used to assist humans in plagiarism detection which attempts to detect whether a piece of work is a replica of an original piece (Gray et al., 1997). The issue of plagiarism has been extensively studied such as the works of Clough (2003) and zu Eissen and Stein (2006) and many approaches for automatic plagiarism detection are being explored. Compared with the other two subfields of authorship analysis, similarity detection is regarded as the least studied area, argue Abbasi and Chen (2008).

Below, we first look at the features used in authorship analysis, and followed by the classification techniques that use these features.

2.1.1 Linguistic Features for Authorship Analysis

A typical approach to authorship analysis essentially involves defining a relevant set of stylistic features (so-called *style markers*). In the view of the survey study for authorship analysis conducted by Stamatatos (2009), feature selection is the essence of authorship analysis in which unique stylistic features serve as effective discriminators for authorship. Most of the previous studies in the field of authorship analysis assume, implicitly or explicitly, that the writing style for a given author is distinguishable by a set of stylistic features that usually remain constant across a large number of writings produced by the same author. Based on a review of literature, stylistic features can generally be classified as lexical, syntactic, semantic, structural, content-specific, and idiosyncratic (Zheng et al., 2006; Abbasi and Chen, 2008; Stamatatos, 2009). The following subsections detail each of these features.

Lexical features. In many of the studies on authorship analysis, *lexical features* are commonly adopted. Lexical features were introduced to complement primitive measures used in the early work on authorship for text complexity analysis (Stamatatos et al., 2001; Koppel et al., 2009). The most common primitive measures are average word length (in terms of letters or syllabus), and average number of words per sentence (or sentence length). In view of the unreliability of primitive measures, lexical-based measures involving type-token ratio and number of words occurring with a specific frequency were introduced; this is termed *vocabulary richness*. Some formal statistical measures for vocabulary richness which have been proposed by researchers in the field of quantitative linguistics include Yule's K-measure, Simpson's D-measure, Sichel's S-measure, Brunet's W-measure, and Honoré's R-measure (Tweedie and Baayen, 1998). For example, K-measure is defined by Yule (1944) as below:

$$K = 10^4(-\frac{1}{N} + \sum_{i=1}^{v} V(i,N)(\frac{i}{N})^2) \tag{2.1}$$

with N as the number of tokens, $V(i,N)$ as the number of types which occur i times in a sample of N tokens, and v as the highest frequency of occurrence.

Unfortunately, vocabulary richness measures are highly dependent on content as well as text length. As such they might not be applicable for discriminating authors across various contexts. Most of the evaluation results in previous studies showed that vocabulary richness measures generally perform worse than other lexical features such as function word counts as discussed below (Tweedie and Baayen, 1998; Zheng et al., 2006; Grieve, 2007).

Function words,[1] which are commonly recognised to be context-independent, then came into play in authorship studies. For a (function) word to be context-free and a good stylistic indicator, its occurrence would follow a Poisson distribution, claimed Damerau (1975) (cited in Tweedie et al. (1996)). Function words were proven to be reliable (and also context-free) for the first time when Mosteller and Wallace (1964) successfully attributed the twelve disputed *Federalist Papers* to Madison with the use of frequency probabilities of the most frequent function words (such as prepositions, conjunctions, and articles). The conclusion derived from this seminal work of theirs was found to be in agreement with the historical findings (Coulthard and Johnson, 2007). Many subsequent authorship studies further confirmed the efficacy of function words as a good discriminator (Merriam and Matthews, 1994; Holmes and Forsyth, 1995; Tweedie and Baayen, 1998; de Vel et al., 2001; García and Martín, 2007).

Various sets of function words were selected in these studies; but it is unclear exactly how the selection was conducted. The most frequent function words were used in Holmes and Forsyth (1995) and Tweedie and Baayen (1998); the former extracted 50 function words from the *Federalist Papers* and the latter extracted 100 function words from a work of children's literature — *Alice's Adventures in Wonderland*. Merriam and Matthews (1994) and García and Martín (2007) have employed function words in the form of a ratio to content words in two different contexts: literary texts of Shakespeare and Marlowe for the first study, gospels for the second one. 122 function words were proposed by de Vel et al. (2001) for e-mail authorship attribution. It is uncertain whether function words are credible when dealing with online texts as unconvincing results were observed by Zheng et al. (2006). The inclusion of function words as the part of the feature set for the English dataset used in Zheng et al.'s study demonstrated insignificant improvement on

[1] de Vel et al. (2001) and Zheng et al. (2006) both classified function words as 'syntactic' claiming that such word usage determines how a sentence is formed; but they are referred to as 'lexical' here.

accuracy.

Other lexically-based approaches have adopted distributions over n-grams. Two n-gram-oriented authorship studies were conducted at the byte level (Keselj et al., 2003) and the character level (Peng et al., 2003). With the use of simple features like byte-level and character-level n-grams, non-trivial natural language processing analysis can be avoided since these features seem to be fairly language-independent. Three types of languages were investigated in these two studies, namely English, Greek, and Chinese. Although improvement in performance (ranging between 10-20%) was observed in Keselj et al. (2003) and Peng et al. (2003) particularly for the Greek data set as used in the previous related work (Stamatatos et al., 2000), their approaches deserve further evaluation with a wider range of languages in order to justify whether n-grams are truly language-dependent. Nonetheless, the use of character n-grams was first introduced by Kjell (1994) back in the mid nineties. In his work, only letter pairs of alphabetic characters were used for the authorship analysis task on the Federalist Papers.

Syntactic features. Using lexical-based features results in quite a shallow analysis. For deeper linguistic analysis, one could consider features that can reflect some form of syntactic or semantic information. Part-of-speech (PoS) n-grams (Stamatatos et al., 2001; Diederich et al., 2003; Argamon et al., 2006) and punctuation symbols (de Vel et al., 2001; Baayen et al., 2002), have been applied in authorship research as *syntactic features*; but they do not truly reflect syntactic structure *per se*. A more sophisticated work of authorship research that explicitly investigates the syntactic structure in texts is that of Baayen et al. (1996), which made use of the frequencies of the syntactical rewriting rules for each sentence. Each sentence is represented by a semi-automatically produced parse tree. Syntactical rules derived from the parse tree are rewritten by adding functional meaning to each constituent. For instance, a syntactic rule representing a noun phrase (such as *the article*) comprising two constituents: [NP → DTP + N] is rewritten as [OD:NP → DT:DTP + NPHD:N],[2] where the label to the right of the colon denotes the syntactic category and the label to the left of the colon denotes the corresponding functional meaning. Baayen et al. (1996) highlighted that their syntactic-based approach potentially performs better compared to purely lexical-based features since the use of syntactic rules is generally less variable within texts than the use of words.

Contra to this, Hirst and Feiguina (2007) argued that the approach of Baayen et al. (1996) suffers from a large space of rewriting rules, and it was assumed to be only applicable on large texts. As an attempt at improvement, syntactic-label bigrams were proposed for authorship attribution

[2] In this rewritten syntactical rule, NP (noun phrase), DTP (determiner phrase), and N (noun) are the syntactic labels; whereas OD (direct object), DT (determiner), and NPHD (noun phrase head) are the functional labels. Note that these labels are based on the TOSCA annotation scheme (Oostdijk, 1991).

of short literary texts (as short as 200 words). In the approach of Hirst and Feiguina (2007), an output of the Cass partial parser (Abney, 1996) is transformed into an ordered stream of syntactic labels. For instance, a noun phrase such as *a good article* would produce a stream of labels [NX DT JJ NN] denoting a noun phrase consists of a determiner, an adjective, and a noun. Bigrams are then extracted from this stream of syntactic labels to represent syntactic features for attribution analysis. Such an approach achieved a reasonably convincing result with close to 85% accuracy using syntactic-label bigrams alone for short texts of 200 words.

Nonetheless, syntactic features adopted by Gamon (2004) in another authorship attribution study are much deeper in comparison, where distributional differences in context-free grammar (CFG) productions were explored. With the use of the parser provided by the Microsoft NLP system NLPWin, context-free grammar production rules of each sentence were extracted. An example is the prepositional phrase, represented as [PP → PP DETP NOUN] in NLPWin. For each observed production, the per-document frequency of each production is measured and is used as the feature for classification analysis. Gamon (2004) tested his ideas on literary texts by the three Brontë sisters — Charlotte, Anne, and Emily — with a total of 1441 documents of 20 sentences each. A high accuracy close to 98% is achievable when other deep linguistic features such as semantics are taken into account as the style markers.

More recently, sophisticated ideas utilising syntactic information have emerged for stylometric analysis and authorship attribution. Raghavan et al. (2010) posit that the writing style of each author could be considered as a 'distinct' language on its own and proposed learning a probabilistic context-free grammar (PCFG) for each of 23 authors to be examined for an authorship attribution task on five different datasets, comprising of news articles and poetry. Each test document is then attributed with respect to the author whose PCFG produces the highest likelihood score for that document. Instead of relying on single-level CFG production rules as used in Gamon (2004), Bergsma et al. (2012) and Feng et al. (2012) both simultaneously attempted different approaches to capture more characteristic syntactic patterns. Tree fragments returned from Tree Substitution Grammar (TSG) were demonstrated to be helpful by Bergsma et al. (2012) for distinguishing native and non-native English writers by capturing syntactic structures where CFG alone are incapable of. For instance, a tree fragment such as [NP → (DT *that*)], representing a stand-alone noun phrase with a determiner, is indicative for non-native writing; such syntactic pattern cannot be obtained through CFG rules as this is broken down into two separate CFG rules: [NP → DT] and [DT → *that*]. Similarly, Feng et al. (2012) attempted to construct similar syntactic patterns by augmenting CFG rules with the 'grandparent' nodes and the augmented rules are found to be often performing better than simple CFG rules in their authorship attribution tasks. Taking the previous example, an augmented CFG rule would look like [DT^NP → *that*].

Semantic features. *Semantic features*, as compared to both lexical and syntactical features, are greatly understudied. As far as semantics is concerned, Gamon (2004) extracted two kinds of semantic information from the semantic dependency graphs: binary semantic features and semantic modification relations. (There is no information provided as to how semantic dependency graphs are produced; this is probably through the NLPWin system.) Some examples of binary semantic features include number and person features on nouns and pronouns, as well as tense and aspectual features on verbs. There were 80 such semantic features extracted in total. Semantic modification relations describe the semantic relations between a node and all its descendants within a semantic graph. One of the common modification structures according to Gamon is a nominal node with a pronominal possessor, represented as [`Noun Possr Pron`]. Similarly, the per-document frequency of all the modification structures are measured; approximately 9400 such structures were observed. Despite semantic features in isolation not performing as well as other feature types, in combination with syntactic features (CFG rules as above) they resulted in error reduction rates within the range of 10–30% in relation to lexical-based features (function words and PoS tri-grams).

A brief review of the literature has demonstrated the potential of syntactic and semantic features as significant stylistic markers for authorship discrimination. These features are worth further investigation, in particular for syntactic features — analysis of long-distance grammar dependencies between constituents (such as subject-verb agreement and verb-adverb placement) can possibly be explored for authorship analysis.

Structural features. Another potential style marker — *structural features* — is more applicable to authorship analysis of online texts as well as software source code. Text composition and layout (such as sentence structure, paragraph length, use of indentation, use of signature, and choice of font as well as its size) can be highly author-dependent. Individual composition style is evident from these features. de Vel et al. (2001) proposed a set of structural features specifically for e-mail authorship identification, and achieved a 4% higher accuracy over purely linguistic features. In other related work, structural features appear to be a good stylistic discriminator between authors for online texts (e-mails, for example) in the authorship studies by Zheng et al. (2006) and Estival et al. (2007). However, it should be noted that structural features may be constrained by genres depending upon the intentions of the authors. The stylistic degree of freedom can be negligible when composing a restricted text as opposed to an unrestricted one. In some circumstances, structural features could not be gainfully exploited as they might turn out to be artifacts of the study. This is particularly true if restrictions are imposed on the structure of as well as the topic of the writing, in which individual stylistic markers would not be easily recognised.

Content-specific features. Some important thematic *keywords* and *phrases* may emerge to be useful discriminators in certain contexts of authorship analysis although they are inherently context-dependent and are widely assumed to be unreliable. Nonetheless, there were some promising studies in authorship attribution with the use of keywords and phrases in terms of word-based n-grams. For instance, content-specific words were used in the cybercrime context (Zheng et al., 2003) where certain content-specific words were helpful in improving the overall attribution performance. 'Topical' words have been found useful by Bergsma et al. (2012) for stylometric tasks in the domain of scientific writing, distinguishing between top-tier conference papers and workshop-styled papers. They also demonstrated that gender prediction on scientific writers can take advantage of topical words as different genders tend to embark on specific areas of research. Since content-specific features are usually in the form of word-based n-grams, intensive efforts are required to identify word and phrase boundaries in the texts of different languages. Furthermore, word and phrase segmentations *per se* remains as a difficult disambiguation task in Asian languages particularly (such as Chinese and Japanese) as argued by Keselj et al. (2003) and Peng et al. (2003).

Idiosyncrasies. In the conventional way for authorship analysis, human experts like forensic linguists usually seek for *idiosyncratic usage* of a given author that serve as the unique fingerprints (perhaps the 'writeprints') of that author (Koppel et al., 2005; Li et al., 2006). The stylistic idiosyncrasies may encompass *orthography*, *grammar*, as well as *neologisms* (usage of unusual words and multiword expressions — not necessarily collocations). The idiosyncratic usage appears to be useful particularly for authorship profiling in identifying author classes that share a common native language as these features essentially exhibit cultural differences and deliberate choices of the authors (Koppel et al., 2005). In somewhat related work (Garera and Yarowsky, 2009), features drawn from socio-linguistic literature (such as frequency of pronoun, passive, and filler word usage) appear to be helpful in gender prediction for conversational speech transcripts.

As a summary, there is a myriad of features that have been exploited for authorship analysis. It is evident that there is no single best feature set that is plausible for various contexts of authorship discrimination. In order to achieve high accuracy of performance, feature sets are to be incorporated with classification methods of relatively high computational capability.

2.1.2 Authorship Classification Techniques

Ever since the first kind of authorship-related problem was proposed in the late 19th century (Mendenhall, 1887), there has been an enormous number of approaches applied to authorship

analysis. According to the literature, classification techniques for authorship analysis can be categorised either along one dimension as statistics-based and machine learning-based (Zheng et al., 2006; Koppel et al., 2009); or along another dimension as supervised and unsupervised (Abbasi and Chen, 2008). Within the statistics-based approaches, further division is possible, where approaches can be univariate and multivariate. Taking these together, approaches to authorship analysis being discussed here are divided into three classes: univariate statistical approaches, multivariate statistical approaches, and machine learning approaches.

Univariate statistical approaches. In the early days of authorship analysis, most of the studies revolved around the task of identification or attribution. The analytical approaches adopted were primarily *univariate*, in which a single numerical feature of a text is used to discriminate between authors. Typically, authorial styles are distinguished based on the probability distribution of word usage. The early work of Mendenhall (1887) was merely based on characteristic curves depicting relationships between different word lengths and their corresponding frequency of occurrence. Mendenhall's idea of using word-length distribution did not turn out to be a useful discriminator (Holmes, 1998), in part because the visualisation of the significant differences between characteristic curves are somewhat subject to individual interpretations. This gave way to more objective statistical measures, such as Yule's K-measure (Yule, 1944), with the assumption that the occurrence of a particular word is largely by chance and can be modelled by a Poisson distribution.

Around the same time, *cusum statistics* — based on cumulative sum charts representing the deviations of the measured variable — were adopted for authorship analysis (Farringdon, 1996). The idea of cusum analysis was used by Morton and Michaelson (1990) in an authorship test in the early 1990s (Holmes, 1998). The underlying assumption is that every individual possesses a unique set of personal styles — *habits* — which are consistent throughout that individual's utterances; and these habits form statistically identifiable patterns in his or her sentences. This statistical technique was the first to find its way into linguistic forensics and had been used in several high-profile court cases (Coulthard and Johnson, 2007). According to Holmes (1998), several later independent studies found that cusum analysis was unreliable and concluded that it should not be used as a definitive indicator for authorship.

Multivariate statistical approaches. By and large, univariate statistical approaches are of limited capability in handling multi-dimensional features. This demanded a shift to *multivariate* statistical methods, in which discriminant analysis is used on multiple numerical features — referred to as stylistic features here. The essence of these methods is to find the most plausible attribution

by assigning an anonymous text to the (supposedly) known author whose texts are closest to it in relation to some similarity or dissimilarity measures. One of the earliest adoptions of multivariate approach to authorship analysis was the *Bayesian* classification model used by Mosteller and Wallace (1964) during their seminal work on the disputed *Federalist Papers*. Their legacy was to demonstrate that the Bayesian classifier could produce significant discrimination results where frequency counts of function words are employed as the variables for multivariate analysis. As mentioned earlier, the outcome of this study yielded a conclusion that is in agreement with the prevailing views reached by historians (Coulthard and Johnson, 2007).

The success of Mosteller and Wallace's work has encouraged many subsequent research efforts to explore other forms of multivariate techniques in the field of authorship analysis. One such method is *principal components analysis* (PCA) — first employed by Burrows (1989) on a large set of function word frequencies. PCA is capable of capturing deviations across high-dimensional observed features by projecting them into a reduced dimensional space spanned by the principal components in order to find a good separation between texts known to be written by different authors. PCA has since been adopted in numerous other authorship studies with promising results. For instance, Baayen et al. (1996) used it with the syntactical rewriting rules and Holmes and Forsyth (1995) revisited the *Federalist Papers* by applying PCA with vocabulary richness (see details on vocabulary richness in Section 2.1.1 under lexical features).

Since then, a few other multivariate methods have been established. *Clustering analysis* was introduced by Ledger and Merriam for the authorship attribution of plays written by Shakespeare and Fletcher (Ledger and Merriam, 1994). More recently, linear discriminant analysis was explored by Baayen et al. (2002) and Chaski (2005) for multi-genre texts and digital evidence investigation, respectively. These techniques were suggested as an attempt to improve PCA in view of the limitations posed by PCA. It was argued that loss of important information (such as salient indicators distinguishing between individual authors' stylistic properties) may happen in PCA as a result of data reduction (Ledger and Merriam, 1994). Clustering analysis, however, allows all features to be deployed for analytical purposes without, necessarily, reducing the features into some smaller dimensions; and it only relies on clustering algorithms to form clusters separating authors in question, making no use of prior knowledge (of the known work) of the authors. On the other hand, Baayen et al. (2002) claimed that *discriminant analysis* would be more appropriate than PCA when predicting the authorship of an unknown (heldout) text based on known (training) texts of which the authorial fingerprints do exist. In a discriminant analysis, a linear function is created to maximise the difference between groups (authors in the case of authorship attribution); the coefficients of the resulting function are then used to predict the group membership of heldout cases.

The use of *factor analysis* in authorship-related work has also been proposed to model variations across different languages and to identify different registers or genres of texts (Biber, 1995). Being similar to the approach of PCA, a large number of original variables (linguistic features in the case of authorship analysis) are reduced to a small set of derived variables known as factors. Each linguistic feature is assumed to have a certain amount of variability across texts (represented by factor loadings) and such variability is possibly shared by other features represented by the same factor. The shared variance is said to reflect the co-occurrence pattern amongst the linguistic features. If two features tend to be frequent in some texts but rare in other texts, they are said to co-occur and have a high amount of shared variance. Factor analysis seems to be able to identify different feature dimensions based on the derived factors and to model the underlying dimensions of variation. The inferred dimensions of linguistic variation are more easily interpretable with respect to their functionalities (as opposed to PCA). For example, as per Biber (1995), linguistic features of English can be grouped under seven dimensions such as involved versus informational production, narrative versus non-narrative discourse, situation-dependent versus elaborated reference, overt expression of argumentation, and so forth.

Machine learning approaches. Authorship analysis in general can be seen as another form of classification problem. Text classification techniques rooted in machine learning have been applied in numerous authorship studies. Essentially, machine learning methods require a set of training data and a set of testing data represented by vectors of features. Whilst the training data is used by learning methods to develop the classification model, the testing data is subsequently used for evaluation of the resulting model. *Neural networks* were one of the earliest methods to be applied, with the first use by Merriam and Matthews (1994). Their findings demonstrated that a multi-layer perceptron neural network is capable of identifying individual work as well as collaborative work. The attempt to re-attribute the *Federalist Papers* using this type of neural network by Tweedie et al. (1996) achieved consistent results with all the previous studies of the *Federalist* problem. Neural networks were further proven to be useful for stylistic analysis by some recent studies (Graham et al., 2005; Tearle et al., 2008).

Following their promising performance in text categorization (Joachims, 1998), *support vector machines* (SVM) have been a popular choice of classifier in the field of authorship analysis. Diederich et al. (2003) experimented with SVM on authorship attribution and they were able to detect 60-80% of the seven target authors for 2600 newspaper articles. An average accuracy close to 80% was achieved by de Vel et al. (2001) who attempted to conduct author-topic classification on 150 e-mail documents on three different topics produced by three authors. Since then, SVM has been widely adopted particularly in e-mail forensic analysis (Corney et al., 2002; Teng et al.,

2004; Ma et al., 2008). In another authorship-related work, Li et al. (2006) first made use of *genetic algorithms* to select optimal feature sets before applying SVM as the classifier. With this additional phase of feature selection, higher classification accuracy was achieved.

Other machine learning techniques that have been suggested for authorship-related studies include *decision trees* (Abbasi and Chen, 2005), *k-nearest neighbour algorithms* (Hoorn et al., 1999), and *Markov chain models* (Khmelev and Tweedie, 2001). Along with SVM, decision trees were examined by Abbasi and Chen (2005) on web forum messages depicting cybercrime and homeland security issues in Arabic. An accuracy of approximately 72% was achieved by using decision trees in classifying 400 Arabic messages. Although it has been demonstrated that SVM may outperform decision trees by more than 20%, decision trees may be more effective in identifying the salient features for classification and thus provide more insight into the effect of individual features. Likewise, Hoorn et al. (1999) performed a comparative study on three different classification techniques — neural networks, k-nearest neighbour algorithms, as well as Bayesian models — for an authorship attribution task involving 90 poems produced by three Dutch poets. Results show that neural networks were the best relative to the other two techniques tested. From these studies, k-nearest neighbour algorithms seem to be more effective than the Bayesian classifiers, roughly 5-10% better in terms of classification accuracy. Markov chain models, which have been widely used in speech recognition, were proposed by Khmelev and Tweedie (2001). A result of 74% accuracy was obtained when the first-order Markov chain was tested with a large dataset of more than 380 texts from the Project Gutenberg, written by 45 authors.

A consensus was found in the previous studies that machine learning methods on the whole perform better than statistics-based methods. This is simply because the machine learning approach is more tolerant to high-dimensional feature spaces (Zheng et al., 2003, 2006) and requires less parameter tuning (Holmes, 1998). Nevertheless, the choice of features by which the texts in question are to be represented is of comparable importance to the choice of classification method.

2.2 Native Language Identification

Identifying the native language of an author from his or her written utterances in a second or a foreign language can be regarded as a form of authorship analysis. More specifically, this is a subtask of *authorship profiling* (authorship characterisation), one of the subfields of authorship analysis. In the absence of a predefined set of candidate authors, authorship profiling exploits sociolinguistic observation to predict the characteristics or traits of a specific class of authors in order to derive a generic author profile. The resulting author profiles can then serve as the basis for authorship identification. Author characteristics may comprise both demographics and

psychometrics, and the native language of an author falls under the former category (Estival et al., 2007). It is plausible to classify authors into distinguished classes with respect to their native language, on the assumption that authors with different language backgrounds tend to exhibit certain stylistic idiosyncrasies in their second language writing, which might have been influenced by their native language to some extent. This idea underlies the concept of *contrastive analysis* for second language acquisition (Lado, 1957; Wardhaugh, 1970) as well as subsequent work in the field, which we discuss in Section 2.3.

The attempt to profile authors based on their native language is little studied compared to other author characteristics; there is only a handful of research work that had attempted to identify native language for the purpose of authorship profiling (Koppel et al., 2005; Tsur and Rappoport, 2007; Estival et al., 2007), prior to the commencement of the studies carried out in this thesis. Since then, the field of native language identification (NLI) has gained much attention and much related work has emerged; for example, Kochmar (2011), Brooke and Hirst (2012), Swanson and Charniak (2012), and Bergsma et al. (2012).

This presents an unusual challenge for the arrangement of the work in the thesis, in that some of the subsequent work is based on our published work in some ways; the question then is whether to present the subsequent work here or only after we present our own work in the relevant chapter. We have opted to discuss the broad aims of the related work here, and then discuss any implications or insights based on our work in the relevant chapter.

Much of the existing work on NLI is based on supervised machine learning approaches to text classification. The earliest notable work in this classification paradigm is that of Koppel et al. (2005), using features that are mostly lexical in nature, which include function words, character n-grams, and PoS bi-grams, together with a range of spelling mistakes (such as repeated or missing letters, letter inversion, and conflated words) and some neologistic terms. Five different groups of English authors (with native languages Bulgarian, Czech, French, Russian, and Spanish) were selected from the first version of *International Corpus of Learner English* (ICLE) (as described in Section 1.3) — each native language group contributed 258 essays with word length in the range of 580–850. Under a ten-fold cross validation setting, Koppel et al. (2005) gained a relatively high classification accuracy of around 80% against a major baseline of 20% by deploying a SVM classifier with a total number of 1,035 features: 400 function words, 200 character n-grams, 250 rare bi-grams, and 185 frequent error types primarily spelling mistakes (detected mostly through the spelling and grammar checkers of Microsoft Word along with self-designed scripts). Of all the feature sets, combining just function words and character n-grams led to greater than 75% accuracy. Although syntactic features — *syntactic errors* in particular — were highlighted by Koppel et al. (2005) as potentially useful features, they seemed to only investigate this idea at a

surface level, by considering the rare PoS bi-grams as indications of ungrammaticality.

The work of Koppel et al. (2005) was later replicated and analysed by Tsur and Rappoport (2007). With the idea of using character n-grams to approximate a language's syllables, they hypothesised that the choice of words in second language writing is highly influenced by the frequency of native language syllables — the *phonology* of the native language. Tsur and Rappoport (2007) claimed that, rather than manipulating sophisticated linguistic features, character bi-grams alone can potentially lead to a good classification accuracy for authorship profiling based on the native language of the authors. Deploying only the 200 most frequently occurring character bi-grams, they achieved 66% accuracy against a baseline of 46.8% (achieved through unigrams comprising of letters and punctuation marks) over a similar set of ICLE data used in Koppel et al. (2005) with the same five native language groups but based on their own random sample. The number of texts randomly chosen in this study was 238 for each native language group with word length between 500 and 1,000. To further test their claims, two other native languages which were not tested in Koppel et al. (2005) — Dutch and Italian — were used to replace French and Spanish in the initial setting, and essentially a similar accuracy of approximately 65% was obtained. The issue of content bias was raised in Tsur and Rappoport (2007), and the approach taken to investigate this was conducted through removing content words selected based on the *tf-idf* (term frequency-inverted document frequency) measure; the resulting classification performance declined only 2-3% and this suggests that the effect of content bias was minimal even if present.

Native language is also among one of the characteristics (gender, age, geographic origin, level of education and psychometric traits) investigated in the authorship profiling task by Estival et al. (2007). Unlike the approach of Koppel et al. (2005), linguistic errors in written texts are not of concern here; rather this study focuses on a range of lexical features (including function words, PoS categories, punctuation symbols, word length, and case attributes of characters) and document-based structural features (such as paragraph breaks and various HTML tags). Using a random forest classifier instead, the classification task yields a relatively good performance of 84% accuracy against the most frequent baseline of 62.9% when the native language alone is used as the profiling criterion with 689 features in total. However, it should be noted that a smaller number of native language groups were examined in this study — Arabic, English, and Spanish. It is also worth noting that this study was carried out on a collection of 9,836 English emails that is not publicly available. As such, comparing this work with the two aforementioned studies (Koppel et al., 2005; Tsur and Rappoport, 2007) is not straightforward.

On the whole, the existing work discussed thus far tackled the problem of native language identification with supervised machine learning, focusing mostly on lexical features. Syntactic features, on the other hand, specifically those that capture grammatical errors, which might po-

tentially be useful for this task, have received relatively little attention. It is only since the commencement of the studies in this thesis that aim at exploiting syntactic knowledge for NLI, that a number of related works have subsequently emerged which largely use as a starting point two studies of ours — Wong and Dras (2009) and Wong and Dras (2011) — that serve as the respective fundamental work of Chapter 3 and Chapter 4 to be presented later.

Kochmar (2011) cited the idea put forward by us in Wong and Dras (2009) of using second language acquisition (SLA) knowledge (discussed next in Section 2.3) for NLI, as the basis for work which extended it in a number of ways. This study involves a number of binary classifications between two Indo-European language groups — Germanic versus Romance, and between closely related languages of each language group, which include Western Germanic versus Northern Germanic, High German versus Low German, German versus Swiss German, Swedish versus Danish, and Catalan versus Spanish. Kochmar (2011) noted that the approach attempted by us in Wong and Dras (2009) of exploring error analysis (to be discussed further in Chapter 3) as a promising direction to identify suitable error types that might be indicative of a particular native language. Features investigated in her work were thus grouped into two types: distributional and error-based. The former include function words, character n-grams, PoS n-grams, and error type rates as well as phrase structure rules (an idea suggested in Wong and Dras (2009) but not then explored). The latter are various types of grammatical and spelling errors systematically identified in the corpus investigated, aiming to address the limitation of ours (Wong and Dras, 2009) that only investigated three manually identified common errors observed in non-native English speakers — namely, subject-verb disagreement, noun-number disagreement, and misuse of determiners. On a publically-available subset of the Cambridge Learner Corpus (CLC) (Yannakoudakis et al., 2011), character-based n-grams were found to be the best indicator, and the best combination set of features for the classification of each language pair contains at least this feature type. The best binary classification accuracy achieved for all the language pairs range between 68–97% against the majority baseline of 50%. Surprisingly, syntactic structures characterised by phrase structure rules (extracted from the RASP parser (Briscoe et al., 2006)) perform poorer than PoS-based or character-based n-grams, although significantly better than the majority baseline (with accuracy rates between 55–65%). Interestingly, a greater number of error types identified in the corpus (in particular those that selected based on information gain) appear to perform better than three error types used by us (Wong and Dras, 2009); and Kochmar (2011) claimed that her findings demonstrated that the three error types are not discriminative for most of her classification tasks except for the Romance language group.

Much more recently, another attempt by us at characterising syntactic structure with parse production rules in the form of context-free grammar (CFG) as published in Wong and Dras (2011)

and to be discussed in Chapter 4 in detail has been replicated and further explored by Swanson and Charniak (2012). Tree Substitution Grammar (TSG) was proposed for NLI tasks through automatically induced TSG fragments — rules of arbitrary parse tree structures with any number of non-terminal and terminal leaves — as classification features. Two different TSG induction algorithms were investigated: one uses a non-parametric Bayesian model and the other is inspired by tree kernels; and the former was found to be more robust in terms of classification performance. Following our experimental setting in Wong and Dras (2011) (i.e. using ICLE essays from seven native language groups), Swanson and Charniak (2012) demonstrate that TSG features induced from the Bayesian approach result in the highest accuracy of around 78% against the baseline of 72.6% obtained through CFG parse rules as well as the tree kernel-induced TSG features (73–77%) with a logistic regression classifier. To avoid potential biases as the artifact of content topics or geographic locations, only terminal symbols of the 598 selected function words were retained for the parse rules.

A somewhat different perspective to NLI is seen in the work of Brooke and Hirst (2012), whose goal is to explore alternative sources for reliable training corpora for tasks in the NLI paradigm. They suggest exploring web-based L1 corpora as a training source to minimise topic biases that could possibly present in existing learner corpora, such as ICLE. In their presentation in the conference meeting of Learner Corpus Research organised by University of Louvain, Brooke and Hirst (2011) claimed that the ICLE corpus might have subject-based clues, leading to some degree of distortion when used for NLI classification tasks. The approach that they took to investigate this effect was to compare the classification performance of two subsets of ICLE essays, in which one is based on a random selection and the other is carefully chosen essays based on a similar topic. The hypothesis is that if content bias is present in the ICLE corpus, the latter data subset would result in a drop in performance. The outcomes show that the classification performance using function words or PoS n-grams (which are supposedly topic-independent) were unexpectedly affected, although only quite moderately. Such a claim would need further support given that Brooke and Hirst (2011) only examined a small subset of ICLE (with four native languages French, Spanish, Chinese, and Japanese) and lack sufficient details in terms of the experimental setup. In any case, as we are ignoring content words in all of our approaches, such an effect should be small.

As an aside, features for NLI tasks can either include content words or not; but this type of feature has largely been avoided by the exisiting work of NLI. A rather interesting related piece of research that demonstrates the usefulness of content words is that of van Halteren (2008), which has demonstrated the possibility of identifying the source language of medium-length translated texts (between 400 and 2500 words). On the basis of frequency counts of word-based n-grams, high classification accuracies from 87% to 97% are achievable in identifying the source language

of speeches from the European Parliament (EUROPARL) corpus (Koehn, 2005). In fact, what van Halteren (2008) uncovered was that there were some salient markers present for each source language in both the linguistic and cultural aspects. Many of these were in fact tied to the content as well as the domain of the speeches. For example, the greeting to the European Parliament is always translated in a particular manner from German to English in comparison with other European languages; and the Dutch speakers seem to be exaggerating their viewpoints both positively and negatively with the Dutch vocabulary. These markers may have suggested an explanation for the high classification accuracy rates.

In relation to this, in reviewing work in the general area of authorship attribution (including NLI), Koppel et al. (2009) discussed the (arguably unreasonable) advantage that content word features can provide, giving clues based on the subject matter of a text, and commented that consequently they "are careful ... to distinguish results that exploit content-based features from those that do not". For this reason, the existing studies we have discussed thus far generally deploy approaches to NLI that similarly do not use content words features. Likewise, we are not using any content word-based features in our NLI studies to be presented in the subsequent chapters.

2.3 Second Language Acquisition: Cross-linguistic Influence

Cross-linguistic influence or *language transfer* has for a long time been one of the key topics of debate in the field of second language acquisition (SLA), centering around the influence of the first (native) language on the acquisition and production of a second language; and in particular the extent to which errors made by learners during their second language learning are the results of the *interference* of their native language. Two early conflicting views of cross-linguistic influence in language learning are *contrastive analysis* and *error analysis*, which have since been amongst the core perspectives in SLA studies (Ellis, 2008; Gass and Selinker, 2008).

2.3.1 Contrastive Analysis and Error Analysis

Contrastive analysis (Lado, 1957) was an early attempt in the field of second language acquisition to study the effects of language transfer in language learning, focusing on predicting the kinds and sources of error that non-native speakers make by conducting comparisons between the first (native) and the second (foreign) languages. It arose out of *behaviourist psychology* (Bloomfield, 1933) that viewed language learning as an issue of habit formation that could be inhibited by previous habits inculcated in learning the native language. The theory was also tied to *structural linguistics*, comparing the syntactic structures of the native and the second languages to find differences that might cause learning difficulties.

Taking a predictive perspective, Lado's work led to the Contrastive Analysis Hypothesis (CAH) framework, which claimed that:

> ...those elements which are similar to [the learner's] native language will be easy to learn, and those elements that are different will be difficult.

The consequence is that there will be more errors made with respect to those difficult elements, hence equating learning difficulties with errors. In this regard, the greater the difference between the first and the second languages, the greater the effect of interference or negative transfer from the first language are to be expected — more errors are predicted.

While contrastive analysis was influential at first (especially in the 1960s), it was increasingly noticed that many errors were common across all language learners regardless of native language, and these could not be explained under the contrastive analysis framework. Furthermore, there was evidence that many errors are not the results of transfer (Dulay and Burt, 1974a) and some predicted errors did not actually occur (Odlin, 1989; Gass and Selinker, 2008). Corder (1967) then presented an alternative linguistic analysis that focuses on the errors that learners make, known as *error analysis*. Under this framework, errors based on contrastive analysis were seen as only one type of errors, termed 'interlanguage' or 'interference' errors; other types were 'intralingual' and 'developmental' errors, which are not specific to the native language but are considered as universal in nature (Richards, 1971). (Section 2.3.1.1 discusses each of these error types in more detail.)

The basis of error analysis is that errors are viewed as the reflection of the learner's competence at a particular stage in the learning process. In other words, errors reflect the knowledge level of the new language that the learner has acquired, regardless of the learner's language background. According to Corder (1967), many errors produced by learners are due to inappropriate application and faulty inferences regarding the grammatical system of a new language to be acquired. Rather than being influenced by the native language, errors are in fact ascribed to the mutual interference of forms and functions within the new language itself (Richards, 1971). Nonetheless, error analysis was not without its own limitations. According to Ellis (2008), error analysis fails to provide a complete picture of language learning as this framework focuses on only what learners do incorrectly but has no account for what learners do correctly. More importantly, as error analysis only investigates what learners do, it has no way to investigate the phenomenon of *avoidance* as discovered by Schachter (1974) — one of the negative effects of cross-linguistic inflluence (see Section 2.3.1.2 for a detailed discussion).

In an overview of contrastive analysis after the emergence of error analysis, Wardhaugh (1970)

noted that there were two interpretations of the CAH, termed the *strong* and the *weak* forms.[3] Under the strong form, all errors were attributed to the native language, and clearly that was not tenable in light of error analysis evidence. In the weak form, these differences have an influence but are not the sole determinant of language learning difficulty. Wardhaugh noted claims at the time that the hypothesis was no longer useful in either the strong or the weak version: "Such a claim is perhaps unwarranted, but a period of quiescence is probable for CA itself".

Such a period of quiescence did in fact occur, starting perhaps around the frequently cited study of Dulay and Burt (1974b), which attempted to count the number of transfer errors in sample texts, and found them to constitute as little as 5% of the total. However, their approach only classified an error as being a transfer one if it could be unambiguously attributed to transfer, and so almost certainly greatly underestimated the effect of L1. In his comprehensive overview of SLA, Ellis (2008) notes that "there can be little doubt that some scholars were too ready to reject transfer as a major factor in L2 acquisition", and that consequently a reappraisal of transfer effects occurred in the 1980s that remains the position today. To sum up, contrastive analysis and error analysis focus on somewhat different aims, although there is naturally an overlap. While contrastive analysis aims to predict (*a priori*) and explain (*a posterori*) transfer errors to account for learning difficulties during the second language acquistion process, error analysis can only be in full support when errors actually manifest in production — otherwise, it might underestimate certain difficulties when learners choose to avoid.

2.3.1.1 Categorisation of Errors

Corder (1967) was one of the first to introduce the distinction between systematic and non-systematic errors in a language. Non-systematic errors generally occur in one's native language and are due to memory lapses, physical states and psychological conditions. Corder named these sorts of errors mistakes and pointed out that they are of no significance in the process of language acquisition, as learners are in general immediately aware of these *mistakes* and make almost instant correction. On the other hand, the sort of systematic errors that are commonly detected during the acquisition of a second language are termed *errors* by Corder. Systematic errors reflect one's competence of a new language at a particular stage of acquisition. This significant distinction in errors made by language learners subsequently led to a more concentrated framework for language errors. Amongst the many research studies examining competence errors of language learners, Richards (1971) identified three sources of (systematic) errors that could be observed during the acquisition of English as a second or a foreign language: *interlanguage (interference)*

[3] Instead of taking the view of strong versus weak, CAH can be seen from the perspective of *predictive* versus *explanatory* (Gass and Selinker, 2008). In the former, one could either make predictions about learning difficulty based merely on comparisons between two languages or perform analysis based on learners' recurring errors.

errors, *intralingual* errors and *developmental* errors; the first is native language-dependent, and the latter two are considered as universal errors potentially made by all the learners of English irrespective of their native languages.

Interlanguage (interference) errors. In essence, interlanguage errors result from the interference of the learner's native language. By its nature, contrastive analysis focuses only on identifying *interference* of the native (first) language with the target (second) language. Instances of interference can be predicted by contrasting syntax and other subsystems — including pronunciation, vocabulary, as well as semantics — of the two languages. Numerous studies of different language pairs have already been explored; in particular, focusing on learners of English with different native language backgrounds. To give some examples: Dušková (1969) investigated Czech learners of English in terms of various lexical and syntactical errors; Light and Warshawsky (1974) examined Russian learners of English (as well as French learners to some extent) on their improper usage of syntax as well as semantics; Guilford (1998) specifically explored the difficulties of French learners of English in various respects, from lexical and syntactical to idiosyncratic; and both Chan (2004) and Mohamed et al. (2004) targeted grammatical errors of Chinese learners in English.

In addition, many other studies examining interlanguage errors restricted their scope of investigation to a specific grammatical aspect of English in which the native language of the learners might have an influence. Granger and Tyson (1996) examined the usage of connectors in English by a number of different native speakers — French, German, Dutch, and Chinese. Vassileva (1998), investigated the employment of first person singular and first person plural by another different set of native speakers — German, French, Russian and Bulgarian. More recently, Slabakova (2000) explored the acquisition of telicity marking in English by Spanish and Bulgarian learners. Similarly, Yang and Huang (2004) studied the impact of the absence of grammatical tense in Chinese on the acquisition of English tense-aspect system (i.e. telicity marking). Sabourin et al. (2006) discovered the transfer effects of native language in the acquisition of the grammatical gender system of a second language. Bannai (2008), whose focus is on Japanese learners of English, examined the placement of verbs and adverbs within sentences.

It has also been discovered that spelling errors in English can plausibly be attributed to the negative transfer of one's native language. Back in 1970, a comprehensive study on English spelling errors was conducted by Oller and Ziahosseiny (1970), including a contrastive analysis among languages that are of different writing systems (roman versus non-roman). In connection with this, Sun-Alperin and Wang (2008) found that Spanish learners of English specifically produce more spelling errors in English vowels due to different graphemes in the vocabulary of English

and Spanish.

Intralingual errors. While errors owing to the effects of language transfer vary across learners of a specific language, errors that are found to be common are the results of the learning strategies adopted by learners throughout the process of acquiring the target language. Intralingual errors are one type of those common errors regarded by Richards (1971) as errors that reflect the general characteristics of *rule learning* of the target language, which include faulty generalisation, incomplete application of rules, and failure to learn conditions (restrictions) under which rules apply. Over-generalisation occurs when learners create deviant sentence structures based on their previous experience of other regular structures in the target language. This could be seen as the result of learners attempting to reduce their linguistic efforts. Some learners exploit over-generalisation for redundancy reduction in which they find certain grammatical items of the target language carry no significant distinctions in sentence realisation. Over-generalisation errors are usually observed between nouns and verbs in terms of number and tense agreement (Richards, 1971; Mohamed et al., 2004). As an example, the omission of '-s' in the third person singular of English is a common error: *the author **give** us some insight*.

Further to the generalisation of deviant structures, learners sometimes fail to recognise the restrictions of the existing structures of the target language, applying rules to contexts where they do not actually apply. The use of analogy may account for such rule restriction errors. An instance of such errors is apparent in the misuse of prepositions (Richards, 1971; Mohamed et al., 2004). Learners who have discovered a particular preposition with a specific type of verb tend to apply an analogy to use the same preposition for some related verbs. For instance, *my supervisor **talked to** me* would lead to **my supervisor **asked to** me*. Another failure in recognising the rule restrictions can potentially be witnessed in the usage of English articles (Richards, 1971; Mohamed et al., 2004). To some extent, the interference of native language can account for the misuse of articles — in part due to the absence of this grammatical category in some languages, such as Czech, Russian as well as Chinese. However, some studies argue that article misuse is indeed derived from the complexities of the English article system itself (Light and Warshawsky, 1974). In most cases, learners are still not competent in complying with the restrictions of article usage; instead, they simply reason out a deviant usage on the basis of their experience and perhaps of their own intuitions (Richards, 1971).

Nevertheless, difficulties in constructing questions are particularly observed across learners of English regardless of their language backgrounds (Richards, 1971). And to some degree, learners of English are found to be confused over verb usage. Incorrect use of participial clauses — for those clauses whose verb element is non-finite — was noted by Granger (1997) in a comparative

study between native and non-native English writers. Application of the rules of subject-verb agreement is another area where learners of English have difficulty acquiring complete competence (Vigliocco et al., 1996; Franck et al., 2002).

Developmental errors. In addition to interlanguage and intralingual errors, the third type of errors that could possibly be observed amongst learners of a second language are developmental errors. With limited knowledge or experience already obtained at a particular stage of the acquiring process, learners have a tendency to build up false hypotheses about the target language. This is also largely due to faulty comprehension of distinctions in the target language (Richards, 1971). The occurrence of false hypotheses is, nevertheless, expected to decline over a period of time as learners gain more in-depth knowledge of the language. The resulting errors are thus regarded as developmental ones.

Development errors observed during the stages of interlanguage development are similar to those that one observed in first language acquisition. They are generally supported by the *U-shaped behaviour* in language learning. U-shaped behaviour reflects in three stages of linguistic use: in the earliest stage, a learner produces some linguistic forms that conform to target-like norms (without errors); at the second stage, a learner appears to lose what had been acquired at the previous stage, and demonstrates linguistic behaviour that deviates from the target-like norms and manifests errors; but eventually at the third stage, the correct target-like forms resurface (Ellis, 2008; Gass and Selinker, 2008). For example, the acquisition of English morphemes for past tense forms: first *ate* then *eated* and finally back to *ate*. Longitudinal studies can often be utilised to obtain evidence for developmental errors, but a particular concern is that one has to determine the stage of development that each learner being examined has reached. In order to rule out those errors that might due to transfer effect (interference errors), examination should involve groups of learners of different native languages but at a similar stage of development.

A point to be noted is that there are other ways to distinguish errors apart from Richards' categorisation of errors. Many SLA researchers have resorted to only a general distinction between interlanguage errors (those that can be attributed to the native language) and intralingual errors (those that are due to the second language being learned subsuming developmental errors), as discussed in Ellis (2008) and Gass and Selinker (2008).

2.3.1.2 Effects of Cross-linguistic Influence

Work subsequent to contrastive analysis and error analysis has found that the effects of cross-linguistic influence can in fact manifest in learner language in a number of ways. Such transfer

effects need not be solely negative, but can be positive and facilitating. In addition to positive transfer, the types of negative transfer to be discussed below include avoidance (underproduction) and overuse (overproduction). (Interference errors as mentioned in the previous subsection, Section 2.3.1.1, is one type of negative transfer.)

Positive transfer. Cross-linguistic similarities between the native and the second languages can facilitiate learning of the second language. Kellerman (1995) claims that:

> ...the more similar the languages are at some point, the more likely the native language is to influence the development of the second language.

Odlin (1989) argued that the effects of positive transfer can only be identified when comparative studies are performed on learners with different native language backgrounds. Facilitating effects have been evident in the acquisition of many of the subsystems of the target (second) language, including vocabulary, syntax, phonology, and orthography. It is obvious that sharing a large number of cognates between two languages (such as English and French, or German and Dutch) can promote vocabulary learning in a reduced time (Odlin, 1989; Ellis, 2008).[4] Similarly, given the similarity of the Chinese and Japanese writing systems, learners of Japanese from Chinese background would have the advantage over those with English background especially in learning *kanji* (Ellis, 2008).

Avoidance. Avoidance or *underproduction* of certain linguistic structures are the type of phenomenon that error analysis would underestimate since error analysis focuses on what learners do but neglects what learners do not do. Gass and Selinker (2008) claimed that the source of avoidance is indeed controversial. There are claims that differences between the native and second languages are the major source of avoidance, but the opposite holds as well — when there is greater the similarity between the two languages, learners may doubt the reliability of these similarities.

Nonetheless, the influential study of avoidance by Schachter (1974) is a demonstration of the first claim. Her study found that Chinese and Japanese learners of English made fewer (restrictive) relative clause errors compared to Persian and Arabic learners (see Table 2.1) despite the latter group having relative clause structures that are closer to those of English. The underlying reason was that Chinese and Japanese learners avoided using such constructions. Schachter (1974) hypothesised that the difficulty that Chinese and Japanese learners have with relative clauses may be due to the fact that their native languages are predominantly left-branching with prenominal

[4] Cognates between two languages, on the other hand, can also lead to spelling errors as observed by Oller and Ziahosseiny (1970) and Sun-Alperin and Wang (2008).

Table 2.1: English relative clause production by five native language groups (with American as the control group) in Schachter (1974)

Native Language	Correct	Error	Total	% of Errors
Persian	131	43	174	25
Arabic	123	31	154	20
Chinese	67	9	76	12
Japanese	58	5	63	8
American	173	0	173	0

relative clauses, while English is right-branching with postnominal relative clauses (as are Persian and Arabic). The examples below (from Gass and Selinker (2008))[5] illustrate how a noun is modified by a relative clause in English (2.1a), Persian (2.1b), Chinese (2.1c), and Japanese (2.1d), respectively.

(2.1) a. I saw [the woman [who speaks English]$_{S'}$]$_{NP}$.

b. [an zæn-ra [ke inglisi hærf-mi-zæn-e]$_{S'}$]$_{NP}$ did-æm.
that woman-OBJ that-COMP English speak-3SG saw-1SG

c. wǒ kàndào [nèi gè [shuō yīngyǔ de]$_{S'}$ nǚrén]$_{NP}$.
I saw that ge-CLASS speak English de-COMP woman

d. watashi-wa [[eigo-o hana-su]$_{S'}$ josei-o]$_{NP}$ mimashi-ta.
I-NOM English-OBJ speak-PRES woman-ACC see-PAST

Avoidance can be the result of the complexity of the target (second) language structures themselves (Gass and Selinker, 2008). In relation to this, the study of Dagut and Laufer (1985) that focuses on the acquisition of phrasal verbs in English found that Hebrew learners tend to use one-word verbs more frequently than the equivalent phrasal verbs. For example, they preferred verbs like *enter*, *stop*, and *confuse* over phrasal verbs like *come in*, *shut off*, and *mix up*. When choosing to use phrasal verbs, they would resort to those that are semantically more transparent (such as *go out* and *get up* instead of *give in* and *look up to*). On the other hand, Kleinmann (1977) discovered that psychological states such as anxiety could be another factor that accounts for the occurrence of avoidance.

Overuse. Overuse or *overproduction* can be viewed as the consequence of underproduction of some structures that perceived to be difficult by learners (Odlin, 1989; Ellis, 2008). Taking relative clauses as an example, Chinese or Japanese learners would resort to more simple sentences to

[5]Note that we added more detailed glosses on these linguistic examples and with the help of Yasaman Motazedi confirming the Persian one.

avoid using constructions that they find difficult. In the study of Schachter and Rutherford (1979), Chinese and Japanese learners of English were seen to rely heavily on the concept of topic and overproduced sentences of *pseudo-passive*, as in the examples (2.2) below. This discourse level of transfer effect is claimed to be due to the topic-comment structure commonly found in Chinese and Japanese.

(2.2) a. There is a small restaurant near my house in my country.

b. Most of food which is served in such restaurant have cooked already.

Overuse phenomena can occur as a result of an intralingual process such as *overgeneralisation* (Ellis, 2008). It has often been noted that second language learners of English tend to overgeneralise for irregular verbs in English with the regular past tense inflection, such as *eated* versus *ate*. Another potential form of overgeneralisation, as aforementioned, is that second language learners might overgeneralise for the third person singular in terms of subject-verb agreement and consequently omitting the '-s' for the verb, given the endingless form is more common (for both the first and second persons).

2.3.2 Other Perspectives on Second Language Acquisition

In the recent years, a range of other SLA frameworks based on diverse perspectives have emerged to supplement contrastive analysis and error analysis: some are psycholinguistically oriented such as the *competition model* of MacWhinney and Bates (1989) and the *processability theory* of Pienemann (1998) and others are based on *sociocultural theory* (Lantolf, 2000; Lantolf and Thorne, 2006). We do not review them comprehensively in this thesis as we do not draw on any of the specific ideas; but we describe a few of them briefly to convey some more recent directions of the field.

The competition model is a theoretical framework that accounts for language acquisition (both first and second languages) and sentence processing. The model rests on the idea that there are various competitive linguistic cues and their strengths, such as syntax (word order), morphology, and semantics (animacy) that the learner relies on for sentence meaning interpretation (MacWhinney, 1987; MacWhinney and Bates, 1989). During acquisition and usage of a language, the activation of these competitve cues are controlled by a number of counteracting cognitive mechanisms (MacWhinney, 2008). With respect to second language acquisition, learners are often confused between the first and second language cues, and tend to rely on their first language interpretation strategies.

The underlying idea of the processability theory model of second language acquisition is based on the logic that the learner can process only those second language constructions permitted by the processing resources (grammatical information) available at any stage of development (Pienemann, 1998, 2005). Within this framework, the transfer effects of the native language are seen as partially due to the constraints of the processability hierarchy: the first language forms can be transferred to the second language only when they can be processed in the developing system.

The sociocultural approaches to second language acquisition take a rather different view as compared to those frameworks from the cognitive paradigm. Taking the general theory to human mental development as its basis, the sociocultural theory posits that the knowledge of a second language is internalised through sociocultural interactions (Lantolf, 2000). In particular, mediation from the external environment (such as recasts and feedback) is crucial for the learners to recognise what they can handle themselves and when they need guidance and support.

2.3.3 Application to Native Language Identification

In light of the focus of this thesis, with the advancement in natural langauge processing (NLP) techniques, we can now revisit and make use of the weak form (explanatory perspective) of the contrastive analysis as well as the notions of avoidance and overuse by taking a probabilistic approach to native language identification (NLI). Specifically, interlanguage (interference) errors or differences in distribution of usage of syntactic constructions, as represented by syntactic structural differences observed in the form of CFG production rules for instance, may be characteristics of the native language of a learner. We can make use of the occurrence of these to come up with a revised likelihood of the native language.

Although both intralingual and developmental errors as well as other aspects (such as psycholinguistics and sociolinguistics) could be complementary for the task of NLI, they are not of focus here but are interesting areas for future investigation.

Chapter 3

A Preliminary Error-based Analysis

The goal of the current chapter is to make a preliminary investigation on the usefulness of syntactic errors in native language identification (NLI), as motivated by the related work discussed in Chapter 2.[1] Given that it is uncertain whether our intuition of exploiting syntactic features or syntactic errors *per se* is at all useful for the task of NLI, the work presented in this chapter first sets out to test this on a small scale. Amongst the frequently encountered types of syntactic error in non-native English users, which it has been argued can be attributed to language transfer to some extent, the three types of error to be investigated are *subject-verb disagreement*, *noun-number disagreement*, and *misuse of determiners*.

The organisation of the current chapter is two-fold. Firstly, the potential of syntactic errors is explored with respect to its usefulness in determining the native language of the authors (writers) of interest. We motivate this by the intuitions of the Contrastive Analysis Hypothesis (CAH) used in a reverse direction (in relation to transfer analysis), in which based on the syntactic errors, our aim is to detect the native language of the author. We will first conduct statistical analysis (which include ANOVA tests) to gauge whether there is any distributional difference with respect to the errors produced by authors of different native language background prior to using them as features to identifying authors' native language. Secondly, the current chapter will investigate to what extent basic syntactic features (i.e. grammatical errors manually identified) can be useful for native language identification in addition to the other commonly used lexical features (such as function words, character n-grams, and PoS n-grams) from the existing related work — with a particular interest in those that were adopted by Koppel et al. (2005).

[1] The core work in this chapter was published in Wong and Dras (2009).

3.1 Syntactic Errors

For the purpose of the current study, only the three major syntactic error types as mentioned above are explored. Here we detail each of the syntactic error types in turn and provide some examples for each type.

Subject-verb disagreement. This error type refers to situations in which the subject of a sentence disagrees with the verb of the sentence in terms of number or person. Two excerpts adapted from our training data (discussed in Section 3.2) demonstrate this sort of error:

(3.1) a. * If **the situation become** worse ...

b. If **the situation becomes** worse ...

(3.2) a. * There is a suggestion to construct **a second railway** which **link** between ...

b. There is a suggestion to construct **a second railway** which **links** between ...

Noun-number disagreement. This error type refers to situations where a noun is in disagreement with its determiner in terms of number. Two excerpts from our training data that demonstrate this sort of error:

(3.3) a. * They provide **many** negative **image** ...

b. They provide **many** negative **images** ...

(3.4) a. * **These finding** prove that ...

b. **These findings** prove that ...

c. **This finding proves** that ...[2]

Misuse of determiners. This type of error refers to situations in which the determiners (such as articles, demonstratives, and possessive pronouns) are improperly used with the nouns modified by them. Specifically, these are situations involve missing a determiner when it is required as well as having an extra determiner when it is in fact not needed. Here are two excerpts from our training data demonstrating this sort of error:

(3.5) a. * Cyber cafes should not be located outside **airport**.

b. Cyber cafes should not be located outside **the airport**.

c. Cyber cafes should not be located outside **airports**.[3]

[2] This additionally resolves subject-verb disagreement.

[3] The second grammatical form of this example indicates an error of noun-number disagreement instead.

Table 3.1: Presence or absence of the three grammatical features in each native language

Native Language	Subject-verb Agreement	Noun-number Agreement	Use of Determiners
Bulgarian	+	+	+
Czech	+	+	–
Russian	+	+	–
French	+	+	+
Spanish	+	+	+
Chinese	–	–	+
Japanese	–	–	+

(3.6) a. * I have read **nice quotation** . . .

b. I have read **a nice quotation** . . .

It is worth highlighting that the three grammatical phenomena might or might not present in each of the native languages under investigation (see Table 3.1 for a brief summary of these). Generally speaking, a '–' indicates that such a phenomenon does not exist or exists to a much lesser extent in a particular native language; where a '+' indicates that the phenomenon does exist but might not coincide precisely with the English one. Some of the general observations are: Spanish and French have much more extensive use of determiners than in English in which determiners are often inflected to agree with gender and number for the nouns that they determine (Jones, 1996; Zagona, 2002); the presence or absence of determiners in Bulgarian has no effect on aspectual interpretation unlike in English (Slabakova, 2000); and as for Chinese and Japanese, the usage of determiners is far less frequent than that of the other languages and generally more deictic in nature (Robertson, 2000; Butler, 2002). On the other hand, conjugations (and consequently subject-verb agreement) are more extensive in the European languages than in English (Jones, 1996; Comrie and Corbett, 2002; Zagona, 2002).

3.2 Data and Examples

The data to be examined in the current study is taken from the second version of the International Corpus of Learner English (ICLE) compiled by Granger et al. (2009) as mentioned in Section 1.3. This was also the data used by Koppel et al. (2005) and Tsur and Rappoport (2007), although they used the first version of the corpus.

Taking Koppel et al. (2005) as a point of reference, we employed the five languages examined by them — Bulgarian, Czech, French, Russian, and Spanish — and added Chinese and Japanese. Such a combination would give us a fair coverage of different language groups. This also allows

Table 3.2: Average text length of each native language

Native Language	Average Text Length (words)
Bulgarian	668
Czech	747
Russian	639
French	692
Spanish	621
Chinese	570
Japanese	610

us to investigate languages that have quite different grammatical features from English. As noted before, subject-verb and noun-number disagreements are not present in Chinese and Japanese.

For each of the native languages, we randomly chose from amongst the essays with length of 500–1000 words. For the purpose of the current study, for each native language, we have 70 essays for training and 25 essays for testing, with another 15 essays set aside for a specific purpose to be explained later in the chapter. By contrast, Koppel et al. (2005) made full use of all the 258 texts from their version (ICLE, Version 1) for each language and evaluated through the common practice of 10-fold cross validation. Table 3.2 gives the average text length of our sample (ICLE, Version 2) broken down by native language; and Figures 3.1 – 3.4 illustrate some excerpts of the essays selected for Czech, French, Chinese, and Japanese, respectively.

```
<ICLE-CZ-PRAG-0011.3>
. . .
Television offers three basic things are important for modern
man: information, entertainment. These things we can get very
easily, in very comfortable way. We just only sit in room and
press buttom. In a few minutes we get all information about the
war in Jugoslavia, how is the President of Lapland and what will
the weather be like. In other few minutes we can learn Spanish
with charming Spanish teacher and if we are not tired young
couple invite us for dancing course. All these things we are able
to get through within few hours without moving from our house.
. . .
```

Figure 3.1: An excerpt of an ICLE essay whose author is of Czech background

```
<ICLE-FR-ULB-0022.2>
. . .
The different courses students attend to become teacher are
another example which illustrates the theoretical aspect of
university. As a matter of fact it is very different to learn
various pedagogical methods and to be in front of twenty pupils
who are most of the time making fun of you. University will not
tell you how to react; maybe you will be given some advice but in
class you will be alone.
. . .
```

Figure 3.2: An excerpt of an ICLE essay whose author is of French background

```
<ICLE-CN-HKU-0085.1>
. . .
The development of country park can directly elp to alleviate
overcrowdedness and overpopulation in urban area. According
to <R>, the population density in urban area is rising at an
alarming rate with the incoming mainland immigrants and economy
growth. As such policy helps to develop the New Territories (N.T)
and north-east part of Lantau Island, more people can more to
there with the improvement in tis infrastructure facilities and
transportation network. Large scale of public estate and elderly
care-centre can be built there. This will lead to a more balanced
and even distribution of the population.
. . .
```

Figure 3.3: An excerpt of an ICLE essay whose author is of Chinese background

```
<ICLE-JP-SWU-0007.2>
. . .
But causes for those problems (lazy college students and juvenile
crime) exist not only young people but Japanese educational
system, I think. Of course, as each person has his/her
responsibility, when he/she acts something, its responsibility
goes with his/her actions. However we should improve going to
cram school from very young age and very fierce examination
hell to enter university. To continue these things becomes
young people' stress, I guess. My suggestion is that Japanese
education system should follow American style. For, to graduate
from university is more difficult than entering university and
students are to pay their school expenses by themselves. By doing
in that way, they will study harder become independent, and get
to have their own responsibility.
. . .
```

Figure 3.4: An excerpt of an ICLE essay whose author is of Japanese background

3.3 Analysis Procedures

Given the main focus of the current chapter is to uncover whether syntactic features are useful in determining the native language of the authors, we require some sort of analysis procedures that could provide us with some insight as to how different are the distributions of the syntactic features (errors) detected from the seven groups of non-native English users. For the purpose of the current chapter, two widely adopted approaches were conducted: *statistical analysis* with hypothesis testing and *text classification* with supervised machine learning.

Statistical analysis. We first performed some basic descriptive statistics to gain a general idea of the distributions of the three syntactic error types. A range of hypothesis testing with single-factor ANOVA tests (having the native language type as the factor) were then carried out to gauge the extent to which the distribution of each error type in fact significantly differed across the seven groups of non-native English learners. If the ANOVA result is statistically significant, that suggests that non-native English learners with different native language backgrounds tend to produce different proportions of errors for a particular syntactic type.

Text classification. A second approach is to take syntactic errors as a type of feature for text classification through supervised machine learning. A general principle of text classification is that if a feature is a good discriminant, this would result in high classification accuracy rates having more texts classified into the correct groups (or classes). With respect to our study, if we expect syntactic features (such as syntactic errors) to be a good marker, using these features should lead to a reasonable accuracy in classifying the authors according to their own native language. In other words, the resulting classification accuracy is at least some form of indication as to how useful a particular feature is for a specific classification task.

Amongst the available machine learners as classifiers, a Support Vector Machine (SVM) was chosen simply because its performance in classification problems and its ability in handling high dimensional feature space have already been well attested (Joachims, 1998). We deployed one of the SVM online tools — LIBSVM (Version 2.89) by Chang and Lin (2001). It should be noted that all the classifications were conducted under the default settings, where the radial basic function (RBF) kennel was chosen as it is appropriate for learning non-linear relationships between multiple features. The kernel was tuned to search for the best pair of (C, γ) parameters with the default 5-fold cross validation on the training data (i.e. the 490 ICLE essays).

3.4 Error Detection Tools

In addition to the machine learning tool, we also required some form of error detection tools or grammar checkers to assist in detecting the three syntactic error types. Two tools were adopted: the online grammar checker *Queequeg*,[4] and a *tagger* that we trained in a fairly simplistic way for error detection purposes.

Queequeg. This is a very small English grammar checker that specifically detects the three error types that are of concern in our study. Unfortunately, the creator of the software acknowledges that this grammar checker suffers from high false positives, especially for detecting errors of determiner misuse. Further evaluation was performed on Queequeg and we will discuss its limitations in more detail when we make comparisons between the two error detection tools.

Re-training Tagger. In view of the limitations of Queequeg, the alternative approach is to make use of a tagger for error detection through supervised learning. We first extracted a portion of the Brown corpus (with approximately 3600 sentences) and modified it by introducing the three types of syntactic errors in a semi-automatic manner and marked each of the error types with a specific part-of-speech (PoS) error tag.[5] It is worth noting that the idea of training a tagger on an artificial erroneous corpus such that the trained tagger would be able to tag (or detect) ungrammatical sentences is quite similar in spirit to the idea used in Foster (2007) of generating an ungrammatical treebank for training a parser in order to parse ungrammatical sentences. (The reader may refer forward to Section 4.2 for a longer description on this.)

For each error of subject-verb disagreement, the error tag of **#SVD** is assigned to the verb that is in disagreement with its subject. As for noun-number disagreement errors, the error tag of **#NND** is assigned to the determiner that disagrees in number with the noun it modified, and for errors of determiner misuse, the noun itself is tagged with the error tag of **#DTM**. An example of each error type augmented from the Brown corpus is presented below.

(3.7) a. He said evidence **was/BEDZ** obtained in ...

b. * He said evidence **were/#SVD** obtained in ...

(3.8) a. **These/DTS** crucial questions must be answered ...

b. * **This/#NND** crucial questions must be answered ...

[4] Queequeg software is available at `http://queequeg.sourceforge.net/index-e.html` (as of July 19, 2012).

[5] Each sentence generally consists of one error type; only a few have more than one error type.

(3.9) a. There were as **many/AP** as 25,000 schools ...

b. * There were as **much** as 25,000 **schools/#DTM** ...

Briefly, we implemented a separate simple pattern-matching and rule-based system (in Python) for each error type. A segment of each of the original Brown sentences that are presumably grammatical is augmented with one of the three error types and the corresponding PoS tag is also modified to be the erroneous one. For instance, to introduce an error of subject-verb disagreemment, we first locate the subject of a selected sentence and search for the corresponding verb that should be in agreement with it, such as *evidence* and *was* as given in Example 3.7. We then introduce an error by changing the verb from the singular form to the plural form (or vice versa). Likewise, the initial PoS tag of the verb is replaced with the new tag indicating an error of subject-verb disagreement. Hence, for this example, *was* is changed to *were* and its correspoding PoS tag, **BEDZ** is replaced with **#SVD**. Similar procedures are applied to the other two error types, where the system would have to find a noun and its determiner in a sentence and augment the determiner either to disagree with the noun (as an error of noun-number disagreement) or introduce an inappropriate determiner (as an error for determiner misuse). To verify that the errors are correctly introduced through the automated systems (Python scripts), we also conducted a random manual check on the erroneous sentences and rectified them if required.

Having the sub-corpus of Brown augmented with flawed sentences, we trained a tagger on it — the Brill tagger provided by NLTK (Bird et al., 2009)[6] — to recognise the possible patterns of these error types. When the Brill tagger is used for tagging erroneous sentences, it would be able to identify the occurrences of these error types and denote them with the appropriate error tags. We then used the trained Brill tagger on our sample of ICLE. The tagger currently performs at around 95% accuracy per tag having trained on the self-created erroneous corpus (i.e. the augmented Brown corpus, with 90% of the corpus used for training and the remaining for evaluation purposes).

As a side note, one issue that we needed to consider throughout the process of introducing errors for the construction of the erroneous corpus was when dealing with the plurality and countability of nouns. Plurality of nouns is essential as when constructing noun-number disagreement errors, we need to know whether a noun is in its plural or singular form before deciding whether a singular or non-singular determiner should be used to modify that noun. Likewise, when dealing misuse of determiners (articles in particular), knowing whether a noun is countable or uncountable is crucial in deciding whether an article should be present. Unfortunately, extensive lists of such nouns broken down by plurality or countability are quite limited and not publicly available. To

[6] NLTK software is available at `https://sites.google.com/site/naturallanguagetoolkit/Home` (as of July 19, 2012).

Table 3.3: Results of false positive and false negative of each error detection tool (%)

Syntactic Error	Retrained Tagger	Queequeg
Subject-verb disagreement		
False positive	4.36	11.99
False negative	56.06	34.02
Noun-number disagreement		
False positive	7.45	14.14
False negative	62.62	70.00
Misuse of determiners		
False positive	2.49	80.54
False negative	93.05	34.54

this end, we had to semi-automatically induce four separate sets of noun for each lexical-semantic property: singular, plural, countable, and uncountable.[7] Here, we implemented a Python script to classify the nouns based on simple heuristics, and then manually check through the resultant noun sets (particularly the countable and uncountable ones) with an online dictionary, Wikitionary.[8] The basic idea is that countable nouns generally have an inflected form for plurality while uncountable nouns do not.

Evaluations on error detection tools. To get an insight into the ability of each error detection tool to identify the three types of syntactic error, a manual evaluation was carried out on a sample of the training data (of 35 essays). As shown in Table 3.3, the results reveal that the retrained tagger generally exhibits much higher false negative rates, whereas the online grammar checker Queequeg has relatively high false positive rates, particularly in detecting errors of determiner misuse.[9]

It was noticed that Queequeg fails to distinguish between countable and uncountable nouns. Consequently, it tends to generate more false positives especially when determining whether the determiners are in disagreement with the nouns they modify. An example of such false positive generated by the grammar checker is as follows:

(3.10) It could help us to save **some money** ...

The above sentence would be inappropriately detected as ungrammatical. Similarly, the approach of retraining a tagger for error detection is not without its limitations. One particular concern worth mentioning is that the current tagger was trained on a relatively small data set (with

[7] For the current study, we have 12158 singular nouns, 6514 plural nouns, 4444 of countable nouns, and 245 uncountable nouns.

[8] Wikitionary is available at `http://en.wikitionary.org`.

[9] Given that the class sizes are relatively small here, we present only the false positive and false negative rates, rather than the accuracy.

approximately 3600 erroneous sentences modified from the original Brown corpus), and as such it will not have learned all the cases of error pattern for each error type. Nevertheless, they appear to be complementary to one another in terms of the errors detected; they were thus both being used as the syntactic features for the classification task of NLI.

3.5 Syntactic Features in Isolation

Given that the aim of the current chapter is to conduct a preliminary investigation on the potential usefulness of syntactic features in determining the native language of the authors, syntactic features are first examined separately. Here we present both results from statistical analysis and supervised learning.

3.5.1 Statistical Analysis on Syntactic Errors

Boxplots. Boxplots generally represent the median, the quartiles, and the range of the lowest and the highest values. This gives an overview of the distribution of each error type. Figures 3.5 – 3.10 depict the distributions of the three syntactic error types as observed in our ICLE training data of 490 essays by seven distinct groups of non-native English users. The distributions of each of the three syntactic errors are presented in two separate figures: the first contrasts between the actual number of errors (manually counted) and the number of errors detected by Queequeg (-Qg); and the second between the actual number of errors and the number of errors detected by the retrained Brill tagger (-Tg).

Note that the frequency of each error type presented in these figures is normalised by the corresponding text length (measured by the total number of words). This is to ensure that the chances for the occurrence of syntactic errors are equal across the seven sub-corpora of different text lengths, as given in Table 3.2. (An ANOVA test shows that the average text lengths are indeed significantly varied at the confidence level of 95%.)

These boxplots do show some variability amongst non-native English users with different native languages with respect to their syntactic errors. This is most obvious in Figures 3.9 and 3.10, illustrating the distribution of errors concerning misuse of determiners. (The boxplots of actual errors give the clearest picture.) Such a phenomenon could possibly be explained by the interference of native language as indicated in the notion of contrastive analysis. As depicted in the figures, Czech and Chinese as well as Japanese learners seem to have more difficulties when dealing with determiners as compared to French and Spanish learners. This appears plausible since determiners (especially articles) are absent from the language system of Czech and are less frequently used

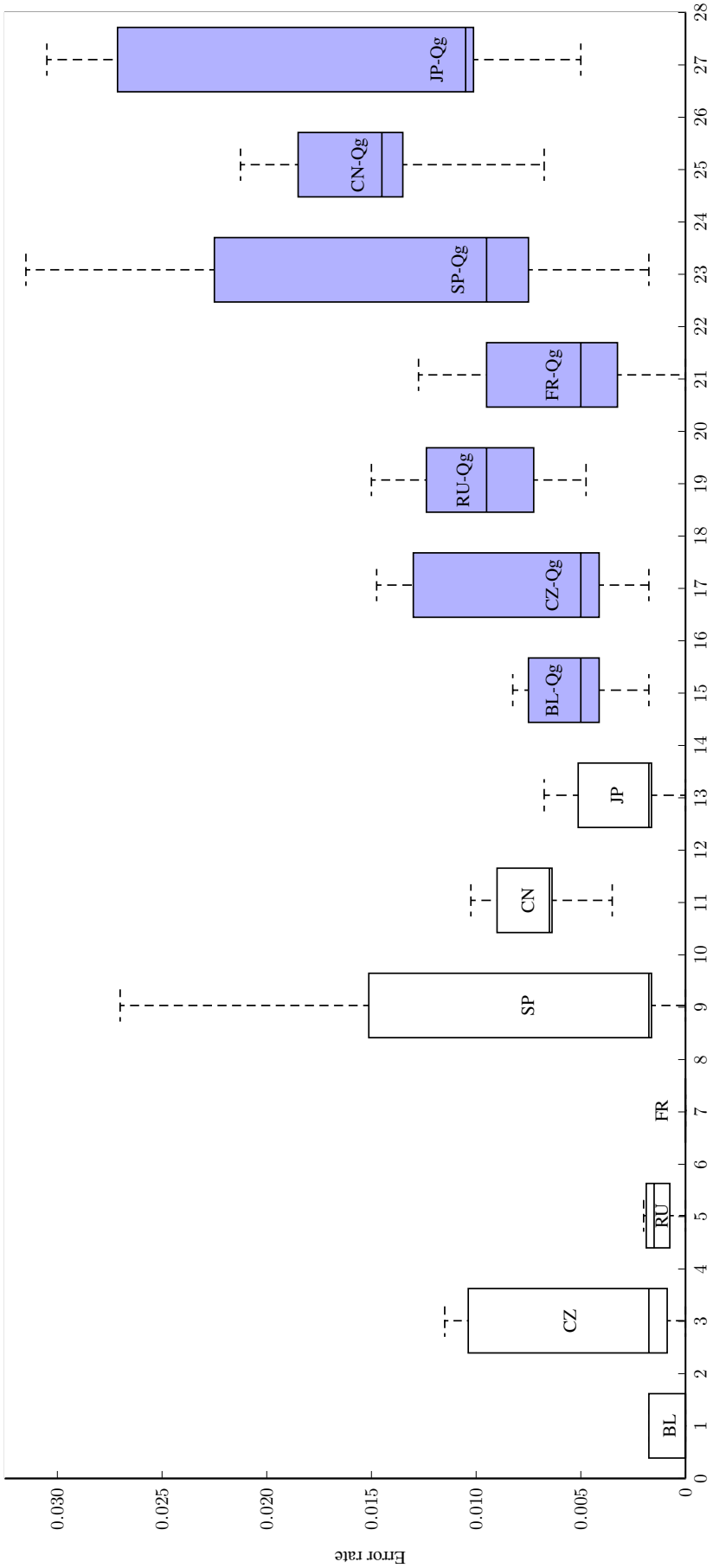


Figure 3.5: Boxplot for subject-verb disagreement — actual errors vs. those detected by Queequeg (-Qg)

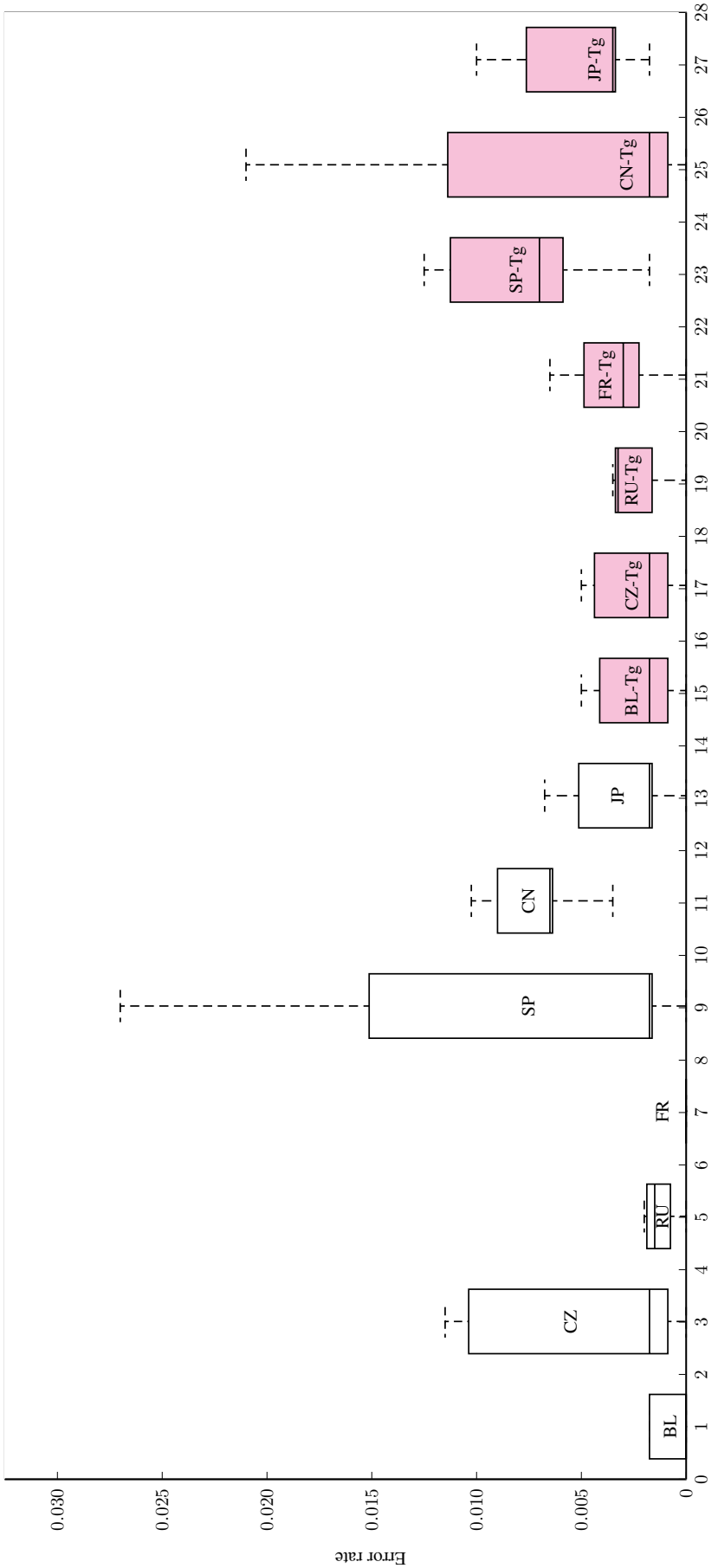


Figure 3.6: Boxplot for subject-verb disagreement — actual errors vs. those detected by the retrained Brill Tagger (-Tg)

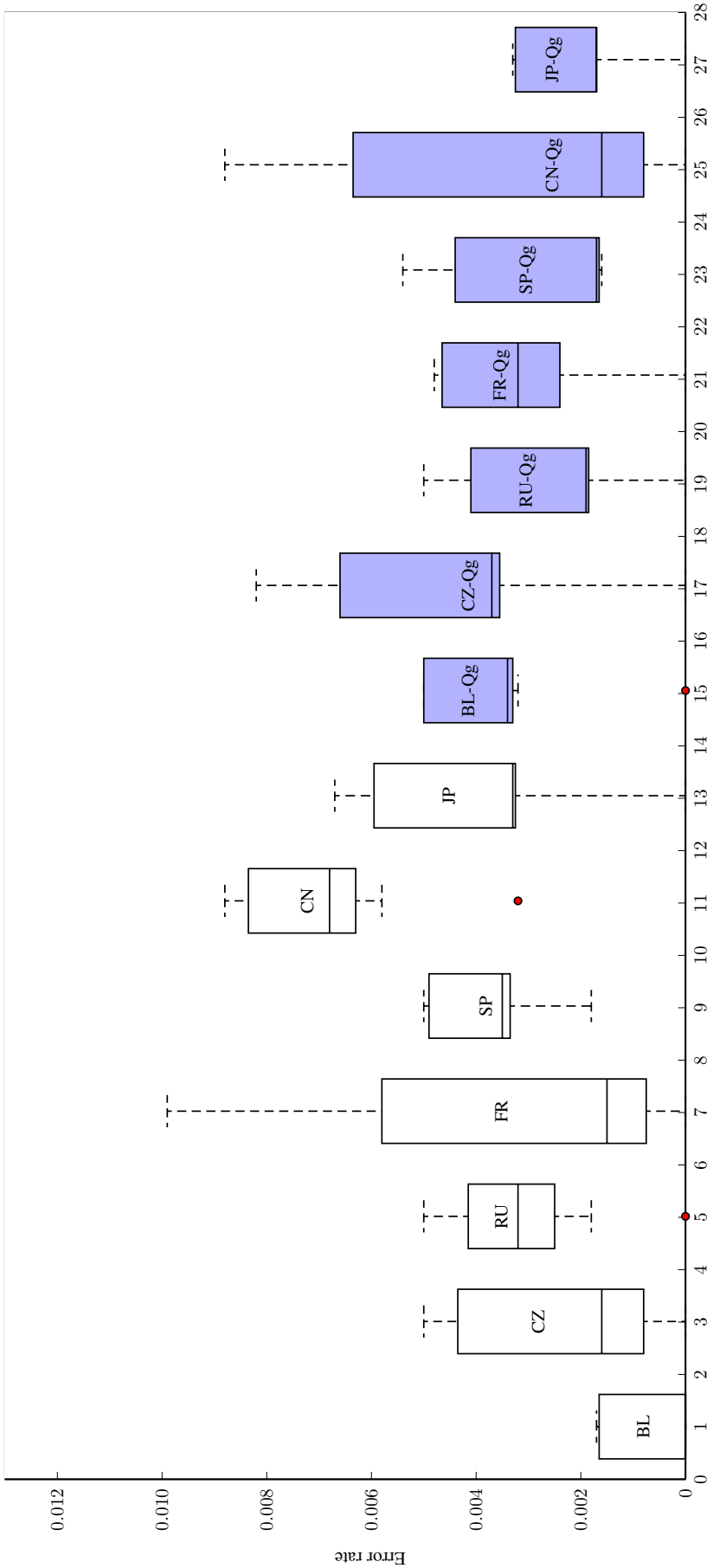


Figure 3.7: Boxplot for noun-number disagreement — actual errors vs. those detected by Queequeg (-Qg)

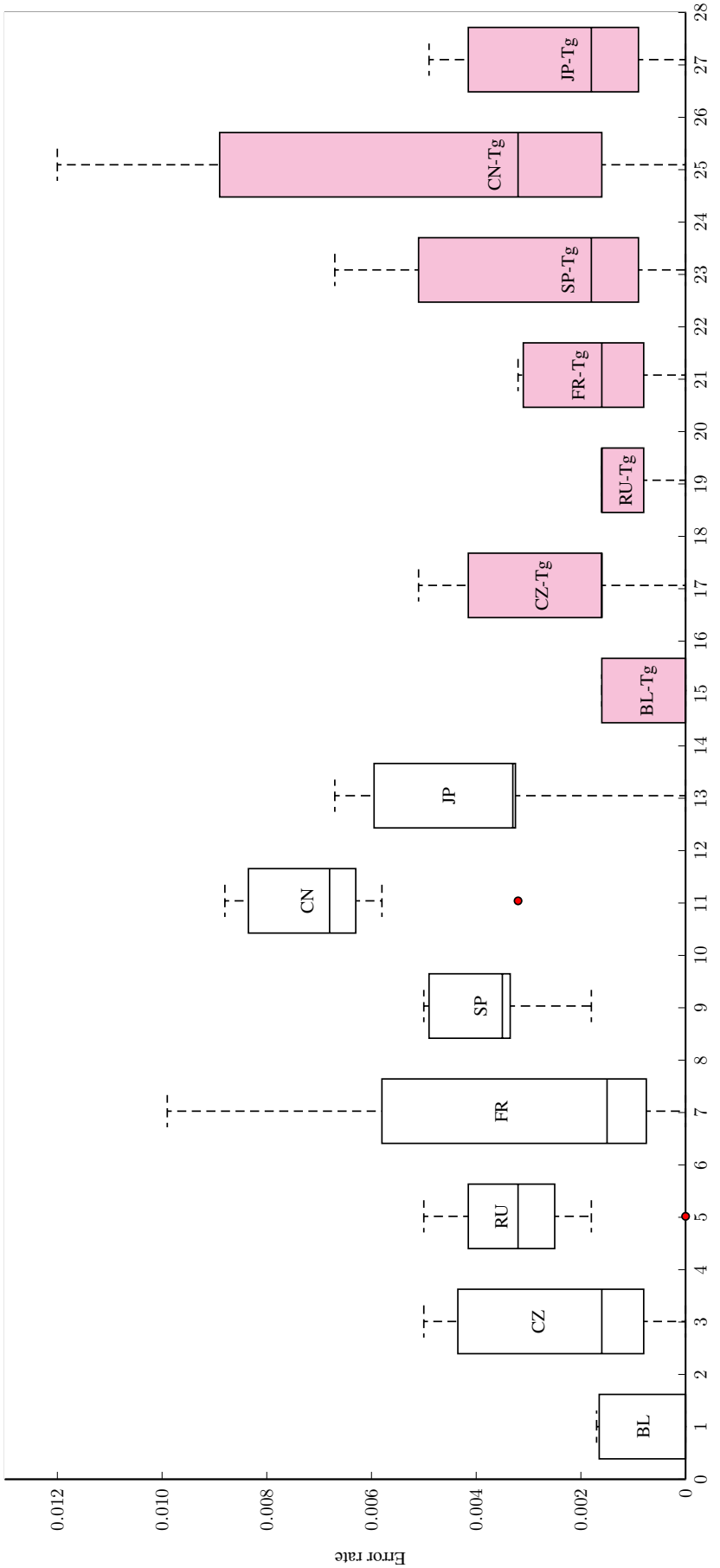


Figure 3.8: Boxplot for noun-number disagreement — actual errors vs. those detected by the retrained Brill tagger (-Tg)

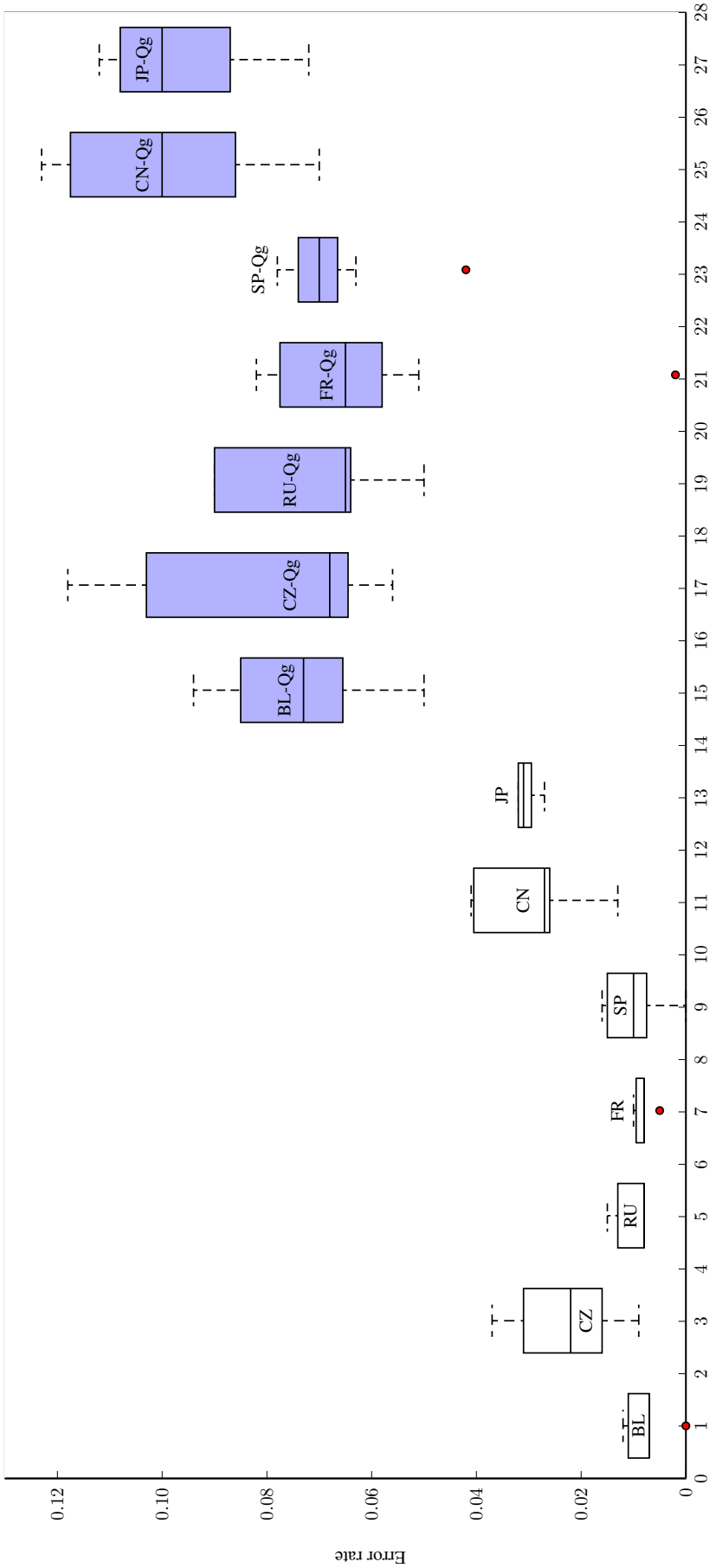


Figure 3.9: Boxplot for misuse of determiners — actual errors vs. those detected by Queequeg (-Qg)

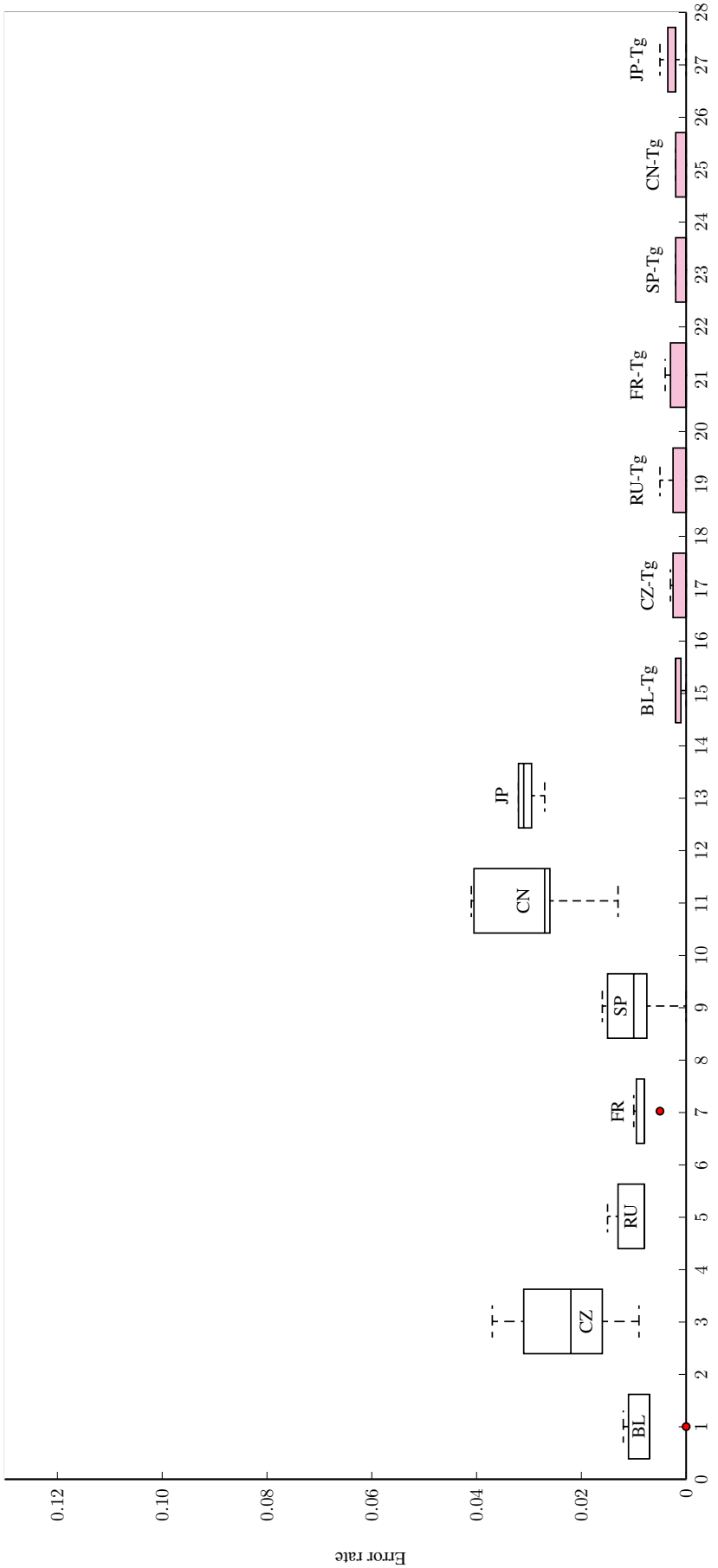


Figure 3.10: Boxplot for misuse of determiners — actual errors vs. those detected by the retrained Brill tagger (-Tg)

Table 3.4: P-values of ANOVA tests for errors detected by Queequeg

Frequency Type	Subject-verb Agreement	Noun-number Agreement	Use of Determiners
Absolute	0.038**	0.114	5.306E-10**
Relative	0.178	0.906	0.006**

**Significant at $\alpha = 0.05$

in Chinese and Japanese; whereas the usage of determiners in French and Spanish is somewhat different from (and generally more extensive than) English.

In addition, there are some modest indications of possible native language transfer in the distributions of the other two errors as well (referring to Figures 3.5 and 3.6 for subject-verb disagreement and Figures 3.7 and 3.8 for noun-number disagreement). In particular, Chinese learners stand out amongst the others with quite a high number of errors, specifically for noun-number disagreement; and this corresponds, according to the transfer effect, to the fact that these two grammatical features are in fact absent from the Chinese language system. (It should be noted that there are a couple of outliers in these distributions, as indicated by the red dots in the figures.)

Another observation that can be drawn from these boxplots is that the distribution of each error type detected by the retrained Brill tagger is generally closer to the distribution of the actual errors, as compared to those by Queequeg. It can be seen that Queequeg has a high false positive rate (as acknowledged by its creator); see Section 3.4 for more details on the evaluation of the two detection tools.

ANOVA tests. On the whole, the boxplots do not suggest an extremely non-Gaussian distribution, so we deployed ANOVA for variance analysis to verify whether the distributions do in fact differ. We note that normality in data distribution is one of the assumptions for ANOVA and other parametric tests. It had been, however, demonstrated by a number of simulation studies conducted on non-normal distributions that, ANOVA is tolerant to a moderate violation of normality, as noted in McDonald (2009) (citing Glass et al. (1972); Harwell et al. (1992); Lix et al. (1996)). In our study with the native language being the factor of interest, single-factor ANOVA tests were carried out for each syntactic error type, for both absolute frequency and relative frequency (normalised to the text length). The results are presented in Table 3.4 and Table 3.5 for both Queequeg and the retrained Brill tagger, respectively.

The most interesting result is for the case of determiner misuse. This is highly significant for both absolute and relative frequencies, be it in the results from Queequeg or from the retrained Brill tagger. This seems to be in line with our expectation as well as the explanation above.

Table 3.5: P-values of ANOVA tests for errors detected by the retrained Brill Tagger

Frequency Type	**Subject-verb Agreement**	**Noun-number Agreement**	**Use of Determiners**
Absolute	0.002**	0.453	1.449E-29**
Relative	0.005**	0.410	1.498E-31**

***Significant at* $\alpha = 0.05$

As for the subject-verb disagreement, significant differences are observed for both frequency types in the results from the tagger; as for Queequeg, on the other hand, only the absolute frequency appears to be significant. This inconsistency in the results of Queequeg could be attributed to the differences in text length. As mentioned earlier, we had additionally carried out another single-factor ANOVA test on the text length from our training data (the average values are given in Table 3.2), which confirms that the text lengths are indeed different.

Noun-number disagreement, however, does not demonstrate any significant dissimilarity among the seven groups of non-native English users (neither for the frequency types nor for the error detection tools), even though again the native languages differ in whether this phenomenon exists. Perhaps our analysis is not fine-grained enough: different L1 speakers may make different sorts of noun-number disagreement errors, but these differences do not show up in aggregate counts. Or perhaps noun-number disagreement is just not an interference (interlanguage) error (refer to Section 2.3.1.1 for discussion on different categories of learner errors); instead, it may be regarded as a developmental error according to the notion of error analysis (Corder, 1967). Developmental errors are largely due to the complexity of the grammatical system of the second language itself. This sort of error would gradually reduce or diminish as learners become more competent.

3.5.2 Learning from Syntactic Features

On the whole, the ANOVA results do demonstrate some support for the claim that non-native English users of different native language backgrounds exhibit different distributions of errors. We therefore hypothesised that with syntactic errors as features, a machine learner would be able to classify the authors (i.e. the non-native English users) according to their native language at rates above chance. The results of classification based on merely syntactic features using both absolute frequency and relative frequency are shown in Table 3.6.[10]

Given that there are seven native languages with an equal quantity of test data (25 essays per native language), the majority class baseline is 14.29%. Since there are only three syntactic error

[10]With the small amount of data used in our study, all the results presented here are after tuning with the default 5-fold cross validation on the training data (i.e. the 490 ICLE essays).

Table 3.6: Classification accuracy for syntactic error features (after tuning)

Baseline (Majority)	Queequeg		Retrained Tagger	
	Absolute Frequency	Relative Frequency	Absolute Frequency	Relative Frequency
14.29% (25/175)	20.57% (36/175)	26.86%** (47/175)	28.0% (49/175)	25.71%** (45/175)

**Significant at $\alpha = 0.05$ with respect to the baseline*

Table 3.7: Classification accuracy for combined syntactic error features from both Queequeg and the retrained tagger (after tuning)

Baseline	Absolute Frequency	Relative Frequency
14.29% (25/175)	29.14% (51/175)	32.0%** (56/175)

**Significant at $\alpha = 0.05$ with respect to the baseline*

types being examined in the current study, it is not unreasonable to expect that the accuracy would not improve to too great an extent. Nonetheless, the classification accuracies with respects to the errors detected by Queequeg and by the retrained tagger are both somewhat higher than the baseline (approximately 12% better than the baseline after tuning based on the relative frequency). It should be noted that these improvements are in effect significant at the 95% confidence level, based on the z-test of two proportions. In addition, combining the syntactic errors detected by both error detection tools leads to a classification accuracy of 32% (based on the relative frequency) as shown in Table 3.7; and this improvement over the baseline is again significant at $\alpha = 0.05$.

3.6 Learning from All Features

Turning to the second part of the current chapter, the focus is to investigate the effects of combining syntactic features with lexical features in determining the native language of the authors. To do this, we broadly replicated the work of Koppel et al. (2005) (as discussed in Section 2.2), which adopted a machine learning approach with features commonly used in authorship analysis — namely, function words, character n-grams, and PoS n-grams. Koppel et al. (2005) also utilised spelling errors as features, although this feature is not explored in our study. Spelling errors would undoubtedly further improve the overall classification performance to some extent as suggested by some of the previous studies (Oller and Ziahosseiny, 1970; Sun-Alperin and Wang, 2008); however, these would largely be tied to content words, and we do not deal with them in our work.

In what follows, each lexical feature being examined in the current study is first explained whilst the classification results for both individual features as well as the combinations of these features are presented next.

3.6.1 Lexical Features

Function words. Koppel et al. (2005) did not specify which set of function words was used, although they noted that there were 400 words in the set. Consequently, three sets of function words were explored. Firstly, a short list of 70 function words was examined; these function words were initially used by Mosteller and Wallace (1964) in their well-known work where they successfully attributed the twelve disputed Federalist papers between Madison and Hamilton (refer to Section 2.1.1 for discussion on lexical features used for authorship analysis in general). Secondly, a long list of 363 function words was adopted from Miller et al. (1958) from which the 70 function words used by Mosteller and Wallace were originally extracted. Considering that Koppel et al. (2005) made use of 400 function words, in an attempt to make up to a list of close to 400 words, we then searched for some stopwords commonly used in information retrieval — where our third list consists of 398 function words integrated with stopwords.[11] (These three sets of function words are listed in Appendix B.)

Character n-grams. Similarly, as Koppel et al. (2005) did not indicate which sort of character n-grams was used, we attempted three different types: unigram, bi-gram, and tri-gram. The 200 most frequently occurring character bi-grams and tri-grams were extracted from our ICLE training data. As for unigrams, only the 100 most frequently occurring ones were used since there were fewer than 200 unique unigrams found in the training data. Note that space and punctuation symbols were considered as tokens when forming n-grams.

PoS n-grams. In terms of PoS n-grams, Koppel et al. (2005) tested on 250 rare bi-grams extracted from the Brown corpus. As for the current study, in addition to these 250 rare Brown bi-grams, we also examined the 200 most frequently occurring PoS bi-grams taken from our training data. For the purpose of PoS tagging, we again used the Brill tagger which performs at approximately 93% accuracy. (Note that the tagger used in this part was trained on the original Brown corpus without modification unlike when it was trained for error detection as mentioned in Section 3.4.)

[11] Stopwords were obtained from Onix Text Retrieval Toolkit available at `http://www.lextek.com/manuals/onix/stopwords1.html` in June 2009. Note that the revised list is now with 429 words as opposed to 398 words as of the time of our experiments.

Table 3.8: Classification accuracy (%) for combining function words with syntactic errors (after tuning)

Function Words	Absolute Frequency – Errors	Absolute Frequency + Errors	Relative Frequency – Errors	Relative Frequency + Errors
				Queequeg
70 words	52.00	50.29	60.57	60.57
363 words	59.43	58.86	65.14	*68.00*
398 words	61.14	61.14	62.86	62.86
				Retrained Tagger
70 words	52.00	50.86	60.57	62.86
363 words	59.43	60.00	65.14	66.29
398 words	61.14	63.43	62.86	*69.14*
				Both error detectors
70 words	52.00	50.86	60.57	65.14
363 words	59.43	58.86	65.14	67.43
398 words	61.14	61.14	62.86	***69.71***

Table 3.9: Classification accuracy (%) for combining character n-grams with syntactic errors (after tuning)

Character N-grams	Absolute Frequency – Errors	Absolute Frequency + Errors	Relative Frequency – Errors	Relative Frequency + Errors
				Queequeg
Unigram	57.14	61.14	61.71	64.00
Bigram	31.43	29.71	69.14	70.29
Trigram	33.71	32.57	71.43	*70.86*
				Retrained Tagger
Unigram	57.14	57.71	61.71	65.14
Bigram	31.43	32.57	69.14	70.85
Trigram	33.71	31.43	71.43	***73.14***
				Both error detectors
Unigram	57.14	60.57	61.71	63.43
Bigram	31.43	32.00	69.14	70.29
Trigram	33.71	32.57	71.43	*72.00*

For each of the lexical features, four sets of classification were performed. The data was examined without normalising, with normalising to lowercase, according to their absolute frequency, as well as their relative frequency (per text length). Only the results without normalising are presented since the classification results with and without normalising to lowercase do not differ much.

3.6.2 Classification Results

Individual results. The classification results (after tuning) for each lexical feature — function words, character n-grams, and PoS n-grams — are presented in Tables 3.8, 3.9 and 3.10, re-

Table 3.10: Classification accuracy (%) for combining PoS n-grams with syntactic errors (after tuning)

PoS N-grams	**Absolute Frequency**		**Relative Frequency**	
	– Errors	**+ Errors**	**– Errors**	**+ Errors**
				Queequeg
Bigram	66.29	66.29	74.28	***78.28***
Trigram	58.86	58.86	62.29	64.00
				Retrained Tagger
Bigram	66.29	62.86	74.28	*77.14*
Trigram	58.86	58.86	62.29	62.86
				Both error detectors
Bigram	66.29	62.86	74.28	*77.71*
Trigram	58.86	58.28	62.29	64.00

spectively. Each table contains results with and without integrating with syntactic features (i.e. the three identified syntactic error types) and there are three sub-tables for each lexical feature representing errors detected from Queequeg itself, from the retrained tagger itself, and from the combination of both. (The best result with each sub-table is italicised and the overall best result among the three is in bold.)

It was found that the classification performance for each lexical feature seems to be better prior to tuning when the absolute frequency of the features is considered. However, using features based on their relative frequency leads to better classification performance after tuning; hence, only the tuned results are presented. These accuracies are in fact much higher — between 5% and 10% improvement over the accuracies achieved based on the absolute frequencies. (There is an exception, in which both character bi-grams and tri-grams are observed to perform much better in their relative frequency in general.)

When integrating with syntactic features, the effect on the classification performance based on the absolute frequency appears to be inconsistent in the sense that adding syntactic errors sometimes improves the classification but at other times the performance decreases. On the other hand, the performance seems to be consistent when the relative frequency is utilised for classification. For the most part, after tuning, adding syntactic errors maintains or slightly improves the classification performance for each lexical feature before combining with any syntactic features (with respect to relative frequency).

Comparing Queequeg and the retrained tagger, the overall picture is that integrating errors detected by the retrained tagger improves the classification for each of the lexical features; but errors from Queequeg sometimes leads to a slight decline in performance (especially when integrating with character tri-grams.) In brief, the best performance for function words, character n-grams, and PoS n-grams with the integration of the syntactic errors detected by both Queequeg and the

retrained tagger are 69.71% (398 function words), 72.0% (character tri-grams), and 77.71% (PoS bi-grams), respectively. However, it should be noted that the overall best result for character n-grams comes from the retrained tagger (73.14% for the tri-grams); as for PoS n-grams, the overall best result is from Queequeg (78.28% for the bi-grams).

As a side note, the classification performance prior tuning of character n-grams — bi-grams and tri-grams (Table 3.9) — is much weaker than the accuracies observed for function words and PoS n-grams as the classification features. It, however, achieves more than 70% after tuning which seems to be on par with the individual best performance of the other two lexical features. Also note that the classification results for the 250 rare bi-grams from the Brown corpus are not presented here since the results are all at around the baseline (14.29%). This is no doubt due to feature sparsity.

Combined results. Table 3.11 presents the classification results of all combinations of lexical features (with and without integrating with syntactic errors) after tuning. We selected the longest list of function words (398 words), character bi-grams, and PoS bi-grams were selected for combination. Although character tri-grams seem to perform better than character bi-grams, combining the former with the rest of the features (including syntactic features) results in weaker classification performance than the integration of the latter with other features. A similar behaviour is observed for function words, where the longest list with stopwords leads to a better overall performance for the combination of all features than the set with solely 363 function words.

The combination of all three lexical features generally results in higher classification accuracy after tuning than any of the combinations of two features — between 75% and 80% accuracies. Some observations from the combined results are that integrating syntactic errors in general does not seem to improve the classification accuracy after tuning; however, adding errors from Queequeg sometimes causes a slight decline in performance with respect to relative frequency — this is observed in the combination of function words with character n-grams (refer to the first section of Table 3.11). Conversely, with the integration of errors from the retrained tagger somehow improves the classification, in particular when character n-grams and PoS n-grams are combined (refer to the second section of Table 3.11).

In summary, our overall best accuracy after tuning thus far is at 81.14% (with 142 out of the 175 essays correctly classified to their respective native language.) The confusion matrix presented in Table 3.12 provides an overall picture illustrating where most of the misclassifications occurred; they largely occur in two of the Slavic languages — Bulgarian and Russian — as well as Spanish. Furthermore, the classifier is able to identify the two Asian (Oriental) languages — Chinese and Japanese — with close to 100% accuracy which is perhaps not surprising, given that

Table 3.11: Classification accuracy (%) for combining lexical features with syntactic errors (after tuning)

Feature Combination	Absolute Frequency† – Errors	Absolute Frequency† + Errors	Relative Frequency – Errors	Relative Frequency + Errors
				Queequeg
Function words+Character n-grams	66.29	66.29	77.71	75.43
Function words+PoS n-grams	73.14	73.14	79.43	79.43
Character n-grams+PoS n-grams	66.29	66.29	78.86	78.86
All three lexical types	74.86	74.86	***81.14***	***81.14***
				Retrained Tagger
Function words+Character n-grams	66.29	66.86	77.71	77.71
Function words+PoS n-grams	73.14	73.14	79.43	79.43
Character n-grams+PoS n-grams	66.29	69.14	78.86	80.00
All three lexical types	74.86	74.86	***81.14***	***81.14***
				Both error detectors
Function words+Character n-grams	66.29	66.86	77.71	76.00
Function words+PoS n-grams	73.14	73.71	79.43	78.29
Character n-grams+PoS n-grams	66.29	66.29	78.86	79.43
All three lexical types	74.86	74.86	***81.14***	***81.14***

†*Character n-grams are in their relative frequency*

Table 3.12: Confusion matrix for the best classification accuracy based on the combination of both lexical and syntactic features

	BL	**CZ**	**RU**	**FR**	**SP**	**CN**	**JP**
BL	[14]	3	3	4	1	–	–
CZ	2	[21]	–	2	–	–	–
RU	3	2	[16]	4	–	–	–
FR	1	–	–	[24]	–	–	–
SP	2	1	1	3	[18]	–	–
CN	–	–	–	–	–	[25]	–
JP	–	–	–	1	–	–	[24]

BL:Bulgarian; CZ: Czech; RU:Russian; FR:French; SP:Spanish; CN:Chinese; JP:Japanese

these language systems differ quite significantly from English, in particular with respect to the grammatical features investigated in the current study.

3.7 Discussion

Comparison with Koppel et al. (2005). The classification results prior to tuning for both function words and character n-grams (without taking syntactic features into consideration) are lower than the results obtained by Koppel et al. (2005). Nonetheless, character n-grams (both bi-grams

Table 3.13: Comparison of classification accuracy with Koppel et al. (2005)

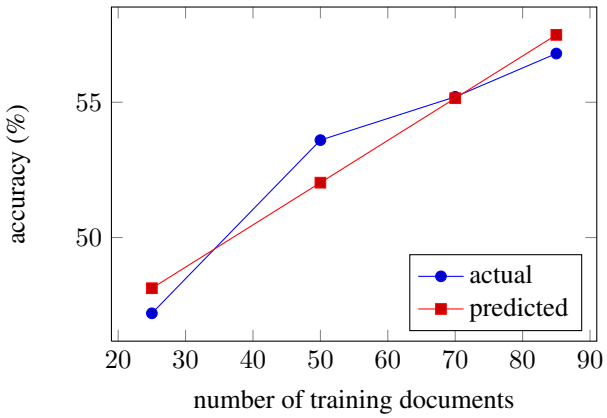
Feature Type	Koppel et al. (2005)	Our Best Result
Function words	71.0%	65.14%
Character n-grams	68.0%	71.43%
All lexical features	80.2%	81.14%

Note: Our best results presented here are based on 363 function words and character trigrams since they are the individual best performance (after tuning)

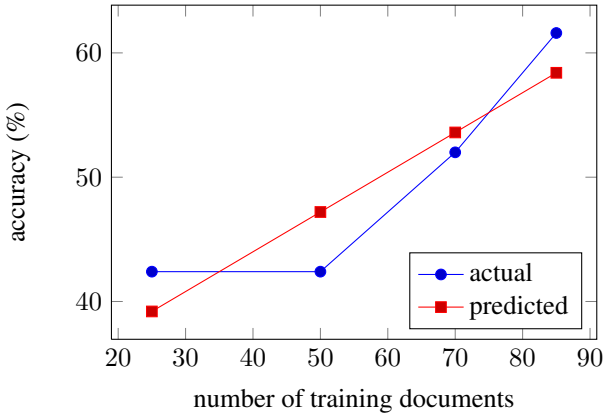
and tri-grams) perform on par with Koppel et al. after tuning, as presented in Table 3.13. The observed difference in classification accuracy, for function words in particular, can basically be explained by the corpus size. As mentioned earlier, we adopted smaller subsets of the ICLE corpus for the current study — only 95 essays were used for each native language (i.e. 70 for training and 25 for testing). In contrast, Koppel et al. (2005) made use of 258 essays for each native language. A simple analysis (looking at a curve of the results for variously sized subsets of our data, and the improvements as the training sample size grows — see Figure 3.11) suggests that our results are indeed consistent with those of Koppel et al. given the sample size. (In order to build up bigger training subsets up to size 85, we made use of the additional 15 essays set aside for each native language group, as mentioned at earlier the chapter under Section 3.2.)

Moreover, it is worth noting that our task is indeed to some extent more difficulty than Koppel et al. (2005) since we examined two additional languages. Overall, we managed to achieve the same level of classification performance as Koppel et al. (2005) for the combination of all features — comparing 81.1% with 80.2%. (Note that the results of PoS n-grams could not be compared here since Koppel et al. (2005) considered these features as part of the errors and did not provide a separate classification result.)

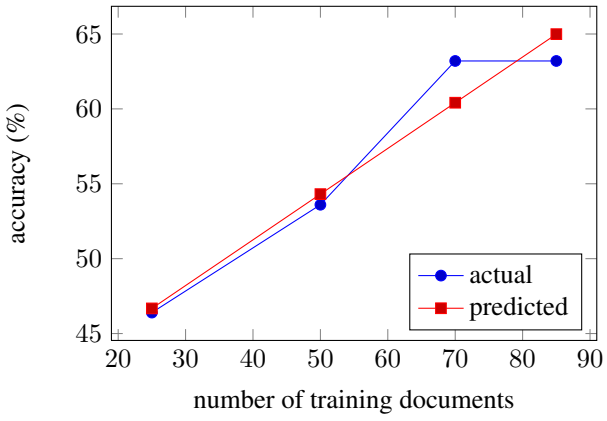
Usefulness of syntactic features. For the best combination of features, our classification results of integrating with syntactic features (i.e. syntactic errors) with the lexical features do not demonstrate any improvement in terms of classification accuracy. In fact, adding syntactic errors decreases the classification performance sometimes. This could possibly be due to the small number of syntactic error types being considered at this stage. Such a small number of features (three in our case) would not be sufficient to add much to the approximately 800 features used in our replication of the work by Koppel et al. (2005). Furthermore, it cannot be ruled out that there might be flaws in error detection as a result of the limitations of the error detection tools.



(a) Function words



(b) Character bi-grams



(c) PoS bi-grams

Figure 3.11: Extrapolation of classification accuracy with respect to training data size
(Note: These results are based on only the five native languages as used by Koppel et al. (2005) for comparison purposes)

Other issues of note. As seen in the individual results of character n-grams (Table 3.9), character n-grams alone are able to achieve a relatively reasonable accuracy of approximately 70%. It can be regarded that character n-grams are to some extent a useful marker and as argued by Koppel et al. (2005) that such features may reflect the orthographic conventions of individual native language. Furthermore, this is also consistent with the hypothesis put forward by Tsur and Rappoport (2007) in their replication of Koppel et al.'s work. It was claimed that the choice of words in second language writing is highly influenced by the frequency of native language syllables (i.e. the phonology of the native language) which can be captured by character n-grams. Such a phenomenon can be observed in Japanese learners of English, where it is well known that they are particularly confused between the English phonemes /l/ and /r/ as the Japanese language does not make such a distinction; and they often assimilate both the phonemes to the Japanese /r/ (Aoyama et al., 2004).

Recent related work. As mentioned in Section 2.2, a number of related works have extended our studies presented in this thesis, in particular the current chapter and the next. Here, one specific piece of work that should be considered is that of Kochmar (2011) which is based on our initial work in this chapter and published in Wong and Dras (2009), where part of her work is aimed at improving upon our approach of using only three syntactic error types.

Kochmar (2011) investigated about 80 types of error identified in the Cambridge Learner Corpus (CLC), and discovered that using the top 10 error types selected through information gain (IG) results led to better performance in most of her binary classification tasks for distinguishing between a pair of closely related languages from the Germanic and Romance language groups. It was found that the three error types considered in our work (Wong and Dras, 2009) were not among the typical error types detected for all the Indo-European languages examined in her work. (These error types are presented in Kochmar (2011) under Table 3.1.) Consequently, the classification results based on just three error types (that we used) only resulted in the classification accuracy slightly above chance (50-60%); while the IG selected error types perform significantly better at the accuracy range of 60-80%. Further, as our results show no significant improvement when integrating errors with widely adopted lexical features, Kochmar (2011)'s results demonstrate that adding her error types with lexical features (and other distributional features) led to some degree of improvement but not higher than 2.5% in general.

It should be noted that there are a number of differences between the work of Kochmar (2011) and ours. In our studies (throughout the entire thesis), we are working on a multiclass classification task involving seven native language groups; while all the classification tasks in Kochmar (2011) are binary ones for pairs of languages from 5 Germanic languages and 5 Romance languages.

With regard to the native languages investigated, we consider languages that are quite different from English in terms of grammatical features (Slavic and Asian in particular), but Kochmar's languages are Germanic and Romance which are quite similar to English. Another aspect that is worth mentioning is the difference in the corpus used (ICLE versus CLC). All these factors may account for some degree of variation in terms of classification performance.

3.8 Concluding Remarks

For the most part, we have found some modest support for the hypothesis that the idea of interference errors from contrastive analysis can help in detecting the native language of a text's author, through a statistical analysis of three syntactic error types and through machine learning using only features based on those error types. Nevertheless, in combining these with features commonly used in other machine learning approaches to the task of native language identification, we did not manage to find an improvement in classification accuracy, only performance at the comparable level of just about 80%.

An examination of the results suggests that using more error types (such as inappropriate use of prepositions and incorrect use of word order), as well as a method for more accurately identifying them (in view of the limitations of the error detection tools used in the current study), might result in classification improvements. A still more useful approach might be to use an automatic means to detect different types of syntactic errors; for example, the idea suggested by Gamon (2004) as inspired by Baayen et al. (1996), in which context-free grammar production rules can be exploited to detect ungrammatical structures. Syntactic parsing would be helpful in this regard. With the help of statistical parsers, parse production rules can be extracted and the resultant different distributions of production rules could serve as useful markers. In adddition, such an approach could also uncover overuse and avoidance of certain syntactic structures that might be the effects of cross-linguistic influence (as seen in Section 2.3.1.2) as these would manifest in the extreme occurrences of certain production rules. We proceed to this second approach in the next chapter.

Chapter 4

Using Statistical Parsing for Syntactic Structures

From Chapter 3, there has been some evidence that error-based differences among texts written in a second language (English) by writers with different native language backgrounds are indeed detectable to some extent through a simple approach of pattern-matching over PoS tags and words. However, the use of these error-based distributions as markers for identifying native language did not lead to an improved classification performance over just the lexical ones. Likely reasons are that we had only explored a relatively small number of error types and perhaps the detected syntactic patterns are in fact inaccurate.

It appears to us that there are two possible directions to pursue: the first is to explore a wider range of error types with more sophisticated error detection techniques; alternatively, syntactic structures themselves can be used as features through deploying statistical parsers with high accuracy. We opt for the latter direction since statistical parsing can be exploited to uncover other phenomena influenced by native language, such as overuse and avoidance of certain syntactic constructions in second language production (to be discussed later in the chapter), in addition to syntactic errors. For the purpose of native language identification (NLI), the goal of this chapter is to investigate how computational linguistic tools — specifically automatic statistical parsers — might detect such syntactic manifestations of native language and use them as features for determining the native language of the writer of a particular text.[1]

We tackle this in two parts. First, it is not clear whether statistical parsers are in fact even useful in characterising syntactic errors at all. In addressing the problem of *sentence grammaticality judgement* as a binary classification task — determining whether a sentence is grammatical or

[1] The core work of this chapter was published in Wong and Dras (2011).

not, that is, whether it contains grammatical errors — by means of supervised machine learning, a number of existing studies have attempted to exploit some form of parser outputs and 'by-products' as features for classification. These include (log) probability of a parse tree, number of partial (incomplete) parse trees, parsing duration, and such (Mutton et al., 2007; Sun et al., 2007b; Foster et al., 2008; Wagner et al., 2009). What we are interested in is the primary parser output, the syntactic tree or parts thereof. As the first task, then, we look at whether features derived from syntactic trees are useful in sentence grammaticality judgement, similar to Sun et al. (2007a) which explored the 'direct' parse trees, in next section; we, in particular, look at horizontal tree slices (i.e. production rules) and at instantiations of more complex templates over trees as used in parse reranking and elsewhere.

Subsequently, we incorporate such syntactic features as they prove to be useful in sentence grammaticality judgement into the NLI task, which we approach using the standard machine learning paradigm. In our earlier work in Wong and Dras (2009) and also (with some more detail) in Chapter 3, we looked at identifying specific errors known from the SLA literature, for example article errors, and using those as features. Here, there is no presupposition about what the errors might be: this then has the potential to discover previously unknown errors, and also to identify when particular constructions are grammatical in the L1 but are overused or avoided in a manner characteristic of a particular L1.[2]

The remaining structure of this chapter is as follows. In Section 4.1, we introduce the key concepts as well as review the related work on sentence grammaticality judgement (the review of relevant NLI work being in Section 2.2). The subsequent two sections detail the classification experiments for each of the two key tasks of this chapter, exploiting parse structures as classification features: Section 4.2 for sentence grammaticality judgement (as a binary classification task) and Section 4.3 for native language identification (as a multiclass classification task).

4.1 Related Work: Sentence Grammaticality Judgement

Research on identifying specific ungrammaticalities, such as articles, prepositions, as well as collocations, has been abundant; and an attempt to give an exhaustive review on these would go beyond the scope of this chapter as well as the thesis. An extensive list of research targeting specific grammatical errors can be found in Leacock et al. (2010). Here we note some prominent examples of work identifying ungrammaticality at the sentence level (i.e. judging whether a sentence is well-formed or not) using parser outputs, including parse trees and parser by-products.

[2] Alternatively, generative grammars, such as the English Resource Grammar (Copestake and Flickinger, 2000) and the ParGram English grammar (Butt et al., 2002), could be an alternative; grammaticality prediction is also realised as a first-order output by these grammars.

An early work using parser outputs is Mutton et al. (2007), who found that outputs from multiple parsers, such as log probability of the most likely parse, number of partial parse trees, and number of invalid parses correlate well with sentence 'fluency'; and that the combination of multiple parser outputs outperformed individual parser metrics. Likewise, Andersen (2007) used grammatical relations obtained from the RASP parser (Briscoe et al., 2006) in addition to n-grams and PoS tags to detect grammatical errors.

N-grams have been explored in other studies as well. Okanohara and Tsujii (2007) used n-grams clustered by a semi-Markov class model as features in grammaticality discrimination; the highest classification accuracy rate reported was 74% for data drawn from the British National Corpus. Sun et al. (2007b) suggested that non-contiguous n-gram-like features which they termed as 'labelled sequential patterns' are more important than parse probabilities in terms of sentence grammaticality classification, achieving an accuracy rate of over 80% on their classification task. In their subsequent work (Sun et al., 2007a), they deployed the full parse trees (also taking into account subsumption relationships between trees) as features, and achieved a similar accuracy rate of up to 80% when combining both the 'labelled tree patterns' and the 'labelled sequential patterns'.

The work presented in this chapter is most similar to the line of work started by Wagner et al. (2009) which based on their earlier work (Wagner et al., 2007). They looked at two approaches to judging sentence grammaticality: one using the precision parser XLE (Maxwell and Kaplan, 1996) with the ParGram English grammar (Butt et al., 2002), a manually constructed grammar in the Lexical Functional Grammar (LFG) formalism; and the other using the statistical reranking parser by Charniak and Johnson (2005) (which we henceforth refer to this throughout the chapter as the C&J parser). In the former, they used n-grams over part-of-speech (PoS) tags, and six features of XLE including a flag indicating use of robustness techniques which indicate that the parser initially failed to parse. In the latter, they used log parse probabilities and "structural differences between the parse trees measured using various parser evaluation metrics".[3] Three variants of the C&J parser were produced, by training on grammatical treebank, a distorted ungrammatical treebank, and the union of these, following the approaches proposed in Foster et al. (2008).[4] The classification accuracy rates reported were within the range of 65–75%, the best result obtained by combining all the feature sets.

Another related work of ours, Wong and Dras (2010), followed a similar approach, training a statistical parser on the same sets of grammatical and ungrammatical treebanks, but deploying the

[3]These were not explicitly described, but were given by citation of Abney et al. (1991) and Sampson and Babarczy (2003).

[4]The methodology used in Foster et al. (2008) for constructing an 'ungrammatical' parser through training with erroneous treebank is discussed in the next section (Section 4.2) as we are deploying the similar methodology for our experiments on judging sentence grammaticality.

Stanford parser (Klein and Manning, 2003) instead of the C&J parser. The idea was to exploit the parse structure itself in the form of CFG production rules to improve grammaticality judgement (as a form of classification task at the sentence level). Compared to a baseline using parse probabilities, we found that considering the trees as sets of context-free production rules (e.g. `S` $\rightarrow$ `NP VP`), and then adding in the production rules as features, resulted in a modest 1.6% improvement overall. Sparsity of training vectors was proposed as a reason for the smallness of the improvement in performance.

Our intuition in this chapter is that the production rules used in our previous work (Wong and Dras, 2010), as a form of representation for syntactic structures, are just a small subset as compared to the feature schemas used in discriminative parse reranking (Charniak and Johnson, 2005). In principle, utilising these templates as classification features for sentence grammaticality judgement could capture much more structure and better characterise ungrammmatical structures, as well as alleviate feature sparsity problems. In this first task of the chapter, we approach the sentence grammaticality judgement task using this broader set of syntactic features.

Finally on sentence grammaticality judgement, we note that there has been recent work by Post (2011), deploying tree fragments from Tree Substitution Grammar (TSG) derivations as features in addition to parse reranking features along with other features for the same classification task: in this it is more similar to our work than those discussed above. The Post (2011) work starts from the BLLIP'99 dataset in contrast to our Wall Street Journal (WSJ) treebank, which constitutes the positive (grammatical) examples. The negative (ungrammatical) examples are, however, less realistic 'pseudo-negative' sentences generated from a trigram language model. The reported classification results based on TSG tree fragments and reranking features are higher than the previous studies, with accuracy in the range 88–100%, with each of them outperforming one another in each dataset (reranking features performed best on the BLLIP data). For the purpose of the work presented in this chapter, we are following Wagner et al. (2009) and Wong and Dras (2010) to examine more 'realistic' ungrammatical data created through modeling real learner errors, and whether parsers are useful in discovering these; with respect to classification accuracy, then, we expect these to be lower compared to Post's.

4.2 Sentence Grammaticality Judgement

Before tackling NLI, the goal of this section (as the first task of this chapter) is to investigate the extent to which parse structures themselves can be deployed for characterising sentence grammaticality in general. We follow the general approach taken by Foster et al. (2008), and our previous work, Wong and Dras (2010). The key idea there is that having a parser that is trained on both

grammatical and ungrammatical data may help in detecting ungrammatical sentences: an 'ungrammatical' parser may, for example, assign higher probabilities to parses of ungrammatical sentences than a parser trained on regular grammatical sentences would, making this a useful feature.

In general, treebanks of ungrammatical data do not exist, however. Foster (2007) therefore proposed an approach for constructing ungrammatical treebanks based on an extensive modelling of errors in real corpora presented in her thesis (Foster, 2005). Ungrammatical sentences are constructed by introducing errors into the original (grammatical) WSJ sentences through the operations of word insertion, substitution, and deletion. Each ungrammatical sentence is then tagged with the gold standard parse tree, a transformation of the original parse tree of its grammatical counterpart with the intended meaning remained intact. The types of errors introduced include missing word, extra word, real-word spelling, agreement, and verb form: according to Foster (2005), these comprise 72% of the analysed errors. Foster et al. (2008) subsequently showed that a combination of regular parsers and ungrammatical parsers was better at detecting ungrammatical sentences than a regular parser alone. Specifically, they took as features the log probabilities of the most likely parse returned by a regular parser and an ungrammatical parser, and used these as features in a classifier.

4.2.1 Experimental Setup

As per this approach, then, the experiments here are carried out in two stages. In the first stage, we induce a parsing model that can (preferably) parse both grammatical and ungrammatical sentences equally well, as in Foster et al. (2008). This robust parsing model is then used for all the subsequent parsing tasks in the second stage, where we take as a baseline the parse probability approach derived from Foster et al. (2008), and then examine the addition of parse production rules and instantiated tree templates (discussed in Section 4.2.2).

4.2.1.1 Scenarios

We investigate two scenarios here. First, we assume an 'ideal' scenario where a gold standard treebank for ungrammatical data is available for training a parser; the parser hence has some insight into how ungrammatical sentences should be parsed. We also investigate a more 'realistic' second scenario, where the gold standard ungrammatical treebank is not available for parser training. Instead, we assume we only have access to a regular parser trained on a grammatical treebank. By assessing these two scenarios, this allows us to gauge the extent to which the use of a gold standard ungrammatical treebank is indeed important for training a parser and subsequently to study its effects on judging the grammatical status of sentences.

4.2.1.2 Corpora

Two types of corpus are needed for the purpose of the present study. For the grammatical sentences, we take the original Wall Street Journal (WSJ) treebank (Marcus et al., 1993) with the assumption that these sentences are grammatically well-formed. We refer to this corpus as *PureWSJ*. The ungrammatical sentences are obtained from the noisy (distorted) versions of WSJ created by Foster (2007).[5] There are two sets of the noisy version of WSJ, which we refer to as *NoisyWSJ* and *NoisierWSJ*; the second set contains two or more errors in each sentence while the first set contains only one error. These grammatically ill-formed WSJ sentences were generated by introducing grammatical errors to the initially well-formed WSJ sentences.

4.2.1.3 Parsers

We employ two PCFG-based statistical parsers: the Stanford parser (Klein and Manning, 2003) and the C&J parser (Charniak and Johnson, 2005). The Stanford parser constructs a generative parsing model within a factored lexicalised parsing framework that uses an unlexicalised PCFG model with a separate lexical dependency model. With carefully defined state splits (parent tree annotation) and other grammar transformations such as markovization, Klein and Manning (2003) demonstrate that the Stanford parser can parse quite accurately compared to early lexicalised models, with a reported f-score of 86.36%.

The C&J parser, on the other hand, adopts a discriminative model which uses a regularised maximum entropy or log-linear model to perform discriminatve reranking on the n-best parses returned by another generative parsing model, the Charniak parser (Charniak, 2000). The n-best parses are produced through heuristic coarse-to-fine refinement of parsing, resulting in high quality parse candidates. The reranking process then selects the best parse that is associated with the highest score of a linear function of a set of linguistically informed reranking features. As a result of reranking, the C&J parser at the time achieved state-of-the-art parsing performance with a reported f-score of 91.0%.

4.2.2 First Stage: Training of Parsers

To enable it to be capable of accurately parsing both grammatical and ungrammatical sentences, a parser ideally has to be re-trained on both grammatical and ungrammatical corpora. Following Foster et al. (2008), we conduct five experiments to re-train the two parsers — the C&J parser

[5]For more detail on how grammatical errors were inserted, the reader can turn to Foster (2007). A special thanks to Jennifer Foster for providing us with the ungrammatical versions of the WSJ treebank; and for her helpful comments on the early draft of this chapter.

(Charniak and Johnson, 2005) as per Foster et al. (2008)[6], and the Stanford parser (Klein and Manning, 2003) as per Wong and Dras (2010) — in order to induce a more robust parsing model capable of parsing both grammatical and ungrammatical sentences. In this first part, we compare the parsing performance of the two parsers, and subsequently their influences on the classification tasks of judging sentence grammaticality.

In the first three experiments, three models for each of the two parsers are induced by training on three different sets of corpora — first on the original WSJ (*PureWSJ*); second on the noisy WSJ (*NoisyWSJ*); and third on the combination of the first two (i.e. *PureWSJ* plus *NoisyWSJ*). We denote these three parsing models as *PureParser*, *NoisyParser*, and *MixedParser*, respectively. To gauge its ability to parse both grammatical and ungrammatical sentences, each of these models is evaluated against the three WSJ test sets (i.e. *PureWSJ*, *NoisyWSJ*, and *NoisierWSJ*) based on the labelled f-score measure.

For the last two experiments we train a 'combination' parser, where each sentence is parsed with two of the abovementioned parsers — one trained exclusively on grammatical data (*PureParser*) and the other trained on some ungrammatical data (either *NoisyParser* or *MixedParser*). The best parse for a sentence is selected based on the one associated with the higher parse probability (Foster et al., 2008). Hence, *PureParser* is integrated with *NoisyParser* for the fourth experiment and with *MixedParser* for the last experiment. All training is performed on Sections 02-21; all testing is on Section 00.

4.2.3 Second Stage: Classification of Sentences

For the main task of grammaticality judgement, we deploy the parser outputs described below as features for classifying sentences, as either grammatical or ungrammatical. Here, we are interested in investigating the degree of influence of the two parsers — Stanford parser and C&J parser — on the overall classification performance, through examining the parse production rules extracted from the parses produced by each of them.

As mentioned above, we intend to evaluate two 'scenarios' as described at the beginning of this section: one where an ungrammatical treebank is available for parser training, in which we have two instances of a parser trained (as in the first stage) on both grammatical and ungrammatical data — *PureParser* with either *NoisyParser* or *MixedParser* (depending on the outcomes from the first-stage experiments); and the more realistic scenario where there is only one instance of the parser being trained on only grammatical data — *PureParser*. To be precise, there will be two sets of features (each coming from any two instances of the parser trained) for the ideal first scenario,

[6]Note that Foster et al. (2008) only retrained the first-stage generative parser of the C&J parser (i.e. the Charniak parser), while we also retrained the reranker at the second stage with assistance from Mark Johnson.

but only one set of features for the realistic second scenario.

4.2.3.1 Classification Features

Parse probability features. Parse probabilities, serving as a baseline, are obtained by conducting similar procedures as in the last two experiments in the first stage. As before, each sentence is parsed with two types of parsing model — *PureParser* and either *NoisyParser* or *MixedParser* (depending on the first-stage outcomes). The parse probability returned by each parsing model is used as a classification feature. With the Stanford parser, there are only two feature values: the parse probability from *PureParser* and the parse probability from either *NoisyParser* or *MixedParser*. For the C&J parser, on the other hand, there are two additional feature values which are the reranking scores as mentioned in Section 4.2.1.3 — i.e. the best parse returned by the C&J is the highest scoring parse amongst the other n-best parse candidates. (Note that this description is specifically for the first scenario mentioned above; as for the second scenario, only *PureParser* is used — hence, there is only a single set of parse probabilities serving as feature values.)

Parse production rule features. The parse production rule features are selected using the two feature selection metrics that we found to perform best in Wong and Dras (2010): Information Gain (IG) (Yang and Pedersen, 1997) and Binormal Separation (BNS) (Forman, 2003). The details of each metric is described as follows (with r representing a parse production rule and c a class, i.e. grammatical or ungrammatical):

- *Information Gain* (IG): The formula below is adopted from Yang and Pedersen (1997), with $m = 2$ (i.e. the number of distinct classes). We pick the 100, 500, and 1000 rules with the highest information gain as classification features. Feature values are of binary type.

$$\begin{aligned} IG(r) = & -\sum_{i=1}^{m} \Pr(c_i) \log \Pr(c_i) \\ & + \Pr(r) \sum_{i=1}^{m} \Pr(c_i|r) \log \Pr(c_i|r) \\ & + \Pr(\bar{r}) \sum_{i=1}^{m} \Pr(c_i|\bar{r}) \log \Pr(c_i|\bar{r}) \end{aligned} \tag{4.1}$$

- *Binormal Separation* (BNS): Forman (2003) defines this metric as below, where *F(x)* is the cumulative probability function of a normal distribution. Similarly, the 100, 500, and 1000 rules with the highest BNS scores are selected as features with binary-typed values.

$$BNS(r, c) = \left| F^{-1}(\Pr(r|c)) - F^{-1}(\Pr(r|\bar{c})) \right| \tag{4.2}$$

Reranking features. To resolve the probable feature vector sparsity problem that might arise from the use of just horizontal parse production rules, we explore more generalised parse structures — features used for discriminative reranking — that might capture much more aspects of ungrammaticality. We use the 13 feature schemas as described in Charniak and Johnson (2005), which were inspired by earlier work in discriminative estimation techniques, such as Johnson et al. (1999) and Collins (2000). Examples of these feature schemas include tuples covering head-to-head dependencies, pre-terminals together with their closest maximal projection ancestors, as well as subtrees rooted in the least common ancestor. (For the full list of feature schemas, the reader can refer to Charniak and Johnson (2005).)

Since the parser keeps track of these features, we can make them perform 'double duty' — both for reranking the parses and as classification features. Based on the 50-best parser outputs returned in the first stage of the reranking parser, features that meet a certain count threshold are selected. There are 3,380,722 features extractable from the training data that we have. Taking the reranker's default count threshold of 5, the number of features was reduced to 1,133,261. Three experiments are to be conducted: first on the 100,000 most frequently occurring features, followed by the 500,000 most frequently occurring ones, and finally all of the 1,133,261 features.[7]

4.2.3.2 Classification Settings

Training. The training set consists of a balance of grammatical and ungrammatical sentences, with 79,664 in total (i.e. 39,832 sentences for each class). As mentioned before, the grammatical sentences are adopted from *PureWSJ 02-21* and the ungrammatical ones are from *NoisyWSJ 02-21*.

Testing. The testing set is also a balanced set of the two classes of sentences. Unlike for training, we have two sets of testing data. While the first set is formed from *PureWSJ 00* and *NoisyWSJ 00*, the second set is from *PureWSJ 00* and *NoisierWSJ 00*. The intention of testing on the latter set is to examine the extent to which the degree of noisiness in the data would affect the overall classification performance. 3,840 sentences are used for testing with the Stanford parser and 3,842 sentences with the C&J parser.[8]

Classifiers. Two types of machine learner are deployed for all the classification tasks: a support vector machine (SVM) as per Wong and Dras (2010), and a maximum entropy (maxent) learner — one of the classifiers that has been gaining much success in text classification (Nigam et al.,

[7] The reason for selecting the 100,000 features as the first set for the reranking features is that there is an issue of feature sparsity even with 50,000 features.

[8] The difference is due to the Stanford parser returning no parses for one of the test sentences in *PureWSJ*.

Table 4.1: Stanford parsing evaluation (labelled f-score %) of five experiments on three versions of WSJ Section 00

Exp	Parser	PureWSJ	NoisyWSJ	NoisierWSJ
1	PureParser	85.61	78.42	72.64
2	NoisyParser	84.31	80.32	76.19
3	MixedParser	82.63	78.69	74.25
4	**Pure-NoisyParser**	85.39	80.43	76.40
5	Pure-MixedParser	85.49	80.04	75.53

Table 4.2: C&J parsing evaluation (labelled f-score %) of five experiments on three versions of WSJ Section 00

Exp	Parser	PureWSJ	NoisyWSJ	NoisierWSJ
1	PureParser	91.09	82.41	76.78
2	NoisyParser	89.73	82.92	78.96
3	MixedParser	90.33	83.80	79.48
4	Pure-NoisyParser	89.79	83.19	78.73
5	**Pure-MixedParser**	90.66	83.64	79.09

1999; Zhu et al., 2005). We use the online SVM tool, *LIBSVM* (version 2.89) by Chang and Lin (2001) and the maxent tool, *MegaM*[9] (fifth release) by Hal Daumé III. Each classifier is tuned to obtain an optimal classification model.

4.2.4 Results

4.2.4.1 Parser Evaluation

Here, we present the parsing results of the five experiments conducted in the first stage where the intention is to induce a more robust parsing model that can handle ungrammatical sentences without compromising its performance on grammatical ones.

Tables 4.1 and 4.2 are the parsing evaluations for the Stanford parser and the C&J parser respectively. In general, the C&J parser performs better than the Stanford parser in all the parsing experiments. For Stanford, *Pure Parser* integrated with *Noisy Parser* in the fourth experiment is the best combination in the sense that this integrated model performs relatively well for both grammatical and ungrammatical data; while for C&J, *Pure Parser* with *Mixed Parser* in the last experiment is best. These two best models are used for all the parsing tasks in the second stage (to produce parse outputs for the purpose of classification).

[9]MegaM software is available at `http://www.umiacs.umd.edu/~hal/megam/`.

Table 4.3: Classification results (accuracy %) with 'parse probabilities' as features on both NoisyWSJ and NoisierWSJ — First scenario

Features	**SVM**		**MegaM**	
Parse Prob	**Stanford**	**C&J**	**Stanford**	**C&J**
Pure-NoisyWSJ	65.46	61.56	66.43	64.24
Pure-NoisierWSJ	74.19	71.76	74.97	74.46

Table 4.4: Classification results (accuracy %) with 'parse probabilities' as features on both NoisyWSJ and NoisierWSJ — Second scenario

Features	**SVM**		**MegaM**	
Parse Prob	**Stanford**	**C&J**	**Stanford**	**C&J**
Pure-NoisyWSJ	50.02	51.54	50.00	52.81
Pure-NoisierWSJ	53.46	53.80	50.00	55.46

4.2.4.2 Classification Performance

In the following tables of classification results, *Pure-NoisyWSJ* denotes the set of testing data consists of *PureWSJ* and *NoisyWSJ*, while *Pure-NoisyWSJ* refers to the testing set containing *PureWSJ* and *NoisierWSJ*. (All tests are performed on Section 00.)

Parse probabilities. For the classification of sentence grammaticality, a reasonably good accuracy (better than chance) is achievable by using just parse probabilities alone as features under the ideal first scenario, in which the gold standard ungrammatical parses are available for parser training (see Table 4.3). Unsurprisingly, noisier data is easier to distinguish — for instance, comparing the SVM classification accuracy of 65.46% (*Pure-NoisyWSJ*) with 74.19% (*Pure-NoisierWSJ*) for the Stanford parser, and 61.56% (*Pure-NoisyWSJ*) with 71.76% (*Pure-NoisierWSJ*) for the C&J parser. Also relatively unsurprisingly, with the use of just *PureParser* in the second scenario, the classifiers perform no better than chance with just a single set of parse probabilities as classification features (as shown in Table 4.4).

Parse rules with Information Gain. Table 4.5 shows that using syntactic information in the form of production rules produces results substantially better than chance. However, results are still below using only parse probabilities alone (Table 4.3). In terms of relative machine learner performance, for the first scenario using IG (see Table 4.5), our previous work (Wong and Dras, 2010) had found that, with an SVM and the Stanford parser, using more production rules led to a reduction in classification performance for *Pure-NoisyWSJ*. However, the MegaM (maxent)

Table 4.5: Classification results (accuracy %) with 'parse rules IG' as features on both NoisyWSJ and NoisierWSJ — First scenario

Features		SVM		MegaM	
n	**IG**	**Stanford**	**C&J**	**Stanford**	**C&J**
100	Pure-NoisyWSJ	60.67	54.01	60.70	54.42
	Pure-NoisierWSJ	63.43	58.02	63.46	58.41
500	Pure-NoisyWSJ	58.07	53.20	63.23	57.13
	Pure-NoisierWSJ	60.81	56.51	67.06	61.97
1000	Pure-NoisyWSJ	56.69	54.61	63.93	58.30
	Pure-NoisierWSJ	59.95	57.65	68.54	63.35

Table 4.6: Classification results (accuracy %) with 'parse rules IG' as features on both NoisyWSJ and NoisierWSJ — Second scenario

Features		SVM		MegaM	
n	**IG**	**Stanford**	**C&J**	**Stanford**	**C&J**
100	Pure-NoisyWSJ	60.31	59.03	60.65	59.63
	Pure-NoisierWSJ	66.88	63.20	67.16	65.20
500	Pure-NoisyWSJ	58.91	58.04	62.03	61.61
	Pure-NoisierWSJ	66.30	63.72	69.74	69.36
1000	Pure-NoisyWSJ	58.77	57.18	62.97	62.41
	Pure-NoisierWSJ	64.45	61.48	70.96	69.91

results of the Stanford parser demonstrate the opposite — as the number of rules increases, the accuracy is gradually improved. Furthermore, in absolute magnitude, these classification scores are a major improvement over those reported in Wong and Dras (2010) (replicated in the SVM-Stanford columns of Table 4.5). As for the C&J parser, there is no noticeable improvement in the SVM classification accuracies; but improvements in classification are observed with MegaM.

Turning to the second scenario (shown in Table 4.6), similar classification performance patterns are observed for both parsers. More precisely, the use of more production rules causes a reduction in the overall SVM classification performance for *Pure-NoisyWSJ*; unlike the cases for MegaM where improvements are observed.

Parse rules with Binormal Separation. BNS performed similarly to IG. For the first scenario with BNS (see Table 4.7), the use of more production rules does not have much impact on the overall SVM classification performance for *Pure-NoisyWSJ* be it the case for the Stanford parser or the C&J parser. MegaM, on the other hand, demonstrates the opposite where the classification performance is gradually improved with more production rules. As for IG, both parsers demonstrate similar classification performance behaviours under the second scenario. As shown in Table 4.8, MegaM shows classification improvements on *Pure-NoisyWSJ* when more production rules are

Table 4.7: Classification results (accuracy %) with 'parse rules BNS' as features on both NoisyWSJ and NoisierWSJ — First scenario

Features		SVM		MegaM	
n	**BNS**	**Stanford**	**C&J**	**Stanford**	**C&J**
100	Pure-NoisyWSJ	57.58	53.85	58.18	55.00
	Pure-NoisierWSJ	61.51	56.59	62.89	58.20
500	Pure-NoisyWSJ	58.77	54.89	62.81	56.64
	Pure-NoisierWSJ	63.88	58.35	67.10	61.17
1000	Pure-NoisyWSJ	57.39	54.71	63.59	57.52
	Pure-NoisierWSJ	61.35	57.78	68.23	62.41

Table 4.8: Classification results (accuracy %) with 'parse rules BNS' as features on both NoisyWSJ and NoisierWSJ — Second scenario

Features		SVM		MegaM	
n	**BNS**	**Stanford**	**C&J**	**Stanford**	**C&J**
100	Pure-NoisyWSJ	60.03	58.12	60.15	57.96
	Pure-NoisierWSJ	65.15	64.18	65.68	63.79
500	Pure-NoisyWSJ	59.09	57.10	62.58	60.38
	Pure-NoisierWSJ	65.96	63.33	69.61	68.01
1000	Pure-NoisyWSJ	58.96	57.05	63.70	61.11
	Pure-NoisierWSJ	64.48	62.44	71.64	69.34

deployed but not for the cases with SVM.

As before, more noisy data is generally easier to distinguish from grammatically well-formed data. All the classification accuracies for *Pure-NoisierWSJ* are, therefore, much higher overall (for both feature selection metrics).

Combining with parse probabilities. In Tables 4.9 and 4.10, we present the results of integrating parse production rules with parse probabilities. (Note that, the number of parse rules used here is based on the top 100 rules since this is generally the best performing set when the SVM setting is taken into consideration.) In the case of the first scenario, combining these two types of features demonstrates some slight improvement in the overall SVM classification performance for both parsers, with the performance of parse probabilities as the point of reference. Improvements occur for the Stanford parser with MegaM (67% vs 66.43%) but not for the C&J parser (64.37% vs 64.24%). As for the second scenario, since there is only one set of parse probabilities, no improvements are observed in terms of the classification performance (as expected); this has in turn worsened the accuracies of using production rules alone in most cases.

Reranking features. Table 4.11 presents the classification results with the reranking features. Here, having established above the consistent relative performance of the two scenarios, we present

Table 4.9: Classification results (accuracy %) with the combinations of 'parse rules and parse probabilities' as features on NoisyWSJ — First scenario

Features	**SVM**		**MegaM**	
Combinations	**Stanford**	**C&J**	**Stanford**	**C&J**
IG + BNS	58.52	53.28	60.83	56.19
IG + Parse Prob	65.60	63.30	67.00	64.37
BNS + Parse Prob	66.02	63.25	66.77	63.35
All three types	66.30	63.33	66.54	63.97

Table 4.10: Classification results (accuracy %) with the combinations of 'parse rules and parse probabilities' as features on NoisyWSJ — Second scenario

Features	**SVM**		**MegaM**	
Combinations	**Stanford**	**C&J**	**Stanford**	**C&J**
IG + BNS	60.26	58.41	61.64	59.60
IG + Parse Prob	56.80	55.13	58.44	59.63
BNS + Parse Prob	57.01	55.57	59.84	58.17
All three types	57.03	56.12	58.31	59.55

results under the realistic second scenario (refer to the discussion section for further details — Section 4.2.5), and only with the MegaM as the SVM implementation could not handle the quantity of reranking features. As shown in the results, reranking features appear to be a better discriminant as compared to parse production rules (be it based on IG or BNS). For instance, comparing the accuracy rate of 64.42% achieved by using the 100,000 most frequently occurring reranking features for *Pure-NoisyWSJ* with the accuracy of 62.41% attained by the use of the 1,000 parse rules with the highest IG (see Table 4.6). If all the reranking features are used, the classification accuracy increases to 65.25% for *Pure-NoisyWSJ* and 71.06% for *Pure-NoisierWSJ*.

In comparison with the accuracies reported by Post (2011) as mentioned in Section 4.1, our reranking feature-based results are indeed lower as expected. Nonetheless, our ungrammatical sentences are of more realistic errors which seem to have posed a more difficult classification problem.

Table 4.11: Classification results (accuracy %) with 'reranking features' as features on NoisyWSJ and NoisierWSJ — Second scenario

Features	**MegaM (2nd scenario)**		
Reranking	**n = 100,000**	**n = 500,000**	**n = 1,133,261**
Pure-NoisyWSJ	64.42	65.02	65.25
Pure-NoisierWSJ	69.99	70.10	71.06

Table 4.12: Classification results (accuracy %) with the combinations of 'reranking features and parse rules' as features on NoisyWSJ — Second scenario

Features	**MegaM (2nd scenario)**		
Combinations	**IG + Reranking**	**BNS + Reranking**	**All three**
Pure-NoisyWSJ	66.37	66.71	66.97

Combining with parse rules. Furthermore, integrating parse rules with reranking features results in greater improvement in the overall classification performance (as shown in Table 4.12). Adding all the reranking features to the 1,000 production rules with either the highest IG or the highest BNS has led to an approximately 4–5% increase in accuracy. (Note that, under the 2nd scenario with the C&J parser, the MegaM accuracy of using only the top 1,000 IG selected rules is 62.41% for *Pure-NoisyWSJ*, while for the top 1,000 BNS selected rules is 61.11%.) Combining all three types of features (i.e. the two feature selection metrics for parse production rules and the reranking features) yields an overall accuracy of 66.97% which is almost 7% higher than the performance of combining just the two feature selection metrics alone for the case of the C&J parser (59.60% as shown in Table 4.10).

4.2.5 Discussion

First scenario versus second scenario. This is perhaps the most striking result, and of particular relevance to the following NLI task. Comparing the two scenarios, taking features of IG selected parse rules for example, the classification results under the second scenario (a single parser) are noticeably higher (compare Table 4.5 with Table 4.6). This is somewhat surprising, in light of the direction of previous work discussed in Section 4.1 and at the start of this section: in terms of statistical parsers, it is not actually beneficial to train two separate parsers on grammatical and ungrammatical data respectively, as far as structural information is concerned. Similar performance patterns are indeed observed amongst the cases with BNS selected parse rules (compare Table 4.7 with Table 4.8).

Stanford parser versus C&J parser. Interestingly, in all of the classifications based on parse production rules as well as parse probabilties, the Stanford parser produces better results, even though by f-score it was worse than the C&J parser (compare Table 4.1 with Table 4.2). Looking at the data, we believe that this is because the C&J parser is better at 'handling' ungrammaticality, effectively ignoring errors and consequently producing fewer structural differences in the parse trees. This may be a consequence of the *causal competition* effect ('explaining-away') inherent

```
(S1
  (S
    (ADVP (RB Nonetheless))
    (, ,)
    (NP (DT the) (NN suit))
    (VP (VBZ seeks)
      (NP (JJ unspecified) (NNS damages)
        (SBAR
          (IN that)
          (S
            (NP (NP (DT man) (NN attorney))
              (PP (IN for)
                (NP (NNP Quickview)))
              (VBD claimed))
            (VP (MD could)
              (VP (VB run)
                (PP (IN into)
                  (NP (NP (DT the) (NNS millions))
                    (PP (IN of)
                      (NP (NNS dollars)))))))))))
    (. .)))
```

Figure 4.1: An example of parse from C&J parser illustrating the effect of 'explaining-away'

in maxent models — the C&J parser's reranker was implemented under the maxent framework. Under causal competition in the context of a parser, an implausible part of speech, say, may be assigned because of strong n-gram or structural associations. Compounding this, both the Stanford and C&J parsers have quite strong built-in linguistic constraints: for example, a sentence will be forced to have a verb, even if none of the words in an ungrammatical sentence are likely candidates.[10] Figure 4.1 demonstrates this phenomenon with an ungrammatical sentence extracted from the NoisyWSJ. The noun phrase *man attorney* was the result of substitution of an article *an* with a noun *man* in a presumably grammatical WSJ sentence *Nonetheless, the suit seeks unspecified damages that* ***an attorney*** *for Quickview claimed could run into the millions of dollars*. Despite the presence of ungrammaticality, the C&J parser still produced the parse that is identical to that of the grammatical sentence given that `NP -> DT NN` is seemingly much more probable in contrast to `NP -> NN NN`.

Table 4.13 provides a closer look at the parse production rules produced by the two parsers. It illustrates the ratio between the number of occurrences of a particular parse rule in both the grammatical and ungrammatical corpora. It can be argued that the greater the ratio for a particular parse rule, the more prominent a feature it might make for classifying sentences based on their grammaticality. Referring to the examples given, the ratios seem to be higher for the parse rules returned

[10] This comment comes from Mark Johnson (personal communication). To our knowledge, this is not documented in the literature, although it may be noted by inspecting the code.

Table 4.13: Examples of parse rules returned by Stanford and C&J parsers

Parse Rules	Stanford		C&J	
IG	**+**	**–**	**+**	**–**
NP → DT DT JJ NN	2	225	1	141
VP → TO TO VP	0	89	0	24
PP → NN IN NP	0	70	0	1
NP → NP PP VBD	0	54	0	26
BNS	**+**	**–**	**+**	**–**
PP → IN IN NP	105	1858	334	917
VP → VBZ VBZ NP	0	157	1	61
NP → DT DT NN	2	531	6	383
S → NP VBD VP .	0	242	0	40

+: grammatical; –: ungrammatical

by the Stanford parser as compared to those by the C&J parser. It seems that the C&J parser is successfully ignoring the ungrammaticality in order to produce a parse. Figure 4.2 presents an example that could conceivably provide some insight into why the higher ratios are observed amongst the Stanford parse rules as compared to those from the C&J. Comparing the two parses derived for an ungrammatical sentence with two infinitivals *to* occuring together, Stanford produced an uncommon parse rule `VP -> TO TO VP` (reflecting some form of ungrammaticality) while C&J produced the typically grammatical rule `VP -> TO VP`.

Further analysis of some other parses reveals that the C&J parser is not always immune to ungrammaticality. Figure 4.3 shows an ungrammatical sentence with the presence of two third-person verbs *has* occuring in sequence. Although the C&J parser produced the more probable `VP -> VBZ NP` in contrast to the less probable `VP -> VBZ VBZ NP` by the Stanford parser, the C&J parse also resulted in another typically ungrammatical rule `NP -> VBZ DT JJ JJ NN`.

Summary. We have shown that using syntactic structural information in the form of parse reranking templates can identify ungrammaticality at rates substantially greater than chance. We also found that the maxent-based machine learner MegaM, allowing the useful incorporation of larger numbers of features, contributes strongly in terms of prediction. On the other hand, having an ungrammatical treebank does not in fact lead to improved classification (as depicted by the classification results of the two scenarios). Furthermore, the better C&J parser, a robust statistical parser that is capable of parsing both grammatical and ungrammatical data equally well, does not necessarily lead to the best results in terms of classifying sentences with respect to their grammaticality: the Stanford parser performs better here by not ignoring ungrammaticality to the extent of the C&J parser. This suggests for the NLI task that, in addition to these two parsers, another parser that has

```
(ROOT
  (S (`` ``)
    (S
      (NP (PRP We))
      (VP (VBP consider)
        (NP (PRP$ our) (JJ internal) (NNS controls))
        (S
          (VP (TO to) (TO to)
            (VP (VB have)
              (VP (VBN worked) (ADVP (RB well))))))))
    (, ,) ('' '')
    (NP (PRP she))
    (VP (VBD was) (VBD said) (, ,)
      (S
        (VP (VBG adding)
          (SBAR (IN that)
            (S
              (NP (DT some) (NNS procedures))
              (VP (VBP have)
                (VP (VBN been) (VP (VBN strengthened))))))))
    (. .)))
(S1
  (S (`` ``)
    (S
      (NP (PRP We))
      (VP (VBP consider)
        (S
          (NP (PRP$ our) (JJ internal) (NNS controls))
          (VP (TO to)
            (VP (TO to) (VB have)
              (VP (VBN worked) (ADVP (RB well))))))))
    (, ,) ('' '')
    (NP (PRP she))
    (VBD was)
    (VP (VBD said) (, ,)
      (S
        (VP (VBG adding)
          (SBAR (IN that)
            (S
              (NP (DT some) (NNS procedures))
              (VP (VBP have)
                (VP (VBN been) (VP (VBN strengthened))))))))
    (. .)))
```

Figure 4.2: Parses by Stanford (top) and C&J (bottom) parsers exhibiting grammatical-ungrammatical ratio difference

```
(ROOT
  (S
    (NP (NNP Warner))
    (VP (VBZ has) (VBZ has)
      (NP
        (NP (DT a) (JJ five-year) (JJ exclusive) (NN contract))
        (PP (IN with)
          (NP (NNP Mr.) (CC and) (NNP Mr.) (NNP Peters)))
        (SBAR
          (WHNP (WDT that))
          (S
            (VP (VBZ requires)
              (S (NP (PRP them))
                (VP (TO to)
                  (VP (VB make)
                    (NP (NNS movies))
                    (ADVP (RB exclusively))
                    (PP (IN at)
                      (NP (DT the)
                      (NNP Warner) (NNPS Bros.) (NN studio)))))))))))
    (. .)))
(S1
  (S
    (NP (NNP Warner))
    (VP (VBZ has)
      (NP
        (NP (VBZ has) (DT a) (JJ five-year) (JJ exclusive) (NN contract))
        (PP (IN with)
          (NP (NP (NNP Mr.)) (CC and) (NP (NNP Mr.) (NNP Peters))))
        (SBAR
          (WHNP (WDT that))
          (S
            (VP (VBZ requires)
              (S (NP (PRP them))
                (VP (TO to)
                  (VP (VB make)
                    (NP (NNS movies))
                    (PP (ADVP (RB exclusively))
                      (IN at)
                      (NP (DT the)
                         (NNP Warner) (NNPS Bros.) (NN studio)))))))))))
    (. .)))
```

Figure 4.3: Parses by Stanford (top) and C&J (bottom) demonstrating different forms of ungrammatical parse rule

weaker linguistic constraints should be investigated.

4.3 Native Language Identification

Given the demonstration in the previous section that parse structures (such as parse production rules and more generalised parse reranking features) appear to be useful in characterising sentence grammaticality, it seems promising to apply a similar approach to the task of native language identification (NLI), exploiting syntactic errors as classification features. The goal of this section (as the second task of this chapter) is, therefore, to investigate the influence to which syntactic features including syntactic errors represented by parse structures would have on the classification task of identifying an author's native language relative to, and in combination with, typical lexical features.

4.3.1 Models for NLI

In this subsection, we describe the three basic models to be investigated for NLI: the first is the lexical model, based on Koppel et al. (2005), serving as the baseline; followed by the two models that exploit syntactic information — parse production rules and reranking features. Then we look at the classification performance of each model independently and also in combination: to combine, we just concatenate feature vectors linearly.

4.3.1.1 First Model: Lexical Features

As per Chapter 3 (similarly to Wong and Dras (2009)), we replicate the features of Koppel et al. (2005) to produce our LEXICAL model. These are of three types: function words,[11] character n-grams, and PoS n-grams.

We follow the approach of Chapter 3 in resolving some unclear issues from Koppel et al. (2005). Specifically, we use the same list of function words, left unspecified in Koppel et al. (2005), that were empirically determined in Chapter 3 to be the best of three candidates. We used character bi-grams, as the best performing n-grams, although this also had been left unspecified by Koppel et al. (2005); and we used the most frequently occurring PoS bi-grams,[12] obtained by using the Brill tagger provided in NLTK (Bird et al., 2009) trained on the Brown corpus.

In total, there are 798 features of this class with 398 function words, 200 most frequently occurring character bi-grams, and 200 most frequently occurring PoS bi-grams. Both function

[11] As noted in Section 4.1, we use only function words, so that the result is not tied to a particular domain, and no clues are obtained from different topics that different authors might write about.

[12] We omitted the 250 rare PoS bi-grams used by Koppel et al. (2005), as an ablative analysis showed that they contributed nothing to classification performance.

words and PoS bi-grams are of binary-typed feature values; while for character bi-grams, the feature value is of the relative frequency. (These types of feature value are the best performing ones for each lexical feature.)

4.3.1.2 Second Model: Production Rules

Under this model (PROD-RULE), we take as features horizontal slices of parse trees, in effect treating them as sets of CFG production rules (as used in the experiments for sentence grammatical judgement). Feature values are of binary type. We look at all possible rules as features, but also present results for subsets of features chosen using feature selection. For each language in our dataset, we identify the n rules most characteristic of the language using Information Gain (IG). We adopt the formulation of IG of Yang and Pedersen (1997) given in Section 4.2.3.1.

We also investigated other feature selection metrics such as simple frequencies, frequency ratios, and pointwise mutual information, as per our previous work (Wong and Dras, 2010). As in much other work, IG performed best, so we do not present results for the others. Binormal Separation (Forman, 2003), often competitive with IG (as seen in the results presented for sentence grammaticality judgement), is only suitable for binary classification — hence it is not applicable to this multi-classification task.

It is worth noting that the production rules being used here are all non-lexicalised ones (i.e. elements of the right-hand side of each rule are non-terminals), except those rules that are lexicalised with function words and punctuation (i.e. the right-hand side of these are terminals), in order to avoid topic-related clues.

4.3.1.3 Third Model: Reranking Features

Reranking features are cross-sections of parse trees that might capture other aspects of language-specific ungrammatical structures, or of overuse or avoidance. It should be noted that the 13 feature schemas as described in Charniak and Johnson (2005) are not the only possible ones — they were empirically selected for the specific purpose of augmenting the Charniak parser (Charniak, 2000). However, much subsequent work has tended to use these same features, albeit sometimes with extensions for specific purposes: for instance, Johnson and Ural (2010) for the Berkeley parser (Petrov et al., 2006), and Ng et al. (2010) for the C&C parser (Clark and Curran, 2007). We also use this standard set as our third feature model (RERANKING), specifically the set of instantiated feature schemas from the C&J parser (Charniak and Johnson, 2005) trained on the Wall Street Journal (WSJ) treebanks (as described in the previous section of sentence grammaticality judgement — Section 4.2).

4.3.2 Experimental Setup

4.3.2.1 Corpus

As per the previous chapter (Chapter 3), we use the same corpus here — the second version of the *International Corpus of Learner English* (ICLE) (Granger et al., 2009). For comparability reason, we examine the same seven languages, namely Bulgarian, Czech, French, Russian, Spanish, Chinese, and Japanese. For the purpose of the present study, we have 95 essays per native language. Of the 95 texts per language, 70 essays are used for training, where the remaining 25 essays are held out for testing. In total, there are 17,718 training sentences and 6,791 testing sentences. It should be noted that the training and testing sets selected here are the exact same sets as used in Chapter 3 and also Wong and Dras (2009).

It is also worth mentioning that some of the essays from the corpus were annotated with special characters that appear to have some impact on the overall classification performance. To prevent any potential biases towards particular native language groups, such special characters are ignored in our experiments.[13]

4.3.2.2 Parsers

We employ three PCFG-based parsers: the Stanford parser (Klein and Manning, 2003) and the C&J parser (Charniak and Johnson, 2005) as before, together with the Berkeley parser (Petrov et al., 2006).

The Berkeley parser, similar to the Stanford parser which uses an unlexicalised grammar, also produces state-of-the-art parsing results, comparable to the best lexicalised C&J parser, with a reported f-score of 90.2%. The Berkeley parsing model utilises latent variables and a hierachical split-and-merge approach to augment non-terminal categories such that optimal state splits can be learned automatically using the Expectation-Maximization (EM) algorithm. Such a model improves over the Stanford model in which state splits are manually identified in response to linguistic observations in the data. We add the Berkeley parser here because, in contrast to the C&J and Stanford parsers, and perhaps related to its quite fine-grained final categories as a result of its split-and-merge operation, it seems to less strongly coerce a parse to have the properties associated with a grammatical sentence (e.g. that it should contain a verb).

With the Stanford parser, there are 26,234 unique parse production rules extractable from our ICLE training set of 490 texts, while the C&J parser produces 27,610 and the Berkeley parser produces 75,790. For the Berkeley parser, we take fine-grained production rules with latent variables

[13] This was suggested by Joel Tetreault and Daniel Blanchard from Educational Testing Service (ETS) (personal communication).

as features (hence the much higher number of rules). The intention is to investigate whether parse rules with finer-grained non-terminals would capture deeper syntactic structures characterising individual native languages.

In terms of reranking features, we use only the C&J parser (for the same reason as before) — since the parser stores these features during parsing, we can use them directly as classification features as before. With the ICLE training data, there are 6,230 features with frequency >10, and 19,659 with frequency >5.

4.3.2.3 Classifiers

As per the sentence grammaticality judgement experiment, we use the maximum entropy (maxent) machine learner, MegaM (fifth release) by Hal Daumé III. (We also used an SVM for comparison purposes, but the results were uniformly worse, as with grammaticality judgement, and degraded more quickly as the number of features increased, so we only report the maxent results here.) Similarly, the classifier is tuned to obtain an optimal classification model for each feature set.

4.3.2.4 Evaluation Methodology

We give results for two types of test data: cross-validation and held-out. Given our relatively small amount of data, we use k-fold cross-validation on the set of 70 documents per language, choosing $k = 5$. While testing for statistical significance of classification results is often not carried out in NLP, we do so here because the quantity of data could raise questions about the certainty of any effect. In an encyclopedic survey of cross-validation in machine learning contexts, Refaeilzadeh et al. (2009) note that there is as yet no universal standard for testing of statistical significance; and that while more sophisticated techniques have been proposed, none is more widely accepted than a *paired t-test* over folds. We therefore use this paired t-test over folds, as formulated by Alpaydin (2004). Under this cross-validation, 5 separate training feature sets are constructed, excluding the test fold; 3 folds are used for training, 1 fold for tuning and 1 fold for testing.

We also use a held-out test set for comparison, as it is well-known that cross-validation can overestimate prediction errors (Hastie et al., 2009). We do not conduct significance testing here — with this held-out test set size ($n = 175$ with 25 per native language), two models would have to differ by a great deal to be statistically significant. We only use it as a check on the effect of applying the model to completely new data.

Table 4.14: Classification results with parse rules as syntactic features (accuracy %) — 5-fold cross validation

Features	MaxEnt
LEXICAL ($n = 798$)	64.29
PROD-RULE[Stanford] ($n = 1000$)	64.90
PROD-RULE[Stanford]* ($n = 1000$)	72.45
PROD-RULE[Stanford]* ($n = all$)	74.29
PROD-RULE[C&J] ($n = 1000$)	62.65
PROD-RULE[C&J]* ($n = 1000$)	72.04
PROD-RULE[C&J]* ($n = all$)	71.84
PROD-RULE[Berkeley] ($n = 1000$)	68.79
PROD-RULE[Berkeley]* ($n = 1000$)	75.10
PROD-RULE[Berkeley]* ($n = all$)	75.10
PROD-RULE[Stanford+C&J] ($n = 2000$)	67.14
PROD-RULE[Stanford+C&J]* ($n = 2000$)	74.49
PROD-RULE[Stanford+C&J]* ($n = all$)	75.71
PROD-RULE[Stanford+C&J+Berkeley] ($n = 3000$)	72.00
PROD-RULE[Stanford+C&J+Berkeley]* ($n = 3000$)	79.18
PROD-RULE[Stanford+C&J+Berkeley]* ($n = all$)	***80.00***
RERANKING (all features)	67.96
RERANKING (>5 counts)	66.33
RERANKING (>10 counts)	64.90

** Rules lexicalised with function words are included*

4.3.3 Results

Table 4.14 presents the results for the three models individually under the cross-validation setting.[14] The first point to note is that PROD-RULE, under all the three parsers, is a substantial improvement over LEXICAL when (non-lexicalised) parse rules together with rules lexicalised with function words are used (rows marked with * in Table 4.14), with the largest difference as much as 80.0% for PROD-RULE [Stanford+C&J+Berkeley]* ($n = all$) versus 64.29% for LEXICAL; these differences with respect to LEXICAL are statistically significant at the confidence level of 95%. (To give an idea, the paired t-test standard error for this largest difference is 2.78%.) In terms of error reduction, this is over 40%.

There appears to be no significant difference between the Stanford and C&J parsers, regardless of their differing accuracy on the WSJ; but the Berkeley parser generally results in much better performance in comparison to the other two parsers. On the whole, using the selection metric with IG for PROD-RULE without rules lexicalised with function words produces results mostly around those for LEXICAL; except that Berkeley's PROD-RULE[Berkeley] ($n = 1000$) still demonstrates

[14] An earlier version of some of the results in this table were published in Wong and Dras (2011) where the results were based on the experiments that included those parse production rules lexicalised with special symbols that might skew the overall classification performance, as highlighted earlier. The results presented here are, on the other hand, based on the experiments that ignored these symbols. (By and large, these revised results are marginally lower.)

Table 4.15: Classification results with parse rules as syntactic features (accuracy %) — held-out validation

Features	MaxEnt
LEXICAL ($n = 798$)	75.43
PROD-RULE[Stanford] ($n = 1000$)	72.57
PROD-RULE[Stanford]* ($n = 1000$)	78.86
PROD-RULE[Stanford]* ($n = all$)	78.86
PROD-RULE[C&J] ($n = 1000$)	73.71
PROD-RULE[C&J]* ($n = 1000$)	78.86
PROD-RULE[C&J]* ($n = all$)	80.57
PROD-RULE[Berkeley] ($n = 1000$)	73.14
PROD-RULE[Berkeley]* ($n = 1000$)	80.57
PROD-RULE[Berkeley]* ($n = all$)	78.28
PROD-RULE[Stanford+C&J] ($n = 2000$)	77.71
PROD-RULE[Stanford+C&J]* ($n = 2000$)	78.28
PROD-RULE[Stanford+C&J]* ($n = all$)	78.86
PROD-RULE[Stanford+C&J+Berkeley] ($n = 3000$)	80.00
PROD-RULE[Stanford+C&J+Berkeley]* ($n = 3000$)	***82.28***
PROD-RULE[Stanford+C&J+Berkeley]* ($n = all$)	81.71
RERANKING (all features)	77.14
RERANKING (>5 counts)	76.57
RERANKING (>10 counts)	75.43

** Rules lexicalised with function words are included*

modest improvement over LEXICAL (and this is statistically significant at the confidence level of 95%). Using fewer reranking features is worse as the quality of RERANKING declines when feature cut-offs are raised.

Another, somewhat surprising point is that the RERANKING results are also generally around those of LEXICAL even though like PROD-RULE they are also utilising cross-sections of the parse trees (which was expected to be performing better). We consider there might be a number of possible explanations for this. Firstly, the feature schemas used were originally chosen for the specific purpose of augmenting the performance of the Charniak parser; perhaps others might be more appropriate for our purpose here. Secondly, we selected only those instantiated feature schemas that occurred in the WSJ, and then applied them directly to ICLE. As the WSJ is filled with predominantly grammatical text, perhaps those that were not seen on the WSJ are precisely those that might indicate ungrammaticality. In contrast, the production rules of PROD-RULE were selected merely from the ICLE training data. We also note that genre difference may have some impact here: given that the WSJ treebank consists of predominantly newswire texts, while the ICLE corpus is a collection of mostly argumentative essays; perhaps the feature schemas are particularly appropriate for newswire texts but not for essays. In relation to this, the proficiency of writers who contributed to both the WSJ and ICLE might be another confounding factor. From the technical

Table 4.16: Classification results for combined models (accuracy %) — 5-fold cross validation

Features	MaxEnt
LEXICAL ($n = 798$)	64.29
LEXICAL + PROD-RULE[Stanford+C&J] ($n = 2000$)	60.41
LEXICAL + PROD-RULE[Stanford+C&J]* ($n = 2000$)	72.04
LEXICAL + PROD-RULE[Stanford+C&J]* ($n = all$)	71.43
LEXICAL + PROD-RULE[Stanford+C&J+Berkeley] ($n = 3000$)	68.57
LEXICAL + PROD-RULE[Stanford+C&J+Berkeley]* ($n = 3000$)	*76.94*
LEXICAL + PROD-RULE[Stanford+C&J+Berkeley]* ($n = all$)	76.33
LEXICAL + RERANKING ($n = all$)	68.17

** Rules lexicalised with function words are included*

standpoint, it could be the case of 'overfitting', given a relative small training data in comparison with a huge number of reranking features; consequently, there might be the problem of feature vector sparsity.

Table 4.15 presents the results for the individual models on the held-out test set.[15] The results are generally higher than for cross-validation — this is not surprising, as the texts are of the same type, but all the training data was used (rather than the $1 - 2/k$ proportion under cross-validation). Overall, the pattern is still the same as observed under cross validation, with PROD-RULE best, then RERANKING and LEXICAL broadly similar; as expected, no differences are significant with this smaller dataset. The gap has narrowed, but without significance testing it is difficult to conclude whether this is a genuine phenomenon. One point to be noted is that the overall best result for the held-out testing is from PROD-RULE[Stanford+ C&J+Berkeley]* ($n = 3000$) with the accuracy rate of 82.28% rather than PROD-RULE[Stanford+C&J+ Berkeley]* ($n = all$) as demonstrated under cross-validation (refer to Table 4.14). As for the accuracy rate for LEXICAL here, it is indeed in line with Wong and Dras (2009); and given the smaller dataset as well as the larger set of languages (7 versus 5), it is also broadly in line with Koppel et al. (2005).

Tables 4.16 and 4.17 present results for model combinations (i.e. integrating LEXICAL with PROD-RULE and RERANKING). It can be seen that the model combinations do not produce results better than using PROD-RULE alone; but again a slight improvement is observed when integrating with Berkeley's PROD-RULE along with rules lexicalised with function words, under the held-out testing. Combining all feature sets (where results are not presented here) seems to degrade the overall performance even of the MegaM: perhaps we need to derive feature vectors more

[15] It should be noted that the results presented here are not directly comparable with the results presented in the previous chapter (Chapter 3) in particular Table 3.11, given that the types of feature value used for the lexical features are somewhat different. In Chapter 3, the results are based on the relative frequencies of the lexical features; whereas the results presented here are based on the best performing type of feature value for each lexical feature: function words and PoS n-grams are both based on binary-typed of feature value; only char n-grams are based on the relative frequencies.

Table 4.17: Classification results for combined models (accuracy %) — held-out validation

Features	MaxEnt
LEXICAL ($n = 798$)	75.43
LEXICAL + PROD-RULE[Stanford+C&J] ($n = 2000$)	81.14
LEXICAL + PROD-RULE[Stanford+C&J]* ($n = 2000$)	80.57
LEXICAL + PROD-RULE[Stanford+C&J]* ($n = all$)	81.71
LEXICAL + PROD-RULE[Stanford+C&J+Berkeley] ($n = 3000$)	82.28
LEXICAL + PROD-RULE[Stanford+C&J+Berkeley]* ($n = 3000$)	*82.86*
LEXICAL + PROD-RULE[Stanford+C&J+Berkeley]* ($n = all$)	*82.86*
LEXICAL + RERANKING ($n = all$)	76.00

** Rules lexicalised with function words are included*

Table 4.18: Confusion matrix based on the best performing feature set for production rules from all the three parsers — on the held-out set

	BL	CZ	RU	FR	SP	CN	JP
BL	[17]	4	3	1	–	–	–
CZ	2	[21]	2	–	–	–	–
RU	1	4	[17]	3	–	–	–
FR	1	–	–	[24]	–	–	–
SP	2	1	3	2	[17]	–	–
CN	–	–	–	–	–	[24]	1
JP	–	–	–	–	1	2	[22]

BL:Bulgarian; CZ:Czech; RU:Russian; FR:French; SP:Spanish; CN:Chinese; JP:Japanese

compactly than by just using feature concatenation. Another plausible explanation could be that the current training data is not sufficient to support such a large feature set and this might have caused the models to overfit.

4.3.4 Discussion

Confusion matrices. As illustrated in the confusion matrices (Table 4.18 for the PROD-RULE model, and Table 4.19 for the LEXICAL model), misclassifications occur largely in Spanish and Slavic languages, Bulgarian and Russian in particular. Unsurprisingly, Chinese is almost perfectly identified, coming as it does from an entirely different language family, Sino-Tibetan, as compared to the rest of the languages which are from the branches of the Indo-European family (with Japanese as the exception). Japanese and French also appear to be easily identified, which could perhaps be attributed to their word order or sentence structure which are, to some extent, quite different from English. Japanese is a 'subject-object-verb' language; and French, although having

Table 4.19: Confusion matrix based on lexical features — on the held-out set

	BL	**CZ**	**RU**	**FR**	**SP**	**CN**	**JP**
BL	[14]	3	4	2	2	–	–
CZ	6	[16]	2	–	1	–	–
RU	3	2	[16]	3	1	–	–
FR	1	–	–	[24]	–	–	–
SP	1	2	1	3	[17]	–	1
CN	–	–	–	–	–	[24]	1
JP	–	–	–	–	1	3	[21]

BL:Bulgarian; CZ:Czech; RU:Russian; FR:French; SP:Spanish; CN:Chinese; JP:Japanese

the same word order as English, heads of phrases in French typically come before modifiers as opposed to English. Overall, the PROD-RULE model results in fewer misclassifications as compared to the LEXICAL model and demonstrates significant reduction in confusion in the Slavic languages.

Data analysis. We looked at some of the data, to see what kinds of syntactic substructure (characterised by parse production rules) are useful in classifying native languages. Although using feature selection with only 1000 features did not improve classification performance, the information gain (IG) ranking does identify particular constructions as characteristic of one of the languages, and so is useful for inspection.

A phenomenon that the literature has noted as occurring with Chinese speakers is that of the inappriopriate use of determiners (Swan and Smith, 2001).[16] As mentioned in Section 3.1, such a phenomenon reflects misuse of determiners both in situations where a determiner is missing when it is required and also in situations where an extra determiner is used when it is indeed redundant. Consequently, this also relates to problems of determiner-noun agreement dealing with singular and plural nouns. If these phenomena are observed in our data, this could correspond to the distributional difference in NP rules involving `NN`, `NNS`, and `DT`, such as `NP` → `NN NN`, `NP` → `NN NNS`, `NP` → `DT NN NN` and `NP` → `DT NN NNS`. This set of rules may be valid in some other contexts but are also possibly indicators of inappropriate usage of certain syntactic constructions (as for the case of determiner misuse). Looking at the distribution of NP rules presented in Table 4.20 (only for rules taken from the top 1000 with highest IG), the absolute frequencies do not seem to provide a clear picture of distributional differences across the seven native languages. However, comparing the proportion of NP rules involving `NN` with respect to the total number

[16] This does happen with native speakers of some other languages, such as the Slavic ones, but not generally (from our knowledge of the literature) with native speakers of others, such as the Romance ones.

Table 4.20: Stanford rules of NP from the Top 1000 IG list

Stanford Rules	Absolute Frequency						
	BL	**CZ**	**RU**	**FR**	**SP**	**CN**	**JP**
NP → NNP NNP NNS	0	3	0	1	0	31	0
NP → DT NNS CC NNS	8	10	13	8	12	41	7
NP → JJR NN NNS	1	0	0	1	1	18	1
NP → DT NN NNS	17	17	14	15	11	44	12
NP → NNP NNP NN	1	1	1	0	0	12	1
NP → NN NN	32	39	41	22	34	53	44
NP → NN , NNS CC NN NN	0	0	0	0	0	8	0
NP → JJR NN NN	0	0	0	0	0	8	1
NP → NN NN NN	0	3	3	2	2	14	5
NP → NN NN NNS	2	3	4	1	0	12	1
NP → JJ NN NNS	15	15	21	9	8	27	14
NP → DT NN NN	40	38	43	44	44	59	41
NP → JJR NN	23	19	16	13	19	29	11
NP → DT JJS NN	15	26	18	11	22	10	15
NP → DT NN NN NNS	1	0	0	0	1	5	2
NP → DT NN NN NN	2	6	2	1	4	9	8
total rules of NP	157	180	176	128	158	380	163
total rules with NN	113	132	124	93	125	202	126
total rules with NNS	44	48	52	35	33	178	37
% rules with NN	0.719	0.733	0.704	0.726	0.791	*0.536*	0.773
total rules with DT	75	90	77	72	82	158	78
total rules without DT	82	90	99	56	76	222	85
% rules with DT	0.4777	0.500	0.437	0.562	0.519	*0.416*	0.478

of NP rules used, Chinese appears to have a much lower proportion in contrast to the other native languge groups. Similarly, the same situation is observed on the proportion of NP rules with DT: Chinese stands out amongst the rest. Taking these observations altogether, Chinese speakers seem to be using singular and plural noun phrases quite differently from other L1 speakers, perhaps reflecting problems with determiner-noun agreement and using determiners inappropriately. We give an example of this sort of phenomenon in Figure 4.4, taking the parse (returned by the Stanford parser) of the sentence *The development of* ***country park*** *can directly elp to alleviate overcrowdedness and overpopulation in urban area*. The phrase *country park* (corresponds to NP → NN NN) should either have a determiner or be plural (in which case the appropriate rule would be NP → DT NN NN or NP → NN NNS). There is a similar phenomenon with the phrase *in urban area*, although this is an instance of the rule NP → JJ NN.

The above phenomenon may be an example of one of the four measures of cross-linguistic influence based on the comprehensive survey of second language acquisition (SLA) by Ellis (2008) and discussed in Section 2.3.1.2, regarded as *negative transfer*. Given that the absence of deter-

```
(ROOT
  (S
    (NP
      (NP (DT The) (NN development))
      (PP (IN of)
          (NP (NN country) (NN park))))
    (VP (MD can)
        (ADVP (RB directly))
        (VP (VB elp)
            (S
              (VP (TO to)
                  (VP (VB alleviate)
                      (NP (NNS overcrowdedness)
                          (CC and)
                          (NN overpopulation))
                      (PP (IN in)
                          (NP (JJ urban)
                              (NN area))))))))
    (. .)))
```

Figure 4.4: An example of parse from Chinese-speaking authors illustrating missing determiners (also with a typographical error)

```
(ROOT
  (S
   (NP (NNP Hong) (NNP Kong) (NNS people))
   (VP (AUX are)
     (ADJP (JJ non-sensable)
        (PP (IN on)
          (S
            (VP (VBG recycling)
              (SBAR (WHNP (WP who))
                (S
                  (VP (AUX are)
                    (ADJP (JJ rich)))))))))
   (. .)))
```

Figure 4.5: An example of parse from Chinese-speaking authors illustrating difficulties with WH-clauses (also with a typographical error)

miners (articles) in Chinese, this has caused difficulties in article usage amongst Chinese learners of English.

In addition, the cross-linguistic phenomenon of *avoidance* is also noticeable. To minimise errors, learners tend to avoid using certain grammatical aspects if lacking in confidence. We observed that Chinese speakers use a relatively lower number of WH-words (*who* and *which* in particular) as characterised by the Stanford rules of `WHNP` $\rightarrow$ `WP` and `WHNP` $\rightarrow$ `WDT` (occurred in the top 1000 rules with highest IG) respectively in their texts in contrast to other native lan-

Table 4.21: Stanford rules of WHNP and WHPP from the Top 1000 IG list

Stanford Rules	Absolute Frequency						
	BL	**CZ**	**RU**	**FR**	**SP**	**CN**	**JP**
WHNP → WP	67	69	66	61	67	*37*	61
WHNP → WDT	67	68	66	67	70	*49*	58
SBAR → WHPP S	25	17	17	27	25	*2*	9
WHPP → IN WHNP	25	16	17	28	24	*2*	10

Table 4.22: WH-word (and relative clause marker) usage across seven native language groups

Native Languages	Relative Frequency						
	who	*which*	*whose*	*whom*	*when*	*where*	*that*
Bulgarian	0.0021	0.0034	0.0002	3.7819E-05	0.0020	0.0010	0.0132
Czech	0.0021	0.0027	3.4155E-05	8.5388E-05	0.0016	0.0007	0.0118
Russian	0.0032	0.0030	0.0002	9.0940E-05	0.0016	0.0006	0.0136
French	0.0018	0.0039	0.0002	5.9216E-05	0.0012	0.0008	0.0117
Spanish	0.0034	0.0042	0.0001	*2.0291E-05*	0.0023	0.0007	0.0141
Chinese	*0.0011*	*0.0022*	*2.2016E-05*	2.2016E-05	*0.0012*	*0.0002*	*0.0113*
Japanese	0.0023	0.0023	8.0413E-05	0.0001	0.0023	0.0003	0.0118

guage groups (refer to Tables 4.21 and 4.22). This observation is in line with the findings of Schachter (1974): although Chinese learners (and also Japanese in her work) made few errors in relative clause usage compared to learners of Arabic and Persian, they actually used a much lower number of relative clauses (i.e. they are avoiding using such a construction). This is obviously a phenomenon of avoidance. Furthermore, given that Chinese is a WH-*in-situ* language while English is a language with overt WH-movements, this also poses a related problem to native Chinese speakers when dealing with WH-clauses in English. As seen in Figure 4.5, the grammatical counterpart of the given sentence should be something like *Hong Kong people* ***who are rich*** *are not sensible on recycling*.

Another production rule that occurs typically — in fact, almost exclusively — in the texts of native Chinese speakers is PP → VBG PP (by the Stanford parser), which almost always corresponds to the phrase *according to*. In Figure 4.6, we give the parse of a short sentence (*According to <R>, burning of plastic waste generates toxic by-products.* — <R> is an in-text citation that was removed in our preparation of ICLE)[17] that illustrates this particular construction. It appears that speakers of Chinese frequently use this phrase as a translation of *gēn jù*. So in this case, what is identified is perhaps not the sort of error that is of interest to contrastive analysis,

[17] All parse rules lexicalised with the symbol <R> were excluded as classification features to avoid biases towards a specific native language, Chinese in particular. This was noted by Joel Tetreault and Daniel Blanchard (personal communication).

```
(ROOT
  (S
    (PP (VBG According)
        (PP (TO to)
            (NP (NNP <R>))))
    (, ,)
    (NP
       (NP (NN burning))
       (PP (IN of)
           (NP (JJ plastic)
               (NN waste))))
    (VP (VBZ generates)
        (NP (JJ toxic)
            (NNS by-products)))
    (. .)))
```

Figure 4.6: An example of parse from Chinese-speaking authors illustrating *according to*

but just a particular construction that is idiosyncratic of a certain native speaker's language, one that is perfectly grammatical but which is used relatively infrequently by others and has a slightly unusual analysis by the parser. This could be considered as an instance of *overuse* in the SLA sense.

Having seen a number of examples from Chinese speakers, we move on to examine the Indo-European language group. It was suggested in Swan and Smith (2001) that French speakers do encounter difficulties with English article usage but in a rather different manner compared to Chinese speakers. A typical error produced by French speakers as a result of the article usage in French is found in ***The English*** *is a difficult language*, where proper nouns are preceded with definite articles.[18] Looking into the ICLE corpus, we found that the texts produced by native French speakers demonstrate somewhat higher number of production rules involving DT (determiners) and NNP (proper nouns) such as `NP -> DT NNP` and `NP -> DT NNP NNP`. More concretely, Figure 4.7 presents a parse example extracted from a text of a native French speaker illustrating such a case.

Overall, we had expected to see more rules that displayed obvious ungrammaticality; for instance, `VP` $\rightarrow$ `DT IN`. However, as for grammaticality judgement in Section 4.2, the Stanford and C&J parsers particularly appear to be good at 'ignoring' errors, and producing relatively grammatical structures (albeit ones with different frequencies for different native languages). Figure 4.8 gives the C&J parse for *Overall, cyber* ***cafeis*** *a good place as recreational centre with a bundle of* ***up-to-dated*** *information*. The correction of *up-to-dated* rather than *up-to-date* is straightforward, but the simple typographical error of running together *cafe* and *is* leads to more complex problems

[18] Spanish speakers were seen to produce similar article errors in their English usage.

```
(ROOT
  (S
    (NP (DT Another) (NN example))
    (VP (VBZ is)
      (NP
        (NP (DT the) (NN worker))
        (PP (IN of)
          (NP
            (NP (DT an) (NN iron)
              (CC and)
              (NN steel) (NN industry))
            (PP (IN near)
              (NP (DT the) (NNP Baïkal)))))))
    (. .)))
```

Figure 4.7: An example of parse from French-speaking authors, illustrating errors in English article usage

```
(S1
  (S
    (ADVP (RB Overall))
    (, ,)
    (NP (NNP cyber))
    (VP (VBD cafeis)
        (NP (DT a) (JJ good) (NN place))
        (PP (IN as)
            (NP (JJ recreational)
                (NNP centre)))
        (PP (IN with)
            (NP
              (NP (DT a) (NN bundle))
              (PP (IN of)
                  (NP (JJ up-to-dated)
                      (NN information))))))
    (. .)))
```

Figure 4.8: An example of parse from C&J parser, illustrating parser correction

for the parser. Nevertheless, the parser produces a solid grammatical tree, specifically assigning the category VBD to the compound *cafeis*. This would support our previous suggestion that this is because both the Stanford and C&J parsers have implicit linguistic constraints such as assumptions about heads; these are imposed even when the text does not provide evidence for them.

We also present in Table 4.23 the top 10 production rules chosen under the IG feature selection for the Stanford parser on the held-out test set. A number of these are concerned with punctuation: these seem unlikely to be related to native language, but perhaps rather to how English learners of a particular language background are taught. Others are more typical of the sorts of example we illustrated above: PP → VBG PP, for instance, is typically connected to the *according to* con-

Table 4.23: Top 10 production rules for the Stanford parser based on Information Gain — on the held-out set

Stanford Rules	Absolute Frequency						
	BL	CZ	RU	FR	SP	CN	JP
: → -	55	51	39	23	10	9	4
PRN → -LRB- X -RRB-	0	1	2	7	0	42	0
: → :	30	39	46	58	47	11	6
X → SYM	0	2	4	7	4	42	6
NP → NNP NNP NNS	0	3	0	1	0	31	0
S → S : S .	36	34	39	53	41	5	9
PP → VBG PP	9	15	12	16	13	54	13
: → ...	16	13	11	39	24	1	3
NP → NNP NNP	23	31	29	22	27	65	25
S → ADVP, NP VP .	48	27	31	45	36	65	60

Table 4.24: Top 10 fine-grained production rules for the Berkeley parser based on Information Gain — on the held-out set

Berkeley Rules	Absolute Frequency						
	BL	CZ	RU	FR	SP	CN	JP
NP-56 → NNP-53 NNP-2	0	1	0	2	0	44	0
NP-53 → NNP-53 NNP-2	0	1	0	1	0	44	2
:-3 → -	52	47	34	20	9	7	3
PP-21 → VBG-3 PP-11	4	3	7	7	5	51	9
WHNP-1 → WP-0	58	61	51	48	52	6	48
NP-30 → NNP-33	2	1	4	1	0	37	1
:-0 → :	28	38	37	53	45	9	6
PP-11 → TO-0 NP-44	1	1	3	1	2	36	3
@S-24 → S-5 :-1	20	16	34	45	33	1	10
@S-23 → S-5 :-0	17	16	18	39	27	0	2

struction discussed in connection with Figure 4.6, and it can be seen that the dominant frequency count there is for native Chinese speakers (column 6 of the counts in Table 4.23).

We compare this with the fine-grained production rules from the Berkeley parser, looking at the top selected rules of Berkeley (Table 4.24). The rule frequency ratio between the seven languages is greater as compared to those from Stanford (as shown in Table 4.23). The improvement in classification performance suggests that augmenting coarse-grained production rules with latent variables (such as those of Berkeley) can model characteristics of individual native languages more closely.

Recent related work. Before concluding this section, one other aspect that worth mentioning is the very recent work of Swanson and Charniak (2012), which has extended from our initial work of this section published in Wong and Dras (2011). Swanson and Charniak (2012) replicated our

work using CFG rules as features for the same NLI task on ICLE texts from the same seven native language groups. Their reported accuracy for CFG rules derived from a single parser (72.6% using Berkeley parser) appears to slightly underperform in comparison with our best result presented in Wong and Dras (2011) (74.49% for Stanford parser under 5-fold cross validation; the revised result here is 74.29% as shown in Table 4.14). (Note that this still holds even if compared with our current best result for Berkeley parser of 75.71%). Such difference in performance is possibly due to the sample selection (as claimed by Swanson and Charniak (2012)) as their result is based on the averaged score over 5 subsamplings of the full data set of ICLE for each native language group; while we only used a single sampling with 5-fold cross validation. It seems that there is some degree of variance across the sub-corpora in ICLE. However, through the proposed features of using rules based on Tree Substitution Grammar (TSG), Swanson and Charniak (2012) have demonstrated that TSG rules appear to perform better than CFG rules by a margin of around 6% in terms of accuracy.

Summary. In this section we have shown that, using cross-sections of parse trees, we can improve above an already good baseline in the task of native language identification. While we do not make any strong claims for the Contrastive Analysis Hypothesis (CAH) or subsequent work in SLA on cross-linguistic effects, the usefulness of *syntax* in the context of this problem does provide some support.

The best features arising from the classification have been horizontal cross-sections of trees (i.e. parse rules), rather than the more general discriminative parse reranking features that might have been expected to perform at least as well. This relatively poorer performance by the reranking features may be due to a number of factors, all of which could be investigated in future work. One is the use of feature schema instances that did not appear in the largely grammatical WSJ; another is the extension of feature schemas to suit our specific problem.

Examining some of the substructures showed some errors that were expected; and also some other constructions that were grammatical, but were just characteristic translations of constructions that were common in the native language. Utilising more fine-grained rules appears to be useful in enhancing the performance of the classification task. Overall, the use of all these led to an error reduction in over 40% in the cross-validation evaluation with significance testing.

4.4 Concluding Remarks

In this chapter, we have demonstrated that syntactic structures — in particular those that are characterised by horizontal slices of parse trees (i.e. production rules) — are not only useful in terms

of sentence-based grammaticality judgement, but can be beneficial to NLI tasks in identifying the native language of the writers. Furthermore, classification performance can be improved through more fine-grained production rules augmented with latent variables, as are generated by the Berkeley parser.

Reranking features directly extracted from the C&J parser, on the other hand, appear to be helpful only for the simpler task of judging sentence grammaticality but not so much for the more specific classification task of NLI. The feature schemas used in the C&J parser were initially identified specifically for the purpose of improving the Charniak parser. A possible improvement here would be to explore instances of feature schemas produced after retraining the C&J parser with some ungrammatical data, such as the noisy WSJ treebank from Foster (2007). Alternatively, one can extend the existing feature schemas to specifically suit the purpose of NLI, based on the extensive literature from the field of second language acquisition.

The results from the Berkeley parser, however, suggest that an approach that is less constrained by linguistic expectations than parsers, and is more purely distributional, would be interesting to pursue. We propose one such approach in the next chapter.

Chapter 5

Exploring Topic Modeling

Chapter 4 particularly explored syntactic structures, in the form of production rules and parse reranking templates. Two of the conclusions were that: (1) the expected instances of clearly ungrammatical elements of syntactic structure indicating non-native speaker errors were not largely observed, instead there were often just different distributions over regular elements of grammatical structure for different native languages; and (2) it might be useful to avoid imposing linguistic constraints.

Therefore, we investigate two ideas in this chapter. Firstly, our first intuition is that it is several elements together that indicate particular kinds of indicative errors, such as incorrect noun-number agreement; and from this, that there might be coherent clusters of correlated features that are indicative of a particular native language. Secondly, we intend to revisit n-gram features as a form of approximation to syntactic structures through selection of collocational 'quasi-syntactic' n-grams of arbitrary length. To realise these ideas, we adopt a Bayesian approach to the task of native language identification (NLI) through *topic modeling* which can be exploited for the purpose of both feature dimensionality reduction and feature selection.[1]

Latent Dirichlet Allocation (LDA) — a generative probabilistic model for unsupervised learning — was first introduced by Blei et al. (2003) to discover a set of latent mixture components known as *topics* which are representative of a collection of discrete data. The underlying idea of LDA is that each document from a text corpus is constructed according to a specific distribution of topics, in which (in this first application of LDA to texts) words comprising the document are generated based on the word distribution for each selected topic; a topic is typically represented by a set of words such as *species*, *phylogenetic*, *evolution* and so on. Such a model allows multiple

[1] The two studies presented in this chapter had been published in Wong et al. (2011) and Wong et al. (2012), respectively. A special thanks to Mark Johnson, who provided much guidance on the topic modeling work conducted in this chapter, especially on the overall formulation of the adaptor grammars.

topics in one document as well as sharing of topics across documents within a corpus.

LDA can be viewed as a form of feature dimensionality reduction technique. As the first key task investigated in this chapter, we intend to exploit LDA to discover the extent to which a lower dimension of feature space — a set of potentially useful coherent clusters of features (or *latent factors*) — in each document affects the classification performance of NLI. Here we are mapping clusters of features as 'topics' in typical LDA models and the posterior topic distributions inferred are to be used for classifying the native language of the authors against baseline models using the actual features themselves. We are particularly interested in whether the topics appear at all to form coherent clusters, and consequently whether they might potentially be applicable to the much larger class of syntactic features.

As the second key task of this chapter, our intention is to investigate the usefulness of n-grams as features for NLI classification. As seen in Koppel et al. (2005), PoS n-grams can be leveraged to characterise surface syntactic structures, in which they approximated ungrammatical structures with rare PoS bi-grams. Hirst and Feiguina (2007) also proposed replacing syntactic rules with PoS sequences (in the form of bi-grams) on the grounds of data sparsity for the more general problem of authorship analysis (see discussion in Section 2.1.1 under syntactic features). However, to capture sequences of patterns that are characteristic of a particular native language, small n-grams like bi-grams or tri-grams might not suffice. To address this, *adaptor grammars* (Johnson, 2010) which are interpretable as an extension to the basic LDA models, can be helpful as they are capable in discovering useful collocational n-grams of arbitrary length. For example, in his initial work, Johnson's model learnt collocations such as *gradient descent* and *cost function*, under a topic associated with 'machine learning'. Hardisty et al. (2010) subsequently applied this idea to perspective classification, learning collocations such as *palestinian violence* and *palestinian freedom*, the use of which as features was demonstrated to help the classification of texts from the Bitter Lemons corpus as either Palestinian or Israeli perspective. In our context, instead of typical collocations of (content) words, we are interested in 'quasi-syntactic collocations' of either pure PoS (e.g. `NN IN NN`) or a mixture of PoS with function words (e.g. `NN of NN`).[2] Thus, a particular question of interest for this second key task is to investigate whether the power of adaptor grammars to discover collocations extend to features beyond the purely lexical — and whether in turn we can discover ones of arbitrary length that are useful for NLI.

The organisation of the chapter is as follows. In Section 5.1, we introduce the fundamental concepts of LDA and its applications to classification in general. We then present the first key task of this chapter in Section 5.2, demonstrating how LDA models can be used for feature dimension-

[2]Typically in NLI and other authorship attribution tasks, the feature sets exclude content words, to avoid unfair cues due to potentially different domains of discourse. (See the discussion in Section 2.2).

ality reduction to uncover clusters of features that are helpful in the classification task of NLI. In Section 5.3, we review the mechanics of adaptor grammars and the view of them as an extension of LDA. This will then be followed by the second key task of this chapter in Section 5.4, where we describe two appproaches in which adaptor grammars can be deployed for feature selection by finding useful n-gram collocations for the purpose of NLI.

5.1 Latent Dirichlet Allocation

In its basic form, Latent Dirichlet Allocation (LDA) is a Bayesian probabilistic model used to represent collections of discrete data such as text corpora, introduced by Blei et al. (2003). It addressed the limitations of earlier techniques such as *probabilistic latent semantic indexing*, which is prone to overfitting and unable to generalise to unseen documents. LDA is a relaxation of classical document mixture models in which each document is associated with only a single topic, as it allows documents to be generated based on a mixture of topics with different distributions. We discuss the basic mechanisms of LDA and our particular representation, in Section 5.2.1.

LDA has been applied to a wide range of tasks, such as building cluster-based models for ad-hoc information retrieval (Wei and Croft, 2006) or grounded learning of semantic parsers (Börschinger et al., 2011). Relevant to our studies on NLI, it has been applied to a range of text classification tasks.

The original paper of Blei et al. (2003) used LDA as a dimensionality reduction technique over word unigrams for an SVM, for genre-based classification of Reuters news data and classification of collaborative filtering of movie review data, and found that LDA topics actually improved classification accuracy in spite of the dimensionality reduction. This same basic approach has been taken with other data, such as spam filtering of web texts (Bíró et al., 2008), where LDA topics improved classification f-measure, or finding scientific topics from article abstracts (Griffiths and Steyvers, 2004), where LDA topics appear to be useful diagnostics for scientific subfields.

It has also been augmented in various ways: supervised LDA, where topic models are integrated with a response variable, was introduced by Blei and McAuliffe (2008) and applied to predicting sentiment scores from movie review data, treating it as a regression problem rather than a classification problem. Work by Wang et al. (2009) followed from that, extending it to classification problems, and applying it to the simultaneous classification and annotation of images. An alternative approach to joint models of text and response variables for sentiment classification of review texts (Titov and McDonald, 2008), with a particular focus on constructing topics related to aspects of reviews (e.g. food, decor, or service for restaurant reviews), found that LDA topics were predictively useful and seemed qualitatively intuitive.

In all of this preceding work, a document to be classified is represented by an exchangeable set of (content) words: function words are generally removed, and are not typically found in topics useful for classification. It is exactly these that are used in NLI, so the above work does guarantee that an LDA-based approach will be helpful here.

Three particularly relevant pieces of work on using LDA in classification are for the related task of authorship attribution, determining which author wrote a particular document. Rajkumar et al. (2009) claim that models with stopwords (function words) alone are sufficient to achieve high accuracy in classification, which seems to peak at 25 topics, and outperform content word-based models; the results presented in Table 2 of that paper and the discussion are, however, somewhat contradictory. Seroussi et al. (2011) also include both function words and content words in their models; they find that filtering words by frequency is almost always harmful, suggesting that function words are helping in this task.[3] In very recent work, Seroussi et al. (2012) modified the existing author-topic model of Rosen-Zvi et al. (2004) to suit the purpose of authorship attribution and demonstrated the combination of both the basic LDA and author-topic models outperforms each of the individual models as well as SVM-based classification.

5.2 Feature Dimensionality Reduction

One key phenomenon observed in Chapter 4 was that there were different proportions of parse production rules indicative of particular native languages. One example is the production rules of NP, such as `NP` $\rightarrow$ `NN NN`, and `NP` $\rightarrow$ `NN NNS`, in which the NP usage of Chinese appears to be quite different from other native language groups; this is likely to reflect determiner-noun agreement errors, as the latter (`NP` $\rightarrow$ `NN NNS`) is used at the expense of the former which is headed by a singular noun (`NP` $\rightarrow$ `NN NN`). There might be coherent clusters of related features, with these clusters characterising typical errors or idiosyncrasies, that are predictive of a particular native language. As the first task investigated in this chapter, we use the fundamental model of LDA to cluster coherent features as a form of latent factors, to investigate whether clustering shows any potential for our classification task for NLI. Here, we intend to explore the basic lexical features — both function words and PoS n-grams — the latter of which is quite novel to our knowledge in terms of classification using LDA.

[3] They note that for function words the term 'latent factor' is more appropriate than 'topic', with its connotation of semantic content.

5.2.1 Mechanics of LDA

5.2.1.1 General Definition

Formally, each document is formed from a fixed set of vocabulary V and fixed set of topics T ($|T| = t$). Following the characterisation given in Griffiths and Steyvers (2004), the process of generating a corpus of m documents is as follows: (1) first generate a set of multinomial distributions over topics $\boldsymbol{\theta}_j$ for each document D_j according to a T-dimensional Dirichlet distribution with concentration parameter $\boldsymbol{\alpha}$ (i.e. $\boldsymbol{\theta}_j \sim \text{Dir}(\boldsymbol{\alpha})$); (2) then generate a set of multinomial distributions $\boldsymbol{\phi}_i$ over the vocabulary V for each topic i according to a V-dimensional Dirichlet distribution with concentration parameter $\boldsymbol{\beta}$ (i.e. $\boldsymbol{\phi}_i \sim \text{Dir}(\boldsymbol{\beta})$); and (3) finally generate each of the n_j words for document D_j by selecting a random topic z according $\boldsymbol{\theta}_j$ and then drawing a word $w_{j,k}$ from $\boldsymbol{\phi}_z$ of the selected topic. The overall generative probabilistic model can be summarised as below:

$$
\begin{aligned}
&\boldsymbol{\theta}_j \sim \text{Dir}(\boldsymbol{\alpha}) && j \in 1, ..., m \\
&\boldsymbol{\phi}_i \sim \text{Dir}(\boldsymbol{\beta}) && i \in 1, ..., t \\
&z_{j,k} \sim \boldsymbol{\theta}_j && j \in 1, ..., m, k \in 1, ..., n_j \\
&w_{j,k} \sim \boldsymbol{\phi}_{z_{j,k}} && j \in 1, ..., m, k \in 1, ..., n_j
\end{aligned}
$$

Hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are Dirichlet priors, which are used as a form of smoothing to smooth the topic and word distributions where the amount of smoothing is determined by the values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. In cases where $\boldsymbol{\alpha}$ ($\boldsymbol{\beta}$) < 1, these are known as sparse (weak) priors, in which most of the probability mass is concentrated on only a small number of topics (words). This type of distribution is appropriate for modeling human languages that obey the power laws. It was suggested by Steyvers and Griffiths (2007) that $\boldsymbol{\alpha} = 50/t$ and $\boldsymbol{\beta} = 0.01$ work well for most of the text collections in previous research.

From the inference perspective, given a corpus of m documents with n_j words each, the task is to estimate the posterior topic distributions $\boldsymbol{\theta}_j$ for each document D_j as well as the posterior word distributions $\boldsymbol{\phi}_i$ for each topic i that maximise the log likelihood of the corpus. Knowing that the exact inference of these posterior distributions is intractable in general, there is a wide variety of means of approximate inference for LDA models which include approximation algorithms such as *Variational Bayes* (Blei et al., 2003) and expectation propagation (Minka and Lafferty, 2002) as well as the *Markov Chain Monte Carlo* approach with Gibbs sampling (Griffiths and Steyvers, 2004).

5.2.1.2 LDA in PCFG Formulation

Johnson (2010) demonstrated that LDA topic models can be regarded as a specific type of probabilistic context-free grammar (PCFG), and that Bayesian inference for PCFGs can be used to learn LDA models where the inferred distributions of PCFGs correspond to those distributions of LDA. A general schema used for generating PCFG rule instances for representing m documents with t topics is as follows:[4]

$$
\begin{array}{ll}
Sentence \rightarrow Doc'_j & j \in 1, ..., m \\
Doc'_j \rightarrow \_j & j \in 1, ..., m \\
Doc'_j \rightarrow Doc'_j \ Doc_j & j \in 1, ..., m \\
Doc_j \rightarrow Topic_i & i \in 1, ..., t; j \in 1, ..., m \\
Topic_i \rightarrow w & i \in 1, ..., t; w \in V
\end{array}
$$

Each of the rules in the PCFG is associated with a Bayesian inferred probability. While the probabilities associated with the rules expanding $Topic_i$ correspond to the word distributions $\boldsymbol{\phi}_i$ of the LDA model, the probabilities associated with the rules expanding Doc_j correspond to the topic distributions $\boldsymbol{\theta}_j$ of LDA. Similarly, inference on the posterior rule distributions can be approximated with Variational Bayes and Gibbs sampling. We use this PCFG formulation of LDA in this work.

5.2.2 Experimental Models

This subsection describes both the LDA models and the corresponding classification models used for our native language identification task on the ICLE corpus (Version 2) (Granger et al., 2009). Following the experimental setup in the previous chapters, we use the same dataset which consists of 490 essays written by non-native English users from seven different groups of language background — namely, Bulgarian, Czech, French, Russian, Spanish, Chinese, and Japanese. There are 70 essay documents per native language.

Unlike the documents often inferred by LDA topic models which mostly consist of only content words, we represent our documents with *function words* instead, given that this is typical for authorship-related tasks, and does not allow unfair clues based on different distributions of domain discourse. In addition, we also experiment with documents represented by another type of lexical features for NLI, *PoS bi-grams*.

[4] It should be noted that each document is given with a document identifier in which sentences in the document are prefixed with $\_j$.

5.2.2.1 LDA Models for NLI

For each of the LDA models we describe below, we experiment with different numbers of topics, in which $t = \{5, 10, 15, 20, 25\}$. In terms of the total number of PCFG rules representing each LDA model, there are 490 of the first three rules as shown in the schema (Section 5.2.1.2), $490 \times t$ of the rule expanding $\texttt{Doc}_j \rightarrow \texttt{Topic}_i$, and $t \times v$ of the rule expanding $\texttt{Topic}_i \rightarrow \texttt{w}$ (see Table 5.1). All the inferences are performed with the PCFG-based Gibbs sampler implemented by Mark Johnson.[5]

FW-LDA models. The first variant of the LDA models is function word based. The vocabulary used for generating documents with this LDA variant is therefore a set of function words. We adopt the same set as used in the previous chapters which consists of 398 words. An instance of the PCFG rule expanding $\texttt{Topic}_i \rightarrow \texttt{w}$ is $\texttt{Topic}_1 \rightarrow \texttt{the}$. There are 398 such rules for each topic.

POS-LDA models. The second variant of the LDA models is PoS bi-gram based. We choose bi-grams as they have already been explored in Chapter 3 and have been shown to be useful to some degree. Similar to the procedures performed in Chapter 3, by tagging the 490 documents with the Brill tagger (using Brown corpus tags), we extract the 200 most frequent occurring PoS bi-grams to form the vocabulary for this LDA variant. An instance of the PCFG rule expanding $\texttt{Topic}_i \rightarrow \texttt{w}$ is $\texttt{Topic}_1 \rightarrow \texttt{NN\_NN}$. There are 200 such rules for each topic.

FW-POS-LDA models. The third variant of the LDA models combines the first two. We note that this is not typical of topic models: most form topics only over single types, such as content words.[6] The vocabulary then consists of both function words and PoS bi-grams with 598 terms in total, with 398 from the former and 200 from the latter. Therefore, there are 598 instances of the rule expanding $\texttt{Topic}_i \rightarrow \texttt{w}$ for each topic.

Note that for each of these variants of LDA models, there are five versions with a different number of topics respectively (as mentioned in the beginning of this subsection).

[5] Gibbs sampling software for PCFG-based topic modeling is available at `http://web.science.mq.edu.au/~mjohnson/Software.htm`.

[6] Those that include multiple types typically treat them in different ways, such as in the separate treatment of content words and movie review ratings of Blei and McAuliffe (2008).

Table 5.1: Number of PCFG rules for each LDA model with different number of topics t

LDA Models	$t = 5$	$t = 10$	$t = 15$	$t = 20$	$t = 25$
FW-LDA	5,910	10,350	14,790	19,230	23,670
POS-LDA	4,920	8,370	11,820	15,270	18,720
FW+POS-LDA	6,910	12,350	17,790	23,230	28,670

5.2.2.2 Classification Models for NLI

Here, we exploit LDA as a form of dimension reduction in feature space to discover clusters of latent factors (features) as represented by 'topics' for classification. Based on each of the LDA models inferred, we take the posterior topic distributions and use as features for classifying the documents into one of the seven native language groups. All the classifications are performed with one of the machine learners used in Chapter 4: maximum entropy — MegaM (fifth release) by Hal Daumé III.[7]

Baselines. Each LDA classification model (as described in the following) is compared against a corresponding baseline model. These sets of model use the actual features themselves for classification without feature reduction. There are three baselines: function word based with 398 features (FW-BASELINE), PoS bi-gram based with 200 features (POS-BASELINE), and the combination of the first two sets of features (FW+POS-BASELINE). For each of these models, we examine two types of feature value: binary value (presence or absence) and relative frequency.

Function words. Features used in this model (FW-LDA) are the topic distributions inferred from the first LDA model. There are five variants of this based on the different number of topics (Section 5.2.2.1). The feature values are the posterior probabilities associated with the PCFG rules expanding $\texttt{Doc}_j \rightarrow \texttt{Topic}_i$ which correspond to the topic distributions $\boldsymbol{\theta_j}$ of the LDA representation.

PoS bi-grams. Similarly, this set of classification models (POS-LDA) uses the topic probabilities inferred from the second variant of the LDA model as classification features. Five variants of this with respect to the different topic numbers are examined as well.

Combined Features The last set of models combine both the function words and PoS bi-grams as classification features. The feature values are then the topic probabilities extracted from the last variant of the LDA models (i.e. the combined FW+POS-LDA models).

[7] MegaM software is available at `http://www.umiacs.umd.edu/~hal/megam/`.

Table 5.2: Classification performance (%) of the three baseline models

Baselines	Binary	Relative Frequency
FW-BASELINE	*62.45*	33.26
POS-BASELINE	*53.87*	45.92
FW+POS-BASELINE	***64.08***	42.65

Feature types: binary value and relative frequency

5.2.2.3 Evaluation

Often, LDA models are evaluated in terms of *goodness-of-fit* of the model to new data, by estimating the *perplexity* or similar of unseen held-out documents given some training documents (Blei et al., 2003; Griffiths and Steyvers, 2004). However, there are issues with all such proposed measures so far, such as importance sampling, harmonic mean, Chib-style estimation, and others; see Wallach et al. (2009) for a discussion. Alternatively, LDA models can be evaluated by measuring performance of some specific applications such as information retrieval and document classification (Titov and McDonald, 2008; Wang et al., 2009; Seroussi et al., 2011). We take this approach here, and adopt the standard measure for classification models — *classification accuracy* — as an indirect evaluation on our LDA models. For the same reason as presented in the previous chapters, the evaluation uses 5-fold cross-validation on our fairly small data set.

5.2.3 Classification Results

5.2.3.1 Baseline Models

Table 5.2 presents the classification accuracies achieved by the three baseline models mentioned above (i.e. using the actual features themselves without dimension reduction in feature space). These results are aligned with the results presented in Chapter 4 under the subsection of 4.3.1.1, in which binary feature values generally perform much better, although the results are lower because the calculation was made under cross-validation rather than on a separate held-out test set (hence with an effectively smaller amount of training data). Combining both the function words and PoS bi-grams yields a higher accuracy as compared to individual features alone. It seems that both features are capturing different useful cues that are predictive of individual native languages.

5.2.3.2 LDA Models

The classification performance for all the three variants of the LDA models under different parameter settings are presented in Tables 5.3 to 5.6. Three sets of concentration parameters (Dirichlet

Table 5.3: Classification performance (%) of each LDA-induced model — under the setting of uniform priors ($\boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta} = 1$)

LDA Models	$t = 5$	$t = 10$	$t = 15$	$t = 20$	$t = 25$
FW-LDA	44.89	*50.61*	44.29	47.14	49.59
POS-LDA	47.35	*51.02*	50.00	50.61	49.79
FW+POS-LDA	49.79	54.08	***55.51***	52.86	53.26

Feature type: topic probabilities

Table 5.4: Classification performance (%) of each LDA-induced model — under the setting of uniform priors ($\boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta} = 1$)

LDA Models	$t = 5$	$t = 10$	$t = 15$	$t = 20$	$t = 25$
FW-LDA	35.51	41.02	*42.65*	41.02	40.00
POS-LDA	40.20	43.88	47.35	*48.16*	44.28
FW+POS-LDA	42.65	46.33	***51.22***	49.39	44.49

Feature type: binary type with probability threshold $>= 0.1$

priors) were tested on each of the three LDA variants to find the best fitted topic model: Table 5.3 contains results for uniform priors $\boldsymbol{\alpha} = 1$ and $\boldsymbol{\beta} = 1$ (the default setting); Table 5.5 is for $\boldsymbol{\alpha} = 50/t$ and $\boldsymbol{\beta} = 0.01$ (as per Steyvers and Griffiths (2007)); and Table 5.6 is for $\boldsymbol{\alpha} = 5/t$ and $\boldsymbol{\beta} = 0.01$ (since for us, with a small number of topics, the $\boldsymbol{\alpha} = 50/t$ of Steyvers and Griffiths (2007) gives much larger values of $\boldsymbol{\alpha}$ than was the case in Steyvers and Griffiths (2007)). On the whole, weaker priors ($\boldsymbol{\alpha} = 5/t$ and $\boldsymbol{\beta} = 0.01$) lead to a better model as evidenced by the accuracy scores.

As observed in Table 5.3, the model with 10 topics is the best model under uniform priors for both the individual feature-based models (FW-LDA and POS-LDA) with accuracies of 50.61% and 51.02% respectively, while the combined model (FW+POS-LDA) performs best at 55.51% with 15 topics. It should be noted that these are the outcomes of using the topic probabilities as feature values. (We also investigated the extent to which binary feature values could be useful by setting a probability threshold at 0.1; however, the results are consistently lower as shown in Table 5.4.)

By setting a stronger $\boldsymbol{\alpha} = 50/t$ and a much weaker $\boldsymbol{\beta} = 0.01$, the resulting models perform no

Table 5.5: Classification performance (%) of each LDA-induced model — under the setting of stronger priors on $\boldsymbol{\alpha}$ ($\boldsymbol{\alpha} = 50/t$ and $\boldsymbol{\beta} = 0.01$)

LDA Models	$t = 5$	$t = 10$	$t = 15$	$t = 20$	$t = 25$
FW-LDA	32.45	42.45	44.29	45.71	*47.35*
POS-LDA	44.29	46.53	*50.82*	48.76	*50.82*
FW+POS-LDA	47.75	49.39	51.02	***54.49***	50.81

Feature type: topic probabilities

Table 5.6: Classification performance (%) of each LDA-induced model — under the setting of weaker priors on $\boldsymbol{\alpha}$ ($\boldsymbol{\alpha} = 5/t$ and $\boldsymbol{\beta} = 0.01$)

LDA Models	$t = 5$	$t = 10$	$t = 15$	$t = 20$	$t = 25$
FW-LDA	41.63	47.14	48.76	45.51	*52.45*
POS-LDA	43.47	49.79	51.22	52.86	*53.47*
FW+POS-LDA	51.84	50.61	53.88	52.62	***56.94***

Feature type: topic probabilities

better than those with uniform priors (see Table 5.5). The best performing models under this setting are with 25 topics for the individual feature-based models but with 20 topics for the combined model. This setting of priors was found to work well for most of the text collections as suggested in Steyvers and Griffiths (2007). However, given that our topic sizes are just within the range of 5 to 25, we also tried $\boldsymbol{\alpha} = 5/t$. The classification results based on $\boldsymbol{\alpha} = 5/t$ and $\boldsymbol{\beta} = 0.01$ are showed in Table 5.6. This setting leads to the best accuracy (thus far) for each of the models with 25 topics — FW-LDA (52.45%), POS-LDA (53.47%), FW+POS-LDA (56.94%). The overall trajectory suggests that more than 25 topics might be useful.

Overall, the classification performance for each of the LDA-induced models (regardless of the parameter settings) performs worse than the baseline models (Section 5.2.3.1) where the actual features were used, contra the experience of Rajkumar et al. (2009) in authorship attribution. The drop is, however, only small in the case of PoS tags; the overall result is dragged down by the drop in function word model accuracies. And comparatively, they are still well above the majority baseline of 14.29% (70/490), so the LDA models are detecting something useful. On the one hand, it is not surprising that reducing a relatively small feature space reduces performance; on the other hand, other work (as discussed in Section 5.1) had found that this had actually helped. While these results may not be conclusive — a more systematic search might find better values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ — the results of the POS-LDA model suggests some promise for applying the method to a much larger feature space of similar terms: this could either be the unrestricted set of PoS bi-grams, or of syntactic structure features. We investigate this further by looking more deeply at some of the 'topics' (i.e. latent factors) found in the discussion below.

5.2.4 Discussion

Despite the fact that all the LDA-induced models had lower accuracy scores than the baseline models, the inferred topics (clusters of related features) did demonstrate some useful cues that appear to be indicative of a particular native language. Here we present a discussion of three of these.

Table 5.7: Analysis on the best performing FW-LDA model for $Topic_8$

Language	Feature Weight	Relative Frequency of *the*
Bulgarian	*(relative to Bulgarian)*	0.0814
Czech	-0.0457	0.0648
French	0.2124	0.0952
Russian	0.0133	0.0764
Spanish	-0.0016	0.0903
Chinese	*3.2409*	*0.1256*
Japanese	0.4485	0.0661

Under the setting of 25 *topics with* $\boldsymbol{\alpha} = 5/t$ *and* $\boldsymbol{\beta} = 0.01$

Analysis of FW-LDA models. It is often noted in the literature on second language errors that a typical error of Chinese speakers of English is with articles such as *a*, *an*, and *the*, as Chinese does not have these. Looking at the best performing FW-LDA model (weak priors of $\boldsymbol{\alpha} = 5/t$ and $\boldsymbol{\beta} = 0.01$; 25 topics), we observed that for the three topics — $Topic_8$ (the 8th feature), $Topic_{19}$ (the 19th feature) and $Topic_{20}$ (the 20th feature) — each of these is associated with a much higher feature weight for Chinese as compared to other native language groups (Table 5.7 shows the analysis on $Topic_8$). As for the function words clustered under these topics, *the* appears to be the most probable one with the highest probabilities of around 0.188, 0.181, and 0.146 for each respectively (i.e. the PCFG rules of $\texttt{Topic}_8 \rightarrow \texttt{the}$, $\texttt{Topic}_{19} \rightarrow \texttt{the}$, and $\texttt{Topic}_{20} \rightarrow \texttt{the}$); this is a higher weighting than for any other word in any topic. To verify that the topic model accurately reflects the data, we found that the relative frequency of *the* in the documents produced by Chinese learners is the highest in comparison with other languages in our corpus. It seems that Chinese learners have a tendency to misuse this kind of word in their English constructions, overusing *the*: this parallels the example given in Chapter 4 and also noted in the beginning of Section 5.2, in which the different proportions of NP rules such as NP → NN NN, NP → NN NNS, and also NP → DT NN NN (rather than specifically ungrammatical constructions), are a characteristic of Chinese texts. However, there is no obvious pattern to the clustering (at least, that is evident to us) — if the clusters were to be grouping features in a way representative of errors, one of these topics might reflect misuse of determiners. But, none of these appear to: in $Topic_8$, for example, *a* appears only in 5th place, and no other determiners appear at all in the upper end of the distribution.

Analysis of POS-LDA models. On the other hand, there is a different story for POS-LDA, in terms of Chinese error phenomena. As shown in Table 5.8, Chinese has the highest feature weight for the first feature, $Topic_1$ (and also for $Topic_4$). To characterise this, we note that the PoS bi-gram NN_NN appears as the top bi-gram under $Topic_1$ ($\sim$0.18) (and also occurs most frequently

Table 5.8: Analysis on the best performing POS-LDA model for $Topic_1$

Language	Feature Weight	Relative Frequency of NN_NN
Bulgarian	*(relative to Bulgarian)*	0.0126
Czech	0.7777	0.0157
French	0.2566	0.0148
Russian	0.0015	0.0129
Spanish	0.0015	0.0142
Chinese	*2.4843*	*0.0403*
Japanese	0.4422	0.0202

Under the setting of 25 *topics with* $\boldsymbol{\alpha} = 5/t$ *and* $\boldsymbol{\beta} = 0.01$

among Chinese learners as compared to other native language groups). Further, the next four bigrams are NN_IN, AT_IN, IN_NN and NN_NNS, the last of which appears to be in complementary distribution in Chinese errors with NN_NN (i.e. Chinese speakers may tend to use the singular more often in compound nouns, when a plural might be more appropriate). At this point, we note the finding in Chapter 4 that usage of nouns was very distinctive for Chinese speakers. There, however, the finding was that Chinese speakers overused plural nouns relative to singular ones. A possible explanation is that there proportions were calculated over the top 1000 rules by IG: there are many other rules not included. It may be that the use of plural nouns by Chinese speakers in those constructions is particularly characteristic (and more common within those constructions), and still be consistent with an overall overuse of singular nouns. In any case, it is the distinctiveness of Chinese noun distribution that is noteworthy. $Topic_1$ thus seems to be somehow connected with noun-related errors.

Our second instance to look at in some detail is Japanese. Our expectation is that there are likely to be errors related to pronouns, as Japanese often omits them. In Ellis (2008)'s comprehensive survey of second language acquisition (also presented in Section 2.3.1.2), there are four measures of cross-linguistic influence: *error* (negative transfer), where differences between the languages lead to errors; *facilitation* (positive transfer), where similarities between the languages lead to a reduction in errors (relative to learners of other languages); *avoidance*, where constructions that are absent in the native language are avoided in the second language; and *overuse*, where constructions are used more frequently in an incorrect way in the second language, due to overgeneralisation.

A priori, it is difficult to predict which of these types of influence might be the case. The classic study of avoidance by Schachter (1974) examines Persian, Arab, Chinese, and Japanese learners of English, and their performance on using relative clauses (as mentioned in Section 2.3.1.2). It found that even though Persian and Arabic have similar (right-branching) relative clauses to English, and Japanese and Chinese have different (left-branching) ones, the Japanese and Chinese

Table 5.9: Analysis on the best performing POS-LDA model for $Topic_8$

Language	Feature Weight	Relative Frequency of `PPSS_VB`
Bulgarian	*(relative to Bulgarian)*	0.0111
Czech	0.7515	0.0137
French	-0.7080	0.0074
Russian	-0.2097	0.0116
Spanish	-0.3394	0.0117
Chinese	-0.1987	0.0059
Japanese	*2.0707*	*0.0224*

Under the setting of 25 *topics with* $\boldsymbol{\alpha} = 5/t$ *and* $\boldsymbol{\beta} = 0.01$

learners made fewer errors (see Table 2.1); but that that was because they avoided using the construction. On the other hand, for a grammatically less complex phenomenon such as article use, several studies such as those of Liu and Gleason (2002) show that there can be a developmental aspect to cross-linguistic influence, with initial errors or avoidance turning to overuse because of overgeneralisation, which is later corrected — reflecting the U-shaped learning behaviour; intermediate learners thus show the greatest level of overuse.

Looking at $Topic_8$ and $Topic_{20}$ under the POS-LDA model, relative to other topics inferred, top-ranking PoS bi-grams are mostly related to pronouns (such as `PPSS_VB`, `PPSS_MD`, and `PPSS_VBD`). Much higher feature weights are associated to these two topics for Japanese (as seen in Table 5.9 the analysis on $Topic_8$). Bi-grams of `PPSS_VB` and `PPSS_MD` occur much more often in Japanese learners' writings, and they are the first and the fifth terms under $Topic_8$, which seems to capture some of these phenomena.

To understand what these were saying about Japanese pronoun usage, we looked at a breakdown of pronoun use (see Table 5.10 and Figure 5.1 for better visualisation). Most apparently, the texts by Japanese speakers use more pronouns than any others. As the texts in the ICLE corpus are written by intermediate (to advanced) speakers, this could indicate a very strong instance of overuse. Looking at the distribution of pronouns, the Japanese speakers make much more use of the pronoun *I* than others: this has been noted elsewhere by Ishikawa (2011) on different corpora, particularly in the use of phrases such as *I think*. (The phrase *I think* is over-represented among Japanese speakers in our data also.)

Overall, then, POS-LDA seems to provide useful clustering of terms, while FW-LDA does not. This accords with the classification accuracies seen.

Analysis of FW+POS-LDA models. One question about the combined models was whether topics split along feature type — if that were the case, for a rough 2:1 ratio of function words to PoS bi-grams under 15 topics, there might be 10 topics whose upper rankings are dominated by

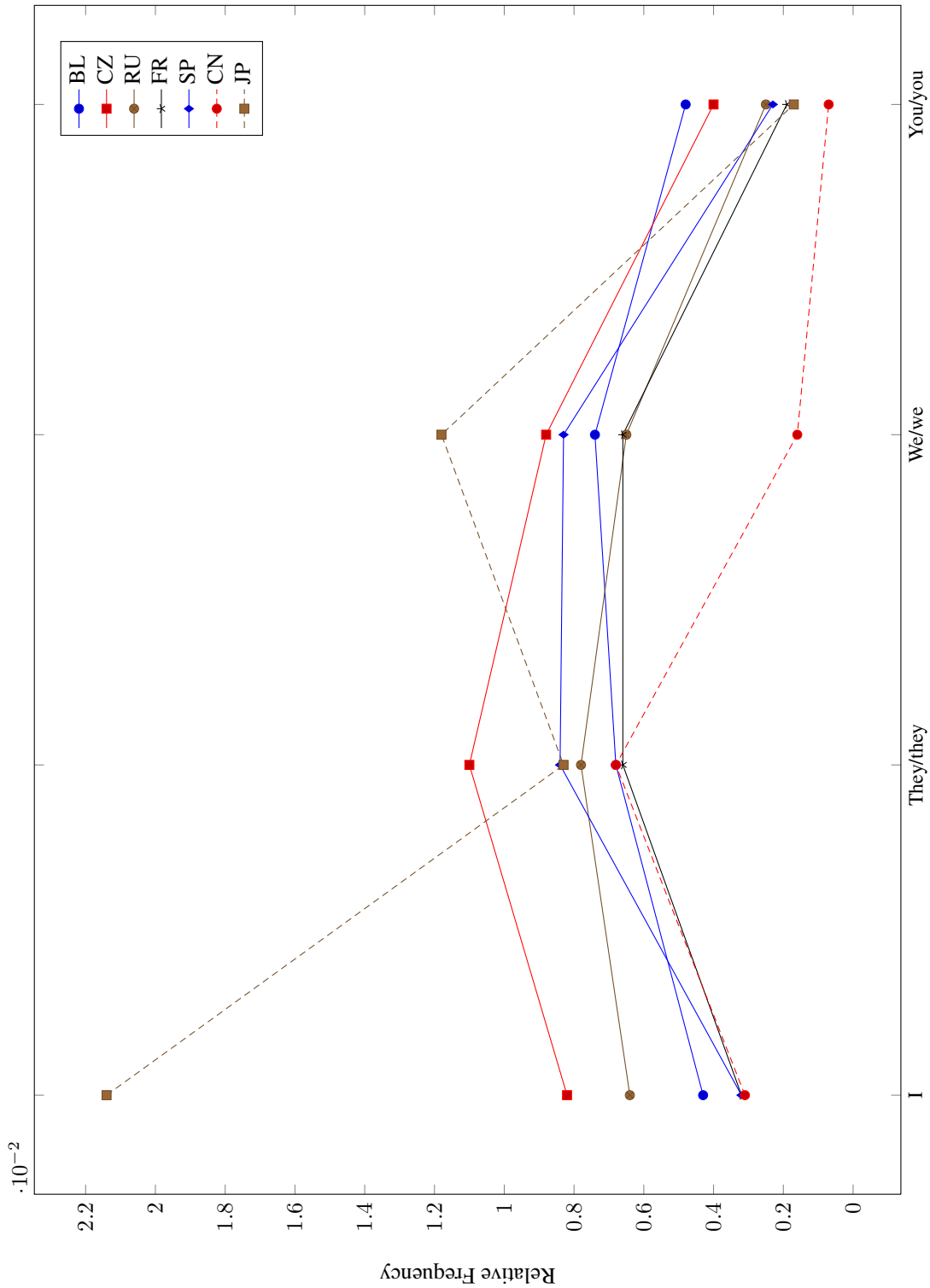


Figure 5.1: Pronoun usage across seven native language groups — words tagged with PPSS

Table 5.10: Pronoun usage across seven native language groups — words tagged with PPSS

Native	**Absolute Frequency**										
Languages	*I*	*They*	*Thou*	*We*	*You*	*it*	*she*	*they*	*we*	*you*	**Total**
Bulgarian	229	66	0	52	*38*	1	0	297	338	*219*	1240
Czech	483	*188*	0	*166*	34	1	0	*459*	348	202	1881
Russian	355	100	*1*	76	28	1	0	332	286	110	1289
French	161	55	0	71	4	2	0	282	261	90	926
Spanish	157	52	0	49	6	2	*1*	361	360	107	1095
Chinese	143	52	0	9	2	2	0	259	66	30	563
Japanese	***1062***	104	0	115	13	*4*	0	310	*473*	71	***2152***

function words, and 5 by PoS bi-grams. However, they are relatively evenly spread: for the top 20 words in each topic (uniform priors; 15 topics), the proportion of function words varied from 0.22 to 0.44, mean 0.339 and standard deviation 0.063. The topics thus appear to be quite mixed.

Looking into the combined model, $Topic_3$ and $Topic_{11}$ inferred by this model are amongst the features that associated with high feature weights for Chinese. Coinciding with our expectation, the two potential terms indicative of Chinese — NN_NN and *the* — topped the lists of $Topic_3$ and $Topic_{11}$ respectively (where *the* also appears as the second most probable in $Topic_3$).

Summary. Although the LDA-induced classification models with feature space reduction somewhat underperformed in relation to the full feature-based models (i.e. the baselines), the 'topics' (latent factors) found appear in fact to be capturing some useful information characterising individual native languages. Given the performance attained by the POS-LDA models, and the fact that the clustering seems more intuitive here, it seems promising to explore LDAs further with larger class of unrestricted PoS bi-grams, or of syntactic features such as the parse tree substructures used in Chapter 4. This could be complemented by using the *adaptor grammars* of Johnson (2010) to capture collocational pairings as a form of feature selection, which we will demonstrate next. But, we will first give an overview of adaptor grammars and discuss its relation to LDA.

5.3 Adaptor Grammars

Adaptor grammars are a non-parametric extension to PCFGs that are associated with a Bayesian inference procedure. Here, we provide an informal introduction to adaptor grammars; Johnson et al. (2007) provide a definition of adaptor grammars as a hierarchy of mixtures of Dirichlet (or 2-parameter Poisson-Dirichlet) Processes to which the reader should turn for further details.

Adaptor grammars can be viewed as extending PCFGs by permitting the grammar to contain an unbounded number of productions; they are non-parametric in the sense that the particular

productions used to analyse a corpus depends on the corpus itself. Because the set of possible productions is unbounded, they cannot be specified by simply enumerating them, as is standard with PCFGs. Instead, the productions used in an adaptor grammar are specified indirectly using a *base grammar*: the subtrees of the base grammar's 'adapted non-terminals' serve as the possible productions of the adaptor grammar (Johnson et al., 2007), much in the way that subtrees function as productions in *Tree Substitution Grammars*.[8]

Another way to view adaptor grammars is that they relax the independence assumptions associated with PCFGs. In a PCFG productions are generated independently conditioned on the parent non-terminal, while in an adaptor grammar the probability of generating a subtree rooted in an adapted non-terminal is roughly proportional to the number of times it has been previously generated (a certain amount of mass is reserved to generate 'new' subtrees). This means that the distribution generated by an adaptor grammar 'adapts' based on the corpus being generated.

5.3.1 Mechanics of Adaptor Grammars

Adaptor grammars are specified by a PCFG G, plus a subset of G's non-terminals that are called the *adapted non-terminals*, as well as a *discount parameter* a_A, where $0 \leq a_A < 1$ and a *concentration parameter* b_A, where $b > -a$, for each adapted non-terminal A. An adaptor grammar defines a two-parameter Poisson-Dirichlet Process for each adapted non-terminal A governed by the parameters a_A and b_A. For computational purposes, it is convenient to integrate out the Poisson-Dirichlet Process, resulting in a predictive distribution specified by a Pitman-Yor Process (PYP). A PYP can be understood in terms of a 'Chinese Restaurant' metaphor in which 'customers' (observations) are seated at 'tables', each of which is labelled with a sample from a *base distribution* (Pitman and Yor, 1997).

In an adaptor grammar, unadapted non-terminals expand just as they do in a PCFG; a production r expanding the non-terminal is selected according to the multinomial distribution θ_r over productions specified in the grammar. Each adapted non-terminal A is associated with its own Chinese Restaurant, where the tables are labelled with subtrees generated by the grammar rooted in A. In the Chinese Restaurant metaphor, the customers are expansions of A, each table corresponds to a particular subtree expanding A, and the PCFG specifies the base distribution for each of the adapted non-terminals. An adapted non-terminal A expands as follows. A expands to a subtree t with probability proportional to n_t, where n_t is the number of times t has been previously generated. In addition, A expands using a PCFG rule r expanding A with probability proportional to $(m_A\, a_A + b_A)\, \theta_r$, where m_A is the number of subtrees expanding A (i.e., the number of tables

[8] For computational efficiency reasons, adaptor grammars require the subtrees to completely expand to terminals. The *fragment grammars* of O'Donnell (2011) lift this restriction.

in A's restaurant). Because the underlying Pitman-Yor Processes have a 'rich get richer' property, they generate power-law distributions over the subtrees for adapted non-terminals.

5.3.2 Adaptor grammars as LDA Extension

With the ability to rewrite non-terminals to entire subtrees, adaptor grammars have been used to extend unigram-based LDA topic models as demonstrated in Johnson (2010). This allows topic models to capture sequences of words with arbitrary length rather than just unigrams of word. It has also been shown that it is crucial to go beyond the *bag-of-words* assumption as topical collocations capture more meaning information and represent more interpretable topics (Wang et al., 2007).

Taking the PCFG formulation for the LDA topic models, it can be modified such that each topic $Topic_i$ generates sequences of words by adapting each of the $Topic_i$ non-terminals (usually indicated with an *underline* in an adaptor grammar). The overall schema for capturing topical collocations with an adaptor grammar is as follows:

$$
\begin{array}{ll}
Sentence \rightarrow Doc_j & j \in 1, \ldots, m \\
Doc_j \rightarrow \_j & j \in 1, \ldots, m \\
Doc_j \rightarrow Doc_j\ Topic_i & i \in 1, \ldots, t; \\
 & j \in 1, \ldots, m \\
\underline{Topic_i} \rightarrow Words & i \in 1, \ldots, t \\
Words \rightarrow Word & \\
Words \rightarrow Words\ Word & \\
Word \rightarrow w & w \in V_{pos} \\
 & w \in V_{pos+fw}
\end{array}
$$

There is a non-grammar-based approach to finding topical collocations as demonstrated by Wang et al. (2007). Both of these approaches learned useful collocations: for instance, Johnson (2010) found collocations such *gradient descent* and *cost function* associated with the topic of 'machine learning'; Wang et al. (2007) found the topic of 'human receptive system' comprises of collocations such as *visual cortex* and *motion detector*.

Related applications. Adaptor grammars have also been deployed as a form of feature selection in discovering useful collocations for perspective classification. Hardisty et al. (2010) argued that indicators of perspectives are often beyond the length of bigrams and demonstrated that the use of the adaptor grammar-inferred lexical n-grams of arbitrary length as features leads to the

start-of-the-art performance for perspective classification on the Bitter Lemons corpus, depicting two different perspectives (Israeli and Palestinian) on various issues. Their goal was to address the limitation of naïve Bayes models in which each class is only associated with a probability distribution over a fixed vocabulary. By adopting the adapted grammars, this allowed them to derive an 'adaptive' naïve Bayes model where the vocabulary can grow as required such that collocations of arbitrary length can be included according to the corpus used. The approach taken by Hardisty et al. (2010) demonstrates how both the collocation selection and classification can be expressed within a single model through adaptor grammar inference, without having to find collocations in a separate process. The overall approach that we are taking in the next section for classifying texts with respect to the author's native language is largely inspired by the general idea of Hardisty et al. (2010) in finding useful collocations. But, the key difference with Hardisty et al. (2010)'s approach will be that our focus is on collocations that mix PoS and lexical elements, rather than being purely lexical.

Another related piece of work that adopts grammatical inference is that of Börschinger et al. (2011), which also inspired our approach in the next section. In that work of Börschinger et al. (2011), their aim was to demonstrate that 'grounded' learning tasks (such as learning of a natural language) can indeed be represented within the scope of grammatical inference for certain cases. This was achieved by reducing the task of learning a semantic parser to a grammatical inference task using the PCFG formalism (rather than the more relaxed representation of adaptor grammars). The general idea was to train a PCFG grammar learning to map a collection of humanly generated comments on a robot soccer game to a set of candidate meanings in different contexts (representing the actions performed in the robot soccer game world); and the results reported were highly competitive. We will be adopting a similar idea to induce an adapted grammar, learning to infer the native language of an author from the written texts. (The grammatical inference procedures will be explained in Section 5.4.2.)

5.4 Feature Selection with Adaptor Grammars

As compared with the first task (Section 5.2), this second task investigated in this chapter differs in that it uses Bayesian techniques to discover collocations of arbitrary length for use in classification, over a mix of both PoS and function words, rather than for use as feature dimensionality reduction.

For the purpose of native language identification (NLI), small n-gram sizes like bi-gram or tri-gram might not suffice to capture sequences that are characteristic of a particular native language. On the other hand, an attempt to represent these with larger n-grams would not just lead to feature sparsity problems, but also computational efficiency issues. Some form of feature selection should

then come into play and adaptor grammars, as discussed above, seem to hold some promise here.

We explore the usefulness of adaptor grammars for n-gram feature selection with two different approaches. We first utilise adaptor grammars for discovery of high performing 'quasi-syntactic collocations' of arbitrary length as mentioned above and use them as classification features in a conventional maximum entropy (maxent) model for identifying the author's native language. In the second approach, we adopt a grammar induction technique to learn a grammar-based language model in a Bayesian setting. The grammar learned can then be used to infer the most probable native language that a given text written in a second language is associated with. The latter approach is actually closer to the work of Hardisty et al. (2010) using adaptor grammars for perspective modeling, in which discovery of collocations is an integral part of the approach rather than a separate process. This alternative approach is also similar in nature to the work of Börschinger et al. (2011) in which grounded learning of semantic parsers was reduced to a grammatical inference task.

5.4.1 Maximum Entropy-based Classification

In this subsection, we first explain the procedures taken to set up the conventional supervised classification task for NLI through the deployment of adaptor grammars for discovery of 'quasi-syntactic collocations' of arbitrary length. We then present the classification results attained based on these selected sets of n-gram features. In all of our experiments, we investigate two sets of collocations: pure PoS and a mixture of PoS and function words. The idea of examining the latter set is motivated by the results of Chapter 4 where inclusion of parse production rules lexicalised with function words as features had shown to improve the classification performance relative to unlexicalised ones.

5.4.1.1 Adaptor Grammars for Supervised Classification

We derive two adaptor grammars for the maxent classification setting, where each is associated with a different vocabulary (i.e. either pure PoS or the mixture of PoS and function words). We use the grammar of Johnson (2010) as presented in Section 5.3.2, except that the vocabulary differs: either $w \in V_{pos}$ or $w \in V_{pos+fw}$. For V_{pos}, there are 119 distinct PoS tags based on the Brown tagset. V_{pos+fw} is extended with 398 function words as per the earlier studies in the thesis. $m = 490$ is the number of documents, and $t = 25$ the number of topics (chosen as the best performing one from the first task of this chapter — feature dimensionality reduction with LDAs, Section 5.2).

Rules of the form $\texttt{Doc}_j \rightarrow \texttt{Doc}_j\ \texttt{Topic}_i$ that encode the possible topics that are associated

with a document j are given similar $\boldsymbol{\alpha}$ priors as used in LDA ($\boldsymbol{\alpha} = 5/t$ where $t = 25$ in our experiments). Likewise, similar $\boldsymbol{\beta}$ priors from LDA are placed on the adapted rules expanding from $\underline{\texttt{Topic}_i} \rightarrow \texttt{Words}$, representing the possible sequences of words that each topic comprises ($\boldsymbol{\beta} = 0.01$). The values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are also based on the established values presented in the first task (Section 5.2). The inference algorithm for the adaptor grammars are based on the Markov Chain Monte Carlo technique made available online by Johnson (2010).[9] The inference algorithm is run for 1,500 iterations for each of the adaptor grammars, with the Pitman-Yor parameters a and b parameters initialised to 0.01 and 10 respectively.

5.4.1.2 Classification Models with N-gram Features

Based on the two adaptor grammars inferred, the resulting collocations (n-grams) are extracted as features for the classification task of identifying the writers' native language. These n-grams found by the adaptor grammars are only a (not necessarily proper) subset of those n-grams that are strongly characteristic of a particular native language. In principle, one could find all strongly characteristic n-grams by enumerating all the possible instances of n-grams up to a given length if the vocabulary is of a small enough closed set, such as for PoS tags; but this is infeasible when the set is extended to PoS plus function words. The use of adaptor grammars here can be viewed as a form of feature selection, as in Hardisty et al. (2010).

Baseline models. To serve as the baseline, we take the commonly used PoS bigrams as per the previous work of NLI (Koppel et al., 2005). A set of 200 PoS bigrams is selected in two ways: the 200 most frequently occur in the training data (as in Chapter 4) and the 200 with the highest values of information gain (IG) in the training data (which have not been evaluated in other work, to our knowledge).

Enumerated n-gram models. Here, we enumerate all the possible n-grams up to a fixed length and select the best of these according to IG, as a generalisation of the baseline. The first motivation for this feature set is that, in a sense, this should give a rough upper bound for the adaptor grammar's PoS-alone n-grams, as these latter should most often be a subset of the former. The second motivation is that, it gives a robust comparison for the mixed PoS and function word n-grams, where it is infeasible to enumerate all of them.

ENUM-POS We enumerate all possible n-grams up to the length of 5, and select those that actually occur (i.e. of the $\sum_{i=1}^{5} 119^i$ possible n-grams, this is 218,042 based on the average of 5 folds).

[9] Adaptor grammar software is available on `http://web.science.mq.edu.au/~mjohnson/Software.htm`.

We look at the top n-grams up to length 5 selected by IG: the top 2,800 and the top 6,500 (for comparability with adaptor grammar feature sets, see below), as well as the top 10,000 and the top 20,000 (to study the effect of larger feature space).

Adaptor grammar n-gram models. The classification features are the two sets of selected collocations inferred by the adaptor grammars which are the main interest of this paper.

AG-POS This first set of the adaptor grammar-inferred features comprise pure PoS n-grams (i.e. V_{pos}). The largest length of n-gram found is 17, but about 97% of the collocations are of length between 2 to 5. We investigate three variants of this feature set: top 200 n-grams of all lengths (based on IG), all n-grams of all lengths ($n = 2,795$ on average), and all n-grams up to length 5 ($n = 2,710$ on average).

AG-POS+FW This second set of the adaptor grammar-inferred features are mixtures of PoS and function words (i.e. V_{pos+fw}). The largest length of n-gram found for this set is 19 and the total number of different collocations found is much higher. For the purpose of comparability with the first set of adaptor grammar features, we investigate the following five variants for this feature set: top 200 n-grams of all lengths, all n-grams of all lengths ($n = 6,490$ on average), all n-grams up to the length of 5 ($n = 6,417$ on average), top 2,800 n-grams of all different lengths, and top 2,800 n-grams up to the length of 5. (Similarly, all the selections are based on IG).

In our models, all feature values are of binary type. For the classifier, we again employ the maximum entropy machine learner — MegaM (fifth release) by Hal Daumé III.[10]

5.4.1.3 Data and Evaluation

The classification experiments are conducted on the same dataset as per the first task on feature dimensionality reduction with LDAs.[11] Hence, the data set used in this second task of the chapter consists of 490 texts written in English by authors of the same seven different native language groups: Bulgarian, Czech, French, Russian, Spanish, Chinese, and Japanese. Each native language contributes 70 out of the 490 texts. As we are again using a relative small data set, we perform k-fold cross-validation, choosing $k = 5$.

[10] MegaM software is available on `http://www.umiacs.umd.edu/~hal/megam/`.

[11] As noted in Chapter 4, Joel Tetreault and Daniel Blanchard from ETS have pointed out (personal communication) that there is a subtle issue with ICLE that could have an impact on the classification performance of NLI tasks; in particular, when character n-grams are used as features, some special characters used in some ICLE texts might affect performance. For our case, this should not be of much issue since they will not appear in our collocations.

Table 5.11: Maxent classification results for individual feature sets — 5-fold cross validation

Features (n-grams)	Accuracy
BASELINE-POS [top200 MOST-FREQ]	53.87
BASELINE-POS [top200 IG]	56.12
AG-POS [top200 IG]	61.02
AG-POS [all $\leq$17-gram] ($n \approx 2800$)	68.37
AG-POS [all $\leq$ 5-gram] ($n \approx 2700$)	68.57
AG-POS+FW [top200 IG]	58.16
AG-POS+FW [all $\leq$19-gram] ($n \approx 6500$)	**74.49**
AG-POS+FW [all $\leq$5-gram] ($n \approx 6400$)	**74.49**
AG-POS+FW [top2800 IG $\leq$ 19-gram]	71.84
AG-POS+FW [top2800 IG $\leq$ 5-gram]	71.84
ENUM-POS [top2800 IG $\leq$ 5-gram]	69.79
ENUM-POS [top6500 IG $\leq$ 5-gram]	72.44
ENUM-POS [top10K IG $\leq$ 5-gram]	71.02
ENUM-POS [top20K IG $\leq$ 5-gram]	71.43

5.4.1.4 Classification Results

Table 5.11 presents all the classification results for the individual feature sets, along with the baselines. On the whole, both sets of the collocations inferred by the adaptor grammars perform better than the two baselines. We make the following observations:

- Regarding ENUM-POS as a (rough) upper bound, the adaptor grammar AG-POS with a comparable number of features performs almost as well. However, because it is possible to enumerate many more n-grams than are found during the sampling process, ENUM-POS opens up a gap over AG-POS of around 4%.
- Collocations with a mix of PoS and function words do in fact lead to higher accuracy as compared to those of pure PoS (except for the top 200 n-grams); for instance, compare the 2,800 n-grams up to length 5 from the two corresponding sets (71.84 vs. 68.57).
- Furthermore, the adaptor grammar-inferred collocations with mixtures of PoS and function words (AG-POS+FW) in general perform better than our rough upper bound of PoS collocations, i.e. the enumerated PoS n-grams (ENUM-POS): the overall best results of the two feature sets are 74.49 and 72.44 respectively.

Given that the AG-POS+FW n-grams are capturing different sorts of document characteristics, they could potentially usefully be combined with the PoS-alone features. We thus combined them with both AG-POS and ENUM-POS feature sets, and the classification results are presented in Table 5.12. We tried two ways of integrating the feature sets: one way is to take the overall top 2,800 of

Table 5.12: Maxent classification results for combined feature sets — 5-fold cross validation

Features (n-grams)	**Accuracy**
AG-POS [all $\leq$ 5-gram] & FW	72.04
ENUM-POS [top2800 $\leq$ 5-gram] & FW	73.67
AG-POS+FW & AG-POS [a]	**75.71**
AG-POS+FW & AG-POS [b]	74.90
AG-POS+FW & ENUM-POS [top2800] [a]	73.88
AG-POS+FW & ENUM-POS [top2800] [b]	74.69
AG-POS+FW & ENUM-POS [top10K] [b]	74.90
AG-POS+FW & ENUM-POS [top20K] [b]	75.10

[a] *Features from the two sets are selected based on the overall top 2800 with highest IG;* [b] *features from the two sets are just linearly concatenated*

the two sets based on IG; the other way is to just combine the two sets of features by concatenation of feature vectors (as indicated by *a* and *b* respectively in the result table). For comparability purposes, we considered only n-grams up to length of 5. A baseline approach to this is just to add in function words as unigram features by feature vector concatenation, giving two further models: AG-POS [all $\leq$ 5-gram] & FW and ENUM-POS [top2800 $\leq$ 5-gram] & FW.

Overall, the classification accuracies attained by the combined feature sets are higher than the individual feature sets. The best performing of all the models is achieved by combining the mixed PoS and function word collocations with the adaptor grammar-inferred PoS, producing the best accuracy thus far of 75.71. This demonstrates that features inferred by adaptor grammars do capture some useful information and function words are playing a role. The way of integrating the two feature sets has different effects on the types of combination. As seen in Table 5.12, method *a* works better for the combination of the two adaptor grammar feature sets; whereas method *b* works better for combining adaptor grammar features with enumerated n-gram features.

Using adaptor grammar collocations also outperforms the alternative baseline of adding in function words as unigrams. For instance, the best performing combined feature set of both AG-POS and AG-POS+FW does result in higher accuracy as compared to the two alternative baseline models, comparing 75.71 with 72.04 (and 75.71 with 73.67). This demonstrates that our more general PoS plus function word collocations derived from adaptor grammars are indeed useful, and supports the argument of Wang et al. (2007) that they are a useful technique for looking into features beyond just the bag of words.

5.4.2 Language Model-based Classification

In this subsection, adaptor grammars are utilised in a different way to the classification task of NLI. We take a *language modeling* approach; the idea here is to adopt grammatical inference to learn a grammar-based language model to represent the texts written by non-English native users. The grammar learned is then used to predict the most probable native language that a document (a sentence) is associated with.

In a sense, we are using a parser-based language model to rank the documents with respect to native language. We draw on the work of Börschinger et al. (2011) for this. In that work (as mentioned in Section 5.3.2), the task was grounded learning of a semantic parser. Training examples there consisted of natural language strings (descriptions of a robot soccer game) and a set of candidate meanings (actions in the robot soccer game world) for the string; each was tagged with a context identifier reflecting the actual action of the game. A grammar was then induced that would parse the examples, and was used on test data (where the context identifier was absent) to predict the context. We take a similar approach to developing an grammatical induction technique, although where they used a standard LDA topic model-based PCFG, we use an adaptor grammar here.[12] We expect that the classification results will likely to be lower than for the discriminative approach of Section 5.4.1 with the maxent-based approach.

However, the approach is of interest for a few reasons: whereas the adaptor grammar plays an ancillary, feature selection role in Section 5.4.1, here the feature selection is an organic part of the approach as per the actual implementation of Hardisty et al. (2010); adaptor grammars can potentially be extended in a natural way with unlabelled data; and for the purposes of the second task investigated in this chapter, it constitutes a second, quite different way to evaluate the use of n-gram collocations.

5.4.2.1 Language Models

We derive two adaptor grammar-based language models. One consists of only unigrams and bigrams, and the other finds n-gram collocations, in both cases over either PoS or the mix of PoS and function words. The assumption that we make is that each document (each sentence) is a mixture of two sets of topics: one is the native language-specific topic (i.e. characteristic of the native language) and the other is the generic topic (i.e. characteristic of the second language — English in our case). The generic topic is thus shared across all languages, and will behave quite differently from a language-specific topic, which is not shared. In other words, there are eight topics, representing seven native language groups that are of interest — Bulgarian (BL), Czech (CZ), French

[12] A special thanks to Benjamin Börschinger for the helpful discussion on the formulation of the adaptor grammars used in this grammatical inference approach.)

(FR), Russian (RU), Spanish (SP), Chinese (CN), and Japanese (JP) — and the second language English itself.[13]

Bi-gram models. The following rule schema for an adaptor grammar is applicable to both vocabulary types of PoS and the mixture of PoS and function words.

$$
\begin{aligned}
&Root \rightarrow \_lang\ langTopics \\
&langTopics \rightarrow langTopics\ langTopic \\
&langTopics \rightarrow langTopics\ nullTopic \\
&langTopics \rightarrow langTopic \\
&langTopics \rightarrow nullTopic \\
&\underline{langTopic} \rightarrow Words \\
&\underline{nullTopic} \rightarrow Words \\
&Words \rightarrow Word\ Word \\
&Words \rightarrow Word \\
&Word \rightarrow w \qquad\qquad w \in V_{pos};\ w \in V_{pos+fw}
\end{aligned}
$$

(Note: $\_lang$ is the native language identifier where it takes the value from the set of 7 native languages {BL, CZ, FR, RU, SP, CN, JP}; $langTopic$ and $nullTopic$ represent the native language specific topic and the generic (second language) topic, respectively.)

N-gram models. The adaptor grammar is generally the same as the above with the only exception that the non-terminal $Words$ is now rewritten as follows in order to capture n-gram collocations of arbitrary length.

$$
\begin{aligned}
&Words \rightarrow Words\ Word \\
&Words \rightarrow Word
\end{aligned}
$$

It should be noted that the two grammars above can in theory be applied to an entire document or on individual sentences. For this present work, we work on the sentence level as the run-time of the current implementation of the adaptor grammars grows proportional to the cube of the sentence length. For each grammar we try both sparse and uniform Dirichlet priors ($\boldsymbol{\alpha} = \{0.01, 0.1, 1.0\}$). The sparse priors encourage only a minority of the rules to be associated with high probabilities.

[13] We could just induce a regular PCFG here, rather than an adaptor grammar, by taking as terminals all pairs of PoS tags. We use the adaptor grammar formulation for comparability.

5.4.2.2 Training and Evaluation

As we are using the same data set as per the previous approach, we perform 5-fold cross validation as well. However, the training for each fold is conducted with a different grammar consisting of only the vocabulary that occur in each training fold. The reason is that we are now having a form of *supervised* topic models where the learning process is guided by the native languages. Hence, each of the training sentences are prefixed with the (native) language identifiers $\_lang$, as seen in the $Root$ rules of the grammar presented above.

To evaluate the grammars learned, as in Börschinger et al. (2011) we need to slightly modify the grammars above by removing the language identifiers ($\_lang$) from the $Root$ rules and then parse the *unlabeled* sentences using a publicly available CKY parser.[14] The predicted native language is inferred from the parse output by reading off the $langTopics$ that the $Root$ is rewritten to. We take that as the most probable native language for a particular test sentence. The parsing performance (i.e. classification outcome) is evaluated at both the sentence and the document levels. At the document level, we select as the class the language predicted for the largest number of sentences in that document.

5.4.2.3 Parsing Results

Tables 5.13 and 5.14 present the parsing results at the sentence level and the document level, respectively. On the whole, the results at the sentence level are much poorer as compared to those at the document level. In light of the results of Section 5.4.1.4 (i.e. maxent-based classification), it is surprising that bi-gram models appear to perform better than n-gram models for both types of vocabulary, with the exception of AG-POS+FW at the document level. In fact, one would expect n-gram models to perform better in general as it is a generalisation that would contain all the potential bi-grams. Nonetheless, the language models over the mixture of PoS and function words appear to be a more suitable representative of our learner corpus as compared to those over purely PoS, confirming the usefulness of integrated function words for the NLI classification task.

It should also be noted that sparse priors generally appear to be more appriopriate; except that for AG-POS+FW n-grams, uniform priors are indeed better and resulted in the highest parsing result of 50.15. Although all the parsing results are much weaker as compared to the results presented in Section 5.4.1.4, they are all higher than the majority baseline of 14.29% i.e. $70/490$.

[14] CKY parser by Mark Johnson is available on `http://web.science.mq.edu.au/~mjohnson/Software.htm`.

Table 5.13: Language modeling-based classification results based on parsing — at the sentence level

Feature Types	**Accuracy**		
(n-grams)	($\alpha = 0.01$)	($\alpha = 0.1$)	($\alpha = 1.0$)
AG-POS [bigrams]	26.84	27.03	26.77
AG-POS [n-grams]	25.85	25.78	25.62
AG-POS+FW [bigrams]	28.58	28.40	27.43
AG-POS+FW [n-grams]	26.64	27.64	28.75

Table 5.14: Language modeling-based classification results based on parsing — at the document level

Features	**Accuracy**		
(n-grams)	($\alpha = 0.01$)	($\alpha = 0.1$)	($\alpha = 1.0$)
AG-POS [bigrams]	41.22	38.88	39.69
AG-POS [n-grams]	36.12	34.90	35.20
AG-POS+FW [bigrams]	47.45	46.94	44.64
AG-POS+FW [n-grams]	43.97	49.39	**50.15**

5.4.3 Discussion

Here we take a closer look at how well each approach does in identifying the individual native languages. The confusion matrix for the best performing model of the two approaches are presented in Table 5.15 and Table 5.16, respectively. Both approaches perform reasonably well for the two Oriental languages (Chinese in particular); this is not a major surprise, as the two languages are not part of the language family that the rest of the languages come from (i.e. Indo-European). Under the supervised maxent classification, misclassifications largely are observed in the Romance ones (French and Spanish) as well as Russian; for the language model-based approach, Bulgarian is identified poorly, and Spanish moreso. However, the latter approach appears to be better in identifying Czech. On the whole, the maxent approach results in much fewer misclassifications compared to its counterpart.

In fact, there is a subtle difference in the experimental setting of the models derived from the two approaches with respect to the adaptor grammar: the number of topics. Under the maxent setting, the number of topics t was set to 25, while we restricted the models with the language modeling approach to only eight topics (seven for the individual native languages and one for the common second language, English). Looking more deeply into the topics themselves reveals that there appears to be at least two out of the 25 topics (from the supervised models) associated with n-grams that are indicative of the native languages, taking Chinese and Japanese as examples (see

Table 5.15: Confusion matrix based on the best performing model — under maxent setting

	BL	**CZ**	**RU**	**FR**	**SP**	**CN**	**JP**
BL	[52]	5	7	4	2	–	–
CZ	5	[50]	5	3	4	–	3
RU	6	8	[46]	5	1	–	4
FR	7	3	5	[43]	8	–	4
SP	7	2	4	9	[47]	–	1
CN	–	–	–	–	–	[70]	–
JP	–	–	2	2	1	2	[63]

BL:Bulgarian, CZ:Czech, RU:Russian, FR:French, SP:Spanish, CN:Chinese, JP:Japanese

Table 5.16: Confusion matrix based on the best performing model — under language modeling setting

	BL	**CZ**	**RU**	**FR**	**SP**	**CN**	**JP**
BL	[20]	32	9	6	–	1	2
CZ	2	[59]	3	1	–	–	5
RU	3	41	[19]	2	1	–	4
FR	8	20	4	[31]	4	–	3
SP	7	27	11	12	[9]	–	4
CN	–	2	–	2	–	[62]	4
JP	–	19	1	2	–	1	[47]

BL:Bulgarian, CZ:Czech, RU:Russian, FR:French, SP:Spanish, CN:Chinese, JP:Japanese

the associated topics in Table 5.17).[15] Perhaps associating each native language with only one generalised topic is not sufficient.

Furthermore, the distribution of n-grams among the topics (i.e. subtrees of collocations derived from the adaptor grammars) is quite different between the two approaches although the total number of n-grams inferred by each approach is about the same. For the language modeling ones, a high number of n-grams were associated with the generic topic $nullTopic$[16] and each language-specific topic $langTopic$ has a lower number of n-grams relative to bi-grams associated with it (Table 5.18). For the maxent-based models, in contrast, the majority of the topics were associated with a higher number of n-grams (Table 5.19). The smaller number of n-grams to be used as features — and the fact that their extra length means that they will occur more sparsely in the documents — seems to be the core of the problem.

Nonetheless, the language models inferred discover relevant n-grams that are representative of

[15] Taking the examples from the first task in this chapter (Section 5.2) as reference, we found similar n-grams that are indicative of Japanese and Chinese.

[16] This is quite plausible as there should be quite a number of structures that are representative of native English speakers that are shared by non-native speakers.

Table 5.17: Top mixture n-grams (collocations) for 4 out of the 25 topics representative of Japanese and Chinese — under maxent setting

Top 10 Mixture N-grams			
Japanese		**Chinese**	
topic$_2$	**topic**$_{23}$	**topic**$_9$	**topic**$_{17}$
.	.	NN	.
we VB	PPSS VB	a NN	NN NN
our NNS	my NN	NN NN	NNS
our NN	CC	VBN by	NN
NN	VBG	NP .	RB ,
PPSS VB	PPSS think	NP	of NN
about	NN	:	JJ NN
because	PPSS VBD	(	NN .
it .	RB	as	VBG NN
we are	PPSS ’ NN	NN NN NN	NN NN NN

N-grams of pronoun with verb are found at the upper end of $Topic_2$ and $Topic_{23}$ reflecting the frequent usage of Japanese; n-grams of noun are top n-grams under $Topic_9$ and $Topic_{17}$ indicating Chinese's common error of determiner-noun disagreement

individual native languages. For instance, the bi-gram NN NN, which we claimed in Section 5.2.4 (and also Chapter 4) may reflect errors of determiner-noun agreement commonly found amongst Chinese learners, was found under the Chinese topic at the top-2 position with a probability of 0.052 as compared to the other languages at the probability range of 0.0005-0.003. Similarly, one example for Japanese, the mixture bi-gram PPSS think, indicating frequent usage of pronouns within Japanese was seen under the Japanese topic at the top-9 position with a probability of 0.025 in relation to other languages within the range of 0.0002-0.006: this phenomenon as characteristic of Japanese speakers has also been noted for different corpora by Ishikawa (2011) as we highlighted in Chapter 4. (Note that this collocation as well as its pure PoS counterpart PPSS VB are amongst the top n-grams discovered under the maxent setting as seen in Table 5.17.) Some excerpts extracted from the corpus that illustrate these two common phenomena are presented in Table 5.20.

To investigate further the issue associated with the number of topics under the language modeling setting, we attempted to extend the adaptor grammar with three additional topics that group the seven native languages of interest: Slavic, Romance, and Oriental. The resulting grammar (with $familyTopic$ representing the three language groups) is presented as follows. However, the parsing result does not improve over the initial setting with eight topics in total.

Table 5.18: Distribution of n-grams (collocations) for each topic — under language modeling setting

	N-gram Frequency																	
Model	**BLTopic**		**CZTopic**		**RUTopic**		**FRTopic**		**SPTopic**		**CNTopic**		**JPTopic**		**NullTopic**	
Types	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**
Bigrams	374	187	352	219	350	211	426	165	351	156	397	351	394	194	867	6169
N-grams	177	159	226	217	148	202	151	152	128	147	357	255	209	226	3089	7794

(a) Subcolumns are for n-grams of pure PoS; (b) subcolumns are for n-grams of mixtures of PoS and function words

Table 5.19: Distribution of n-grams (collocations) for each topic — under maxent setting

N-gram Frequency																				
Topic$_1$		**Topic**$_2$		**Topic**$_3$		**Topic**$_4$		**Topic**$_5$		**Topic**$_6$		**Topic**$_7$		**Topic**$_8$		**Topic**$_9$		**Topic**$_{10}$	
(a)	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**
174	443	145	441	136	245	141	341	236	519	169	748	127	340	182	473	109	339	190	236
Topic$_{11}$		**Topic**$_{12}$		**Topic**$_{13}$		**Topic**$_{14}$		**Topic**$_{15}$		**Topic**$_{16}$		**Topic**$_{17}$		**Topic**$_{18}$		**Topic**$_{19}$		**Topic**$_{20}$	
(a)	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**
57	259	126	455	103	543	211	225	170	459	81	309	238	207	152	475	119	452	333	423
Topic$_{21}$				**Topic**$_{22}$				**Topic**$_{23}$				**Topic**$_{24}$				**Topic**$_{25}$			
(a)		**(b)**		**(a)**		**(b)**		**(a)**		**(b)**		**(a)**		**(b)**		**(a)**		**(b)**	
245		341		168		492		194		472		201		366		195		190	

(a) Subcolumns are for n-grams of pure PoS; (b) subcolumns are for n-grams of mixtures of PoS and function words

Table 5.20: Excerpts from ICLE illustrating the common phenomena observed amongst Chinese and Japanese

Languages	Excerpts from ICLE
Chinese	...the overpopulation problem in *urban area* ...
	...The development of *country park* can directly ...
	...when it comes to *urban renewal project* ...
	...As developing *new town* in ...
	...and reserve *some country park* as ...
Japanese	...*I think* many people will ...
	...*I think* governments should not ...
	...*I think* culture is the most significant ...
	...*I think* the state should not ...
	...*I really think* we must live ...

$$Root \rightarrow \_lang\ langTopics$$
$$langTopics \rightarrow langTopics\ langTopic$$
$$langTopics \rightarrow langTopics\ familyTopic$$
$$langTopics \rightarrow langTopics\ nullTopic$$
$$langTopics \rightarrow langTopic$$
$$langTopics \rightarrow familyTopic$$
$$langTopics \rightarrow nullTopic$$
$$\underline{langTopic} \rightarrow Words$$
$$\underline{familyTopic} \rightarrow Words$$
$$\underline{nullTopic} \rightarrow Words$$
$$Words \rightarrow Words\ Word$$
$$Words \rightarrow Word$$
$$Word \rightarrow w \qquad w \in V_{pos}; w \in V_{pos+fw}$$

Summary. We have shown that the extension of adaptor grammars to discovering collocations beyond the lexical, in particular a mix of PoS tags and function words, can produce features useful in the NLI classification problem. More specifically, when added to a new baseline presented in this paper, the combined feature set of both types of adaptor grammar inferred collocations produces the best result in the context of using n-grams for NLI. The usefulness of the collocations does vary, however, with the technique used for classification.

Follow-up work from here will involve a broader exploration of the parameter space of the adaptor grammars, in particular the number of topics and the value of prior $\boldsymbol{\alpha}$; a look at other non-parametric extensions of PCFGs, such as *infinite PCFGs* (Liang et al., 2007) for finding a set of non-terminals permitting more fine-grained topics; and an investigation of how the approach

can be extended to *semi-supervised learning* to take advantage of the vast quantity of texts with errors available on the Web.

5.5 Concluding Remarks

The two approaches investigated in this chapter demonstrate how topic modeling can be exploited for the classification task of native language identification. Although the first approach of deploying the basic LDA models as a form of feature space reduction for finding clusters of coherent features does not lead to an especially high classification performance, the 'topics' or latent factors inferred from (at least) the PoS-based LDA models do capture some indicative patterns (in the form of PoS bi-grams) that are characteristics for a particular native language.

With the second approach, unigram-based topic models are extended using adaptor grammars for discovering collocational n-grams of arbitrary length that are useful for NLI, treating them as an approximation to the 'deep' syntactic structures (such as production rules and tree templates for parse reranking). Using the conventional classification method with a maxent machine learner, n-grams of beyond just purely PoS — mixtures of PoS with function words — has shown that reasonably good classification performance can be achieved, although it is not as high as the performance attained with the syntactic structures used in Chapter 4, parse production rules in particular.

Before we conclude in the next chapter, one issue of concern that deserves some further discussion is of the linguistic phenomena greatly investigated in this chapter as well as in the previous two chapters: *(mis)use of determiners*. Most of the discussions thus far have been focusing on Chinese speakers of English on the whole. This is largely due to the subset of the ICLE corpus used in this thesis in which this phenomena appears to be evident amongst Chinese, who appear to overuse certain determiners such as *the* as discussed in Section 5.2.4. One would argue that similar determiner-noun errors should also be pronounced for Japanese as well as Slavic speakers, given the fact that these languages lack articles in their language system; however, these were not observed in our dataset. A plausible explanation could be that these non-native English speakers might have chosen to avoid using certain determiners when lack of confidence — one of the several ways of manifesting errors as mentioned in Section 2.3.1.2. It has also been argued that articles (definites in particular such as *the*) is closely related to demonstratives (such as *this* and *that*). As for languages without articles but with demonstratives, it is reasonable to hypothesise that there might be lexical transfer from the semantics of L1 demonstratives to the usage of L2 (English) articles subjected to the learners' interpretation on the semantic environments of the noun phrases in which the articles are used — consequently, a demonstrative such as *that* could be (over)used

Table 5.21: Usage of the article *the* versus the demonstrative *that*

Language	**Relative Frequency of *the***	**Relative Frequency of *that***
Bulgarian	0.0814	0.0133
Czech	0.0648	0.0119
French	0.0952	0.0117
Russian	0.0764	0.0134
Spanish	0.0903	0.0143
Chinese	0.1256	0.0112
Japanese	0.0661	0.0121

for definiteness marking in place of *the*. Table 5.21 shows that Bulgarian, Russian and Japanese; each has a somewhat higher number of usage for *that* in comparison with Chinese as observed in our dataset. In fact, there have been a number of related works attempted to investigate the relationship between L1 demonstratives and English definite articles (Niimura and Brenda, 1996; Robertson, 2000; Butler, 2002; Ionin et al., 2012); yet the results are not conclusive. It is thus an interesting topic of research that worth addressing in the future.

Chapter 6

Conclusion

6.1 Summary of Findings

Prior to the studies presented in this thesis, there were only a limited number of works on native language identification (NLI) in which the usefulness of syntactic features had not been explored extensively. Following our publications that specifically exploit syntactic features for NLI taking cross-linguistic influence as the basis (Wong and Dras, 2009, 2011), work on NLI proliferated and much subsequent related research has emerged — such as the work of Kochmar (2011), Swanson and Charniak (2012), and Bergsma et al. (2012) — which utilise syntactic features in some way. Work has also appeared very recently in the linguist community on exploring the transfer effects of native language in second language learning in an automated manner rather than relying on conventional human judgements, using mainly statistical-based methods focusing on linear discriminant analysis (Jarvis and Crossley, 2012).

The work in the thesis started with, as its motivating principle, the assumptions of Contrastive Analysis Hypothesis (CAH) in its weaker form, although the ways in which *syntax* appears to be useful in native language identification differ quite significantly from the CAH predictions. As seen in Chapter 3, there is some modest support for the contention that syntactic structure can indeed help in detecting the writers' native language, through a statistical analysis of three syntactic error types and through machine learning-based classification using only features based on those error types. This suggests that errors as a result of native language transfer appear to be useful for the NLI task. Despite the fact that there appears to be no improvement in terms of classification performance when combining these with features commonly used in other machine learning approaches to the classification task of NLI, such as Koppel et al. (2005), a comparable level of accuracy rate of around 80% is observed on the classification task of identifying seven

different native language groups based on the texts from the ICLE corpus.

In view of the limitation of the first approach focusing on only three manually detected syntactic errors, Chapter 4 explored an automated means of detecting different types of syntactic errors as well as other cross-linguistic effects through statistical parsing. The outcomes of this second approach suggests that syntactic structures — in particular those that are characterised by horizontal slices of parse trees (i.e. context-free grammar (CFG) production rules) — are not only useful in terms of sentence-based grammaticality judgement, but can be beneficial to NLI tasks in identifying the native language of the writers. Moreover, more fine-grained production rules augmented with latent variables, as are generated by the Berkeley parser, lead to the best classifcation performance. On the other hand, the more generalised syntactic structures represented by the tree templates used for parse reranking (Charniak and Johnson, 2005) which were expected to be more helpful in general but do not appear so, as they seem to be only useful for the simpler task of judging sentence grammaticality. Nonetheless, an interesting aspect worth highlighting is that distributional differences over syntactic structures appear to be representative of syntactic errors (e.g. misuse of determiners by Chinese speakers) as well as other phenomena of cross-linguistic influence, including avoidance (e.g. underuse of relative clauses by Chinese) and overuse (e.g. pronouns by Japanese).

The last approach investigated in this thesis shows some promising outcomes from adopting topic modeling to native language identification in two ways, as discussed in Chapter 5. The clusters of coherent latent factors ('topics') found appear to be capturing useful information for individual native languages when inspecting the topics manually, although the topic modeling-induced classification models with the reduction in feature dimension in general underperform in relation to the full feature-based models (i.e. taking features themselves for classification rather than the topic probabilities inferred from the topic models). With the extension to adaptor grammars, collocations beyond the purely lexical ones, in particular mixtures of PoS tags and function words, can produce informative features that are useful in the NLI classification task. More specifically, adaptor grammar-induced n-grams consist of mixtures of PoS and function words outperform all enumerated PoS n-grams with function words incorporated separately. However, the usefulness of n-gram collocations does vary with respect to the technique used for classification. As supported by the classification results, the conventional machine learning-based approach with maximum entropy learners performs better in comparison with the language modeling-based approach. All in all, classification models with the full syntax represented in the form of production rules still performs better than any of the topic modeling-induced models for the NLI tasks conducted in the thesis.

6.2 Limitations and Future Outlook

The discoveries of the thesis have also led to a recognition of various limitations and caveats. One that requires much further investigation is the corpus used throughout the thesis. There have been some arguments that the ICLE corpus might have subject-based clues that would lead to a distortion when used for tasks in the NLI paradigm, even with features typically regarded as not connected to the subject of the text. However, the effect of this on the classification results presented in this thesis would likely be a slight inflation of all the results, rather than any wholesale revision of relative usefulness. As a result, some other recent research, such as Brooke and Hirst (2012), has resorted to alternative sources for reliable training corpora for NLI-related tasks. Another related issue to be aware of when using learner corpora in general is that all other variables — for example, the writers' proficiency level in the second language and the registers of second language texts — must be certain to be held constant except for the native language factor in order to yield unbiased classification outcomes. The availability of such a corpus is still far from ideal; future research efforts to compiling a fully controlled corpus would certainly be beneficial to NLI-related tasks.

Given that this thesis has taken CAH as the basis to identifying predictive syntactic features for NLI, the types of syntactic errors and patterns uncovered are expected to be primarily influences of the native language. It is undeniable that there is the possibility that some of these features might be in fact the outcomes of the developmental process. As suggested in the extensive literature of second language acquisition (SLA) by Ellis (2008), interlingual (interference) and intralingual (developmental) errors are often not easy to be indubitably distinguished. It is therefore worth exploring constrastive analysis and error analysis hand-in-hand in greater detail in order to rule out syntactic errors and patterns that are irrelevant for the purpose of determining the writers' native language.

With respect to the approaches investigated in this thesis, despite the outcomes of exploiting parse structures and topic modeling-induced features sound promising, the credibility of these approaches needs to be further attested. Admittedly, the number of native languages examined throughout the studies of this thesis is rather limited in size — only seven in total. Another related but orthogonal issue: the subset of the ICLE corpus used in this thesis is also rather small in size. A more extensive study would need to be conducted on a wider range of native languages with a larger corpus to substantiate the credibility as well as the scalability of these approaches. In relation to this, we had extended these approaches to the first ever shared task on NLI organised by researchers from ETS based on a relatively larger corpus comprising 11 native languages with

1,100 essays per language (Malmasi et al., 2013).[1] Furthermore, having the ability to identify a broader range of native language groups would be of help to the CALL practitioners at large in detecting and correcting specific errors produced by learners of different native langauge backgrounds with more efficacy. It is anticipated that there would still be a rise in terms of the number of English learners as it has been reported that more than one billion of people in the world are using English as their second or foreign language (Guo and Beckett, 2007).

As for further improvement on the approaches, one potential avenue of future work is to further explore different variants of topic modeling that might deem suitable. For example, the non-parametric variant — hierarchical Dirichlet processes (Teh et al., 2006) — which allows an unbounded number of topics (coherent latent factors) and sharing of topics across groups of documents. In relation to our task, this could be those syntactic patterns found in the grammatical native English constructions across the different native language groups. Other variants that could be potentially useful include more linguistic-informed syntactic topic modeling (Boyd-Graber and Blei, 2009) as well as supervised topic modeling (Blei and McAuliffe, 2008), which might produce feature clusters that are more closely aligned to native language identification cues.

To wrap up, native language identification ideas can be applied in a number of ways. As highlighted in Chapter 1, SLA researchers and practitioners can benefit from NLI studies from a pedagogical standpoint, specifically in terms of second language assessment as well as second language curriculum planning. As for heightening Internet security in relation to phishing attacks, successfully identifying the phishers' native language would be of help to law enforcement agencies in investigation. And given the recent surge of interest in the field, no doubt many more techniques and applications will be forthcoming.

[1] Further information on NLI Shared Task 2013 is available at `http://www.nlisharedtask2013.org/`.

Appendix A

Subsets of ICLE Corpus

Below are the subsets taken from the seven subcorpora of *International Corpus of Learner English (Version 2)* (Granger et al., 2009), for the purpose of all the NLI classification tasks conducted in the thesis. For each of the subsets (each representing one native language group):

- the first 70 essays are used as the training set in Chapter 3 and for cross validation in Chapter 4 and Chapter 5;
- the last 25 essays are used as the held-out test set in Chapter 3 and Chapter 4;
- the 15 essays in between these two previous sets are those used in Chapter 3 — for the purpose of investigating the impact of the training sample size on classification performance (as described in Section 3.7).

All the essays are of argumentative type and with 500–1000 words.

Bulgarian

The first 70 essays:

BGSU1001	BGSU1002	BGSU1003	BGSU1004
BGSU1005	BGSU1006	BGSU1007	BGSU1008
BGSU1011	BGSU1012	BGSU1013	BGSU1014
BGSU1018	BGSU1022	BGSU1024	BGSU1030
BGSU1031	BGSU1034	BGSU1035	BGSU1036
BGSU1037	BGSU1039	BGSU1042	BGSU1044
BGSU1045	BGSU1047	BGSU1051	BGSU1053
BGSU1055	BGSU1057	BGSU1058	BGSU1059
BGSU1065	BGSU1066	BGSU1067	BGSU1068
BGSU1071	BGSU1072	BGSU1073	BGSU1074
BGSU1076	BGSU1080	BGSU1081	BGSU1082
BGSU1083	BGSU1086	BGSU1087	BGSU1099
BGSU1100	BGSU1102	BGSU1105	BGSU1107
BGSU1108	BGSU1109	BGSU1114	BGSU1115
BGSU1116	BGSU1119	BGSU1126	BGSU1127
BGSU1129	BGSU1130	BGSU1132	BGSU1133
BGSU1136	BGSU1139	BGSU1140	BGSU1145
BGSU1146	BGSU1148		

The 15 essays in between:

BGSU1149	BGSU1150	BGSU1152	BGSU1153
BGSU1154	BGSU1157	BGSU1158	BGSU1159
BGSU1160	BGSU1162	BGSU1164	BGSU1165
BGSU1168	BGSU1172	BGSU1173	

The last 25 essays:

BGSU1174	BGSU1178	BGSU1182	BGSU1183
BGSU1184	BGSU1190	BGSU1191	BGSU1192
BGSU1193	BGSU1194	BGSU1195	BGSU1196
BGSU1198	BGSU1202	BGSU1203	BGSU1205
BGSU1208	BGSU1210	BGSU1211	BGSU1213
BGSU1215	BGSU1216	BGSU1217	BGSU1223
BGSU1228			

Czech

The first 70 essays:

CZKR1001	CZKR1002	CZKR1003	CZKR1004
CZKR1005	CZKR1006	CZKR1007	CZKR1008
CZKR1009	CZKR1010	CZPR2001	CZPR2002
CZPR2004	CZPR2005	CZPR2008	CZPR2014
CZPR2015	CZPR2016	CZPR2017	CZPR2018
CZPR2020	CZPR2022	CZPR2023	CZPR2027
CZPR2029	CZPR2030	CZPR2031	CZPR2032
CZPR2034	CZPR2039	CZPR2043	CZPR2045
CZPR2047	CZPR2048	CZPR2049	CZPR2050
CZPR2051	CZPR2054	CZPR2055	CZPR2057
CZPR2059	CZPR3009	CZPR3010	CZPR3011
CZPR3012	CZPR3013	CZPR3016	CZPR3018
CZPR3022	CZPR3023	CZPR3025	CZPR3026
CZPR3027	CZPR3029	CZPR3036	CZPR3039
CZPR3040	CZPR3047	CZPR3048	CZPR3049
CZPR3050	CZPR3052	CZPR3053	CZPR3057
CZPR3060	CZPR3061	CZPR4004	CZPR4005
CZPR4008	CZPR4010		

The 15 essays in between:

CZPR4011	CZPR4012	CZPR4013	CZPR4014
CZPR4016	CZPR4017	CZPR4018	CZPR4019
CZPR4020	CZPR4021	CZPU1002	CZPU1004
CZPU1005	CZPU1006	CZPU1007	

The last 25 essays:

CZPU1008	CZPU1009	CZPU1011	CZPU1012
CZUN1001	CZUN1004	CZUN1005	CZUN1006
CZUN1009	CZUN1010	CZUN1011	CZUN1012
CZUN1014	CZUN1016	CZUN1017	CZUN1019
CZUN1020	CZUN1023	CZUN1025	CZUN1026
CZUN1027	CZUN1030	CZUN1031	CZUN1032
CZUN1033			

Russian

The first 70 essays:

RUMO1020	RUMO2001	RUMO2002	RUMO2003
RUMO2004	RUMO2005	RUMO2006	RUMO2008
RUMO2010	RUMO2011	RUMO2012	RUMO2013
RUMO2014	RUMO2015	RUMO2017	RUMO2019
RUMO2020	RUMO2021	RUMO3004	RUMO3005
RUMO3006	RUMO3007	RUMO3008	RUMO3009
RUMO3010	RUMO3011	RUMO3012	RUMO4001
RUMO4002	RUMO4003	RUMO4004	RUMO4007
RUMO4010	RUMO4011	RUMO4012	RUMO4013
RUMO4014	RUMO4015	RUMO4016	RUMO4018
RUMO4019	RUMO4020	RUMO4022	RUMO4023
RUMO4024	RUMO4025	RUMO4026	RUMO4029
RUMO5002	RUMO5004	RUMO5007	RUMO5008
RUMO5009	RUMO5011	RUMO5012	RUMO5013
RUMO5014	RUMO5016	RUMO5017	RUMO5018
RUMO5019	RUMO5020	RUMO5021	RUMO5023
RUMO5024	RUMO5026	RUMO5027	RUMO5031
RUMO5032	RUMO5033		

The 15 essays in between:

RUMO5034	RUMO5036	RUMO5037	RUMO5039
RUMO5040	RUMO5041	RUMO5042	RUMO5043
RUMO5045	RUMO5046	RUMO6001	RUMO6002
RUMO6003	RUMO6005	RUMO6006	

The last 25 essays:

RUMO6008	RUMO6009	RUMO6011	RUMO6014
RUMO6017	RUMO6020	RUMO7001	RUMO7002
RUMO7004	RUMO7005	RUMO7008	RUMO7014
RUMO7015	RUMO7017	RUMO7020	RUMO7022
RUMO7023	RUMO7031	RUMO7033	RUMO7035
RUMO7036	RUMO7038	RUMO7040	RUMO7042
RUMO7046			

French

The first 70 essays:

FRUB1002	FRUB1003	FRUB1004	FRUB1005
FRUB1006	FRUB1007	FRUB1008	FRUB1010
FRUB1011	FRUB1012	FRUB1013	FRUB1014
FRUB1015	FRUB1018	FRUB1019	FRUB1021
FRUB1022	FRUB1024	FRUB1026	FRUB1027
FRUB1028	FRUB1029	FRUB1030	FRUB2003
FRUB2004	FRUB2005	FRUB2006	FRUB2007
FRUB2008	FRUB2009	FRUB2010	FRUB2011
FRUB2012	FRUB2014	FRUB2015	FRUB2016
FRUB2017	FRUB2018	FRUB2019	FRUB2020
FRUB2021	FRUB2022	FRUB2024	FRUB2025
FRUB2027	FRUB2028	FRUB2029	FRUB2030
FRUC1001	FRUC1004	FRUC1005	FRUC1006
FRUC1007	FRUC1008	FRUC1009	FRUC1010
FRUC1011	FRUC1013	FRUC1015	FRUC1016
FRUC1021	FRUC1022	FRUC1023	FRUC1024
FRUC1025	FRUC1027	FRUC1028	FRUC1029
FRUC1030	FRUC1032		

The 15 essays in between:

FRUC1033	FRUC1034	FRUC1035	FRUC1037
FRUC1038	FRUC1041	FRUC1042	FRUC1044
FRUC1045	FRUC1047	FRUC1050	FRUC1051
FRUC1053	FRUC1054	FRUC1055	

The last 25 essays:

FRUC1056	FRUC1060	FRUC1061	FRUC1064
FRUC1065	FRUC1067	FRUC1069	FRUC1071
FRUC1072	FRUC1073	FRUC1074	FRUC1076
FRUC1078	FRUC1079	FRUC1085	FRUC1086
FRUC1089	FRUC1090	FRUC1091	FRUC1092
FRUC1093	FRUC1094	FRUC1096	FRUC1098
FRUC1099			

Spanish

The first 70 essays:

SPM01005	SPM01006	SPM01007	SPM01010
SPM01011	SPM01012	SPM01013	SPM01014
SPM01015	SPM01016	SPM01017	SPM01018
SPM02001	SPM02003	SPM02004	SPM02005
SPM02007	SPM02009	SPM02010	SPM02011
SPM02012	SPM02013	SPM02014	SPM02015
SPM03002	SPM03004	SPM03005	SPM03006
SPM03007	SPM03008	SPM03012	SPM03014
SPM03015	SPM03016	SPM03017	SPM03018
SPM03019	SPM03020	SPM03021	SPM03023
SPM03025	SPM03026	SPM03027	SPM03028
SPM03029	SPM03031	SPM03034	SPM03035
SPM03036	SPM03037	SPM03038	SPM03040
SPM03042	SPM03043	SPM03044	SPM03045
SPM03046	SPM03047	SPM03049	SPM03051
SPM03052	SPM03053	SPM03054	SPM04001
SPM04002	SPM04003	SPM04004	SPM04007
SPM04008	SPM04010		

The 15 essays in between:

SPM04011	SPM04012	SPM04013	SPM04015
SPM04016	SPM04017	SPM04018	SPM04019
SPM04022	SPM04023	SPM04026	SPM04028
SPM04029	SPM04030	SPM04031	

The last 25 essays:

SPM04032	SPM04033	SPM04034	SPM04035
SPM04036	SPM04037	SPM04039	SPM04040
SPM04041	SPM04042	SPM04043	SPM04044
SPM04045	SPM04046	SPM04047	SPM04048
SPM04049	SPM04052	SPM04054	SPM04055
SPM04057	SPM05001	SPM05003	SPM05004
SPM05005			

Chinese

The first 70 essays:

CNHK1001	CNHK1002	CNHK1003	CNHK1004
CNHK1005	CNHK1010	CNHK1013	CNHK1017
CNHK1021	CNHK1025	CNHK1027	CNHK1029
CNHK1030	CNHK1031	CNHK1032	CNHK1033
CNHK1035	CNHK1038	CNHK1041	CNHK1042
CNHK1044	CNHK1048	CNHK1051	CNHK1052
CNHK1053	CNHK1055	CNHK1058	CNHK1059
CNHK1060	CNHK1062	CNHK1063	CNHK1065
CNHK1066	CNHK1067	CNHK1072	CNHK1073
CNHK1075	CNHK1076	CNHK1077	CNHK1079
CNHK1081	CNHK1084	CNHK1085	CNHK1086
CNHK1089	CNHK1091	CNHK1092	CNHK1093
CNHK1095	CNHK1098	CNHK1105	CNHK1107
CNHK1109	CNHK1115	CNHK1117	CNHK1121
CNHK1123	CNHK1128	CNHK1134	CNHK1135
CNHK1136	CNHK1137	CNHK1145	CNHK1149
CNHK1152	CNHK1160	CNHK1164	CNHK1175
CNHK1176	CNHK1177		

The 15 essays in between:

CNHK1181	CNHK1185	CNHK1186	CNHK1188
CNHK1190	CNHK1191	CNHK1192	CNHK1193
CNHK1198	CNHK1199	CNHK1200	CNHK1201
CNHK1206	CNHK1207	CNHK1208	

The last 25 essays:

CNHK1209	CNHK1210	CNHK1211	CNHK1212
CNHK1213	CNHK1214	CNHK1216	CNHK1217
CNHK1218	CNHK1219	CNHK1220	CNHK1221
CNHK1222	CNHK1223	CNHK1225	CNHK1226
CNHK1228	CNHK1229	CNHK1230	CNHK1233
CNHK1237	CNHK1240	CNHK1242	CNHK1244
CNHK1246			

Japanese

The first 70 essays:

JPHI1001	JPHI1002	JPHI1003	JPHI1004
JPHI1005	JPKO1004	JPKO1013	JPKO1017
JPKO1018	JPKO2001	JPKO2002	JPKO2003
JPKO2004	JPKO2005	JPKO2006	JPKO2007
JPKO2008	JPKO2009	JPKO2010	JPKO2011
JPKO2012	JPKO2013	JPKO2014	JPKO2015
JPKO2017	JPKO2018	JPKO2019	JPKO2021
JPKO2022	JPKO2025	JPKO2027	JPKO2028
JPKO2029	JPKO2030	JPKO2031	JPKY1001
JPKY1002	JPNH1001	JPSH1001	JPSH1002
JPSH1003	JPSH1004	JPST1001	JPST1002
JPSW1011	JPSW1014	JPSW1017	JPSW1018
JPSW1022	JPSW1023	JPSW1025	JPSW1028
JPSW1032	JPSW1036	JPSW1038	JPSW1039
JPSW2004	JPSW2006	JPSW2007	JPSW2010
JPSW2011	JPSW2012	JPSW2013	JPSW2015
JPSW2018	JPSW2020	JPSW2021	JPSW3001
JPSW3002	JPSW3003		

The 15 essays in between:

JPSW3004	JPSW3005	JPSW3006	JPSW3013
JPSW3015	JPSW3022	JPSW3025	JPSW3026
JPSW3027	JPSW3028	JPSW3029	JPSW3030
JPSW3031	JPSW4004	JPSW4007	

The last 25 essays:

JPSW4008	JPSW4009	JPSW4010	JPSW4011
JPSW4013	JPSW4019	JPSW4021	JPSW4022
JPSW4025	JPSW4029	JPSW4030	JPTF1004
JPTF1012	JPTF1013	JPTF1014	JPTF1015
JPTF1016	JPTF1017	JPTF1019	JPTF1020
JPTF1022	JPTF1024	JPTF1025	JPTF1041
JPTK1001			

Appendix B

Lists of Function Words

Below are the three lists of function words first introduced and used in Chapter 3: the first list consists of 70 function words used in Mosteller and Wallace (1964); the second is the original list of Miller et al. (1958), in which the first list was extracted from, with 363 function words; and the third is the list obtained from the website of Onix Text Retrieval Toolkit including stopwords commonly used in information retrieval, with up to 398 words. Only the last list is used in the studies in the two subsequent chapters, Chapter 4 and Chapter 5.

The First List: 70 function words

a	had	one	was
all	has	only	were
also	have	or	what
an	her	our	when
and	his	shall	which
any	if	should	who
are	in	so	will
as	into	some	with
at	is	such	would
be	it	than	your
been	its	that	
but	may	the	
by	more	their	
can	must	then	
do	my	there	
down	no	things	
even	not	this	
every	now	to	
for	of	up	
from	on	upon	

The Second List: 363 function words

a
about
above
across
ado
aforesaid
after
again
against
alive
all
almost
alone
along
already
also
although
alway
always
am
among
amount
an
and
anon
another
any
anybody
anyone
anything
anywhere
apart
are
around
as
aside
at
away
awful
awfully
awhile
backward
backwardness
be
because
been
before
beforehand
behind
being
below
between
beyond
both
but
by
can
cannot
canst
could
did
do
does
doing
done
dont
down
during
each
eight
eighteen
eighth
eighty
either
eleven
else
elsewhere
enough
etc
even
ever
evermore
every
everyone
everything
everywhere
except
fairly
farther
few
fifty
first
five
for
forasmuch
foregoing
forever
forth
forty
forward
four
fourteen
fourteenth
fourth
from
further
furthermore
get
gets
got
had
halfdozen
hardly
has
have
having
he
height
hence
henceforth
her
here
herein
heretofore
herself
hes
him
himself
his
hither
how
howbeit
however
hundred
i
if
in
inasmuch
indeed
inner
insofar
insomuch
instead
into
inward
is
it
its
itself
just
keep
kept
later
least
less
lest
likewise
many
may
me
middle
midthirties
might
mighty
millenia
million
millions
mine
more
moreover
most
much
must
my
myself
nay
neither
never
nevertheless
next
nine
ninth
no
nobody
noes
none
nor
not
nothing
now
nowadays
nowhere
of
often
oftentimes
oh
on
once
one
ones

The Second List: 363 function words (continue)

only
onto
or
other
others
otherwise
ought
our
ours
ourselves
outside
outward
over
overmuch
past
per
perhaps
please
plus
pretty
quite
rather
real
really
right
same
second
self
selves
seven
seventy
several
shall
shalt
she
should
shouldest
since
six
sixteen
sixth
sixty
so
some
somebody
someday
something
sometimes
somewhat
still
such
ten
than
that
the
thee
their
theirs
them
themselves
then
thence
there
thereafter
thereby
therefore
therein
thereof
thereon
theretofore
therewith
these
they
thine
thing
things
third
thirds
thirty
this
thither
those
thou
though
thousand
three
thrice
through
throughout
thus
thy
thyself
to
together
too
toward
truly
twelve
twenty
twentyseven
twice
two
twothirds
under
underneath
undoing
unless
until
unto
up
upon
upward
us
very
was
wasnt
way
we
weight
well
were
what
whatever
when
whence
whenever
where
whereas
wherefor
wherefore
wherein
whereinsoever
whereof
whereon
wherever
wherewith
whether
which
while
whilst
whither
who
whom
whose
why
will
wilt
with
withal
within
without
won
would
ye
yea
yes
yeses
yet
you
your
yours
yourself
yourselves

The Third List: 398 function words and stopwords

a
about
above
across
after
again
against
all
almos
alone
along
already
also
although
always
among
an
and
another
any
anybody
anyone
anything
anywhere
are
area
areas
around
as
ask
asked
asking
asks
at
away
back
backed
backing
backs
be
became
because
become
becomes
been
before
began
behind
being
beings
best
better
between
big
both
but
by
came
can
cannot
case
cases
certain
certainly
clear
clearly
come
could
did
differ
different
differently
do
does
done
down
downed
downing
downs
during
each
early
either
end
ended
ending
ends
enough
even
evenly
ever
every
everybody
everyone
everything
everywhere
face
faces
fact
facts
far
felt
few
find
finds
first
for
four
from
full
fully
further
furthered
furthering
furthers
gave
general
generally
get
gets
give
given
gives
go
going
good
goods
got
great
greater
greatest
group
grouped
grouping
groups
had
has
have
having
he
her
here
herself
high
higher
highest
him
himself
his
how
however
if
important
in
interest
interested
interesting
interests
into
is
it
its
itself
just
keep
keeps
kind
knew
know
known
knows
large
largely
last
later
latest
least
less
let
lets
like
likely
long
longer
longest
made
make
making
man
many
may
me
member
members
men
might
more
most
mostly
mr

The Third List: 398 function words and stopwords (continue)

mrs
much
must
my
myself
necessary
need
needed
needing
needs
never
new
newer
newest
next
no
nobody
non
noone
not
nothing
now
nowhere
number
numbers
of
off
often
old
older
oldest
on
once
one
only
open
opened
opening
opens
or
order
ordered
ordering
orders
other
others
our
out
over
part
parted
parting
parts
per
perhaps
place
places
point
pointed
pointing
points
possible
present
presented
presenting
presents
problem
problems
put
puts
quite
rather
really
right
room
rooms
said
same
saw
say
says
second
seconds
see
seem
seemed
seeming
seems
sees
several
shall
she
should
show
showed
showing
shows
side
sides
since
small
smaller
smallest
so
some
somebody
someone
something
somewhere
state
states
still
such
sure
take
taken
than
that
the
their
them
then
there
therefore
these
they
thing
things
think
thinks
this
those
though
thought
thoughts
three
through
thus
to
today
together
too
took
toward
turn
turned
turning
turns
two
under
until
up
upon
us
use
used
uses
very
want
wanted
wanting
wants
was
way
ways
we
well
wells
went
were
what
when
where
whether
which
while
who
whole
whose
why
will
with
within
without
work
worked
working
works
would
year
years
yet
you
young
younger
youngest
your
yours

Bibliography

Ahmed Abbasi and Hsinchun Chen. Applying authorship analysis to extremist-group Web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 2005. 24

Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26 (2):1–29, 2008. 7, 8, 15, 21

Steven Abney. Partial parsing via finite-state cascades. *Natural Language Engineering*, 2(4): 337–344, 1996. 18

Steven Abney, Dan Flickinger, Claudia Gdaniec, Ralph Grishman, Philip Harrison, Donald Hindle, Robert Ingria, Fred Jelinek, Judith Klavans, Mark Liberman, Mitch Marcus, Salim Roukos, Beatrice Santorini, and Tomek Strzalkowski. A procedure for quantitatively comparing the syntactic coverage of English grammars. In Ezra Black, editor, *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, pages 306–311, Pacific Grove, CA, USA, February 1991. Association for Computational Linguistics. 69

Ethem Alpaydin. *Introduction to Machine Learning*. MIT Press, Cambridge, MA, USA, 2004. 89

Øistein Andersen. Grammatical error detection using corpora and supervised learning. In Ville Nurmi and Dmitry Sustretov, editors, *Proceedings of the Twelfth Student Session of the European Summer School for Logic, Language and Information*, pages 1–9, Dublin, Ireland, 2007. 69

Katsura Aoyama, James Emil Flege, Susan G. Guion, Reiko Akahane-Yamada, and Tsuneo Yamada. Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/. *Journal of Phonetics*, 32(2):233–250, 2004. 65

Shlomo Argamon, Moshe Koppel, Fine Jonathan, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *Text*, 23(3):321–346, 2003. 1, 3

Shlomo Argamon, Moshe Koppel, James Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123, 2006. 14, 17

R. Harald Baayen, Hans van Halteren, and Fiona J. Tweedie. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11 (3):121–131, 1996. 17, 22, 66

R. Harald Baayen, Hans Halteren Van, Anneke Neijt, and Fiona Tweedie. An experiment in authorship attribution. In *Proceedings of the Sixth International Conference on the Statistical Analysis of Textual Data (JADT)*, pages 29–37, St. Malo, France, 2002. 17, 22

Masanori Bannai. L2 knowledge of verb placement by Japanese learners of English. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 46(1):1–29, 2008. 32

Shane Bergsma, Matt Post, and David Yarowsky. Stylometric analysis of scientific articles. In *Proceedings of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies*, HLT-NAACL '12, pages 327–337, Montréal, Canada, June 2012. 14, 18, 20, 25, 139

Douglas Biber. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge, UK, 1995. 23

Stephen Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., Sebastopol, CA, USA, 2009. 46, 86

István Bíró, Jácint Szabó, and András A. Benczúr. Latent Dirichlet allocation in web spam filtering. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web*, pages 29–32, Beijing, China, April 2008. 105

David Blei and Jon McAuliffe. Supervised topic models. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 121–128. MIT Press, Cambridge, MA, USA, 2008. 105, 109, 142

David M. Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. 103, 105, 107, 111

Leonard Bloomfield. *Language*. Holt, Rinehart and Winston, New York, NY, USA, 1933. 29

Benjamin Börschinger, Bevan K. Jones, and Mark Johnson. Reducing grounded learning tasks to grammatical inference. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1416–1425, Edinburgh, Scotland, July 2011. 105, 121, 122, 127, 129

Jordan Boyd-Graber and David M. Blei. Syntactic topic models. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 185–192. Curran Associates, Inc., 2009. 142

Ted Briscoe, John Carroll, and Rebecca Watson. The second release of the RASP system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pages 77–80, Sydney, Australia, July 2006. 27, 69

Julian Brooke and Graeme Hirst. Native language detection with 'cheap' learner corpora. Presented at the *Conference of Learner Corpus Research*, University of Louvain, Belgium, 2011. 28

Julian Brooke and Graeme Hirst. Measuring interlanguage: Native language identification with L1-influence metrics. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 779–784, Istanbul, Turkey, May 2012. 25, 28, 141

John D. Burger and John C. Henderson. An exploration of observable features related to blogger age. In *Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 15–20, Stanford, CA, USA, March 2006. 14

John F. Burrows. 'An ocean where each kind. . .': Statistical analysis and some major determinants of literary style. *Computers and the Humanities*, 23(4-5):309–321, 1989. 22

Jill Burstein, Martin Chodorow, and Claudia Leacock. Automated essay evaluation: The *criterion online writing services*. *AI Magazine*, 25(3):27–36, 2004. 6

Yuko Goto Butler. Second language learners theories on the use of English articles: An analysis of the metalinguistic knowledge used by Japanese students in acquiring the English article system. *Studies in Second Language Acquisition*, 24(3):451–480, 2002. 41, 137

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. The Parallel Grammar project. In *Proceedings of the 2002 Workshop on Grammar Engineering and Evaluation*, volume 15 of *COLING-GEE '02*, pages 1–7. Association for Computational Linguistics, Taipei, Taiwan, 2002. 68, 69

Alice Y. W. Chan. Syntactic transfer: Evidence from the interlanguage of Hong Kong Chinese ESL Learners. *The Modern Language Journal*, 88(1):56–74, 2004. 32

Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. 44, 76

Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, NAACL 2000, pages 132–139, Seattle, WA, USA, April 2000. 72, 87

Eugene Charniak and Mark Johnson. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 173–180, Ann Arbor, MI, USA, June 2005. 9, 69, 70, 72, 73, 75, 87, 88, 140

Carole E. Chaski. Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1):1–13, 2005. 22

Mei-Hua Chen, Shih-Ting Huang, Hung-Ting Hsieh, Ting-Hui Kao, and Jason S. Chang. FLOW: A first-language-oriented writing assistant system. In *Proceedings of the ACL 2012 System Demonstrations*, pages 157–162, Jeju Island, Korea, July 2012. 6

Stephen Clark and James R. Curran. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552, 2007. 87

Paul Clough. Old and new challenges in automatic plagiarism detection. In *Proceedings of National UK Plagiarism Advisory Service*, pages 391–407, 2003. 15

Michael Collins. Discriminative reranking for natural language parsing. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML'00)*, pages 175–182, Stanford, CA, USA, 2000. 75

Bernard Comrie and Greville G. Corbett, editors. *The Slavonic Languages*. Routledge, London, UK, 2002. 41

Ann Copestake and Dan Flickinger. An open-source grammar development environment and broad-coverage English grammar using HPSG. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 591–600, Athens, Greece, 2000. 68

Stephen P. Corder. The significance of learners' errors. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 5(4):161–170, 1967. 30, 31, 56

Malcolm Corney, Olivier de Vel, Alison Anderson, and George Mohay. Gender-preferential text mining of e-mail discourse. In *18th Annual Computer Security Applications Conference 2002*, volume 13, pages 282–289. IEEE Computer Society, 2002. 23

Malcolm Coulthard and Alison Johnson. *An Introduction to Forensic Linguistics: Language in Evidence*. Routledge, New York, NY, USA, 2007. 16, 21, 22

Menachem Dagut and Batia Laufer. Avoidance of phrasal verbs — A case for contrastive analysis. *Studies in Second Language Acquisition*, 7(1):73–79, 1985. 36

Daniel Dahlmeier and Hwee Tou Ng. Correcting semantic collocation errors with L1-induced paraphrases. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Edinburgh, Scotland, July 2011. 6

Robert Dale, Ilya Anisimoff, and George Narroway. HOO 2012: A report on the preposition and determiner error correction shared task. In *Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educational Applications*, pages 54–62, Montréal, Canada, June 2012. 6

Fred J. Damerau. The use of function word frequencies as indicators of style. *Computers and the Humanities*, 9(6):271–280, 1975. 16

Olivier de Vel, Alison M. Anderson, Malcolm W. Corney, and George Mohay. Mining e-mail content for author identification forensics. *ACM SIGMOD Record*, 30(4):55–64, 2001. 7, 8, 14, 16, 17, 19, 23

Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. Authorship attribution with support vector machines. *Applied Intelligence*, 19(1):109–123, 2003. 17, 23

Heidi C. Dulay and Marina K. Burt. You can't learn without goofing. In Jack C. Richards, editor, *Error Analysis*. Longman, London, UK, 1974a. 30

Heidi C. Dulay and Marina K. Burt. Errors and strategies in child second language acquisition. *TESOL Quarterly*, 8:129–136, 1974b. 31

Libuše Dušková. On sources of error in foreign language learning. *International Review of Applied Linguistics (IRAL)*, 7(1):11–36, 1969. 32

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1365–1374, Portland, OR, USA, June 2011. 14

Rod Ellis. *The Study of Second Language Acquisition*. Oxford University Press, Oxford, UK, 2nd edition, 2008. 5, 6, 29, 30, 31, 34, 35, 36, 37, 95, 115, 141

Dominique Estival, Tanja Gaustad, Son-Bao Pham, Will Radford, and Ben Hutchinson. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 263–272, Melbourne, Australia, September 2007. 4, 9, 14, 19, 25, 26

Jill M. Farringdon. *Analysing for Authorship: A Guide to the Cusum Technique*. University of Wales Press, Cardiff, Wales, UK, 1996. 21

Song Feng, Ritwik Banerjee, and Yejin Choi. Characterizing stylistic elements in syntactic structure. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1522–1533, Jeju Island, Korea, July 2012. 14, 18

Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to detect phishing emails. In *Proceedings of the 16th International World Wide Web Conference*, Banff, Alberta, Canada, May 2007. 7

George Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, 2003. 74, 87

Jennifer Foster. *Good Reasons for Noting Bad Grammar: Empirical Investigations into the Parsing of Ungrammatical Written English*. PhD thesis, Department of Computer Science, Trinity College, University of Dublin, 2005. 71

Jennifer Foster. Treebanks gone bad: Parser evaluation and retraining using a treebank of ungrammatical sentences. *International Journal of Document Analysis and Recognition*, 10(3–4): 129–145, 2007. 45, 71, 72, 102

Jennifer Foster, Joachim Wagner, and Josef van Genabith. Adapting a WSJ-trained parser to grammatically noisy text. In *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technology — Short Papers*, pages 221–224, Columbus, OH, USA, June 2008. 68, 69, 70, 71, 72, 73

Julie Franck, Gabriella Vigliocco, and Janet Nicol. Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, 17(4):371–404, 2002. 34

Michael Gamon. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 611–617, Geneva, Switzerland, August 2004. 14, 18, 19, 66

Michael Gamon. Using mostly native data to correct errors in learners' writing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 163–171, Los Angeles, CA, USA, June 2010. 6

Michael Gamon, Jianfeng Gao, Chris Brockett, Alex Klementiev, William B. Dolan, Dmitriy Belenko, and Lucy Vanderwende. Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP'08)*, pages 449–456, Hyderabad, India, 2008. 6

Antonio M. García and Javier Calle Martín. Function words in authorship attribution studies. *Literary and Linguistic Computing*, 22(1):49–66, 2007. 14, 16

Nikesh Garera and David Yarowsky. Modeling latent biographic attributes in conversational genres. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ACL/AFNLP, pages 710–718, Singapore, August 2009. 14, 20

Susan M. Gass and Larry Selinker. *Second Language Acquisition: An Introductory Course*. 2008. 6, 29, 30, 31, 34, 35, 36

Gene V. Glass, Percy D. Peckham, and James R. Sanders. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42(3):237–288, 1972. 55

Neil Graham, Graeme Hirst, and Bhaskara Marthi. Segmenting documents by stylistic character. *Natural Language Engineering*, 11(4):397–415, 2005. 23

Sylviane Granger. On identifying the syntactic and discourse features of participle clauses in academic English: native and non-native writers compared. In Jan Aarts, Inge de Mönnink, and Herman Wekker, editors, *Studies in English Language and Teaching*, pages 185–198. Rodopi, Amsterdam, the Netherlands, 1997. 33

Sylviane Granger and Stephanie Tyson. Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, 15(1):17–27, 1996. 32

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. *International Corpus of Learner English (Version 2)*. Presses Universitaires de Louvain, Louvian-la-Neuve, Belgium, 2009. 10, 41, 88, 108, 143

Andrew Gray, Philip Sallis, and Stephen Macdonell. Software forensics: Extending authorship analysis techniques to computer programs. In *Proceedings of 3rd Biannual Conference of the International Association of Forensic Linguists (IAFL97)*, volume 13, pages 1–8. Durham, NC, USA, 1997. 13, 15

Jack W. Grieve. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270, 2007. 16

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl. 1):5228–5235, 2004. 105, 107, 111

Jonathon Guilford. English learner interlanguage: What's wrong with it? *Anglophonia French Journal of English Studies*, 4:73–100, 1998. 32

Yan Guo and Gulbahar H. Beckett. The hegemony of English as a global language: Reclaiming local knowledge and culture in China. *Convergence*, 40(1–2):117–132, 2007. 142

Na-Rae Han, Joel Tetreault, Soo-Hwa Lee, and Jin-Young Ha. Using an error-annotated learner corpus to develop an ESL/EFL error correction system. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pages 763–770, Valletta, Malta, May 2010. 6

Eric A. Hardisty, Jordan Boyd-Graber, and Philip Resnik. Modeling perspective using adaptor grammars. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 284–292, Cambridge, MA, USA, October 2010. 104, 120, 121, 122, 123, 127

Michael R. Harwell, Elaine N. Rubinstein, William S. Hayes, and Corley C. Olds. Summarizing Monte Carlo results in methodological research: the one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, 17(4):315–339, 1992. 55

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. 2nd edition, 2009. 89

Matthieu Hermet and Alain Désilets. Using first and second language models to correct preposition errors in second language authoring. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 64–72, Boulder, Colorado, June 2009. 6

Graeme Hirst and Ol'ga Feiguina. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4):405–417, 2007. 17, 18, 104

David I. Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117, 1998. 13, 21, 24

David I. Holmes and Richard S. Forsyth. The Federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing*, 10(2):111–127, 1995. 16, 22

Johan F. Hoorn, Stefan L. Frank, Wojtek Kowalczyk, and Floor van Der Ham. Neural network identification of poets using letter sequences. *Literary and Linguistic Computing*, 14(3):311–338, 1999. 24

Tania Ionin, Soondo Baek, Eunah Kim, Heejeong Ko, and Kenneth Wexler. That's not so different from the: Definite and demonstrative descriptions in second language acquisition. *Second Language Research*, 28(1):69–101, 2012. 137

Farkhund Iqbal, Hamad Binsalleeh, Benjamin C.M. Fung, and Mourad Debbabi. Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation*, 7(1–2): 56–64, 2010. 8

Shun'ichiro Ishikawa. A new horizon in learner corpus studies: The aim of the ICNALE Project. In George R. S. Weir, Shun'ichiro Ishikawa, and K. Poonpon, editors, *Corpora and Language Technologies in Teaching, Learning and Research*, pages 3–11. University of Strathclyde Press, Glasgow, UK, 2011. 116, 132

Scott Jarvis and Scott A. Crossley, editors. *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*. Multilingual Matters, Bristol, UK, 2012. 139

Thorsten Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML)*, pages 137–142. Springer, 1998. 23, 44

Mark Johnson. PCFGs, topic models, adaptor grammars and learning topical collocations and the structure of proper names. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1157, Uppsala, Sweden, July 2010. 9, 104, 108, 118, 120, 122, 123

Mark Johnson and Ahmet Engin Ural. Reranking the Berkeley and Brown parsers. In *Proceedings of Human Language Technologies: the 2010 Annual Conference of the North American*

Chapter of the Association for Computational Linguistics, HLT/NAACL-10, pages 665–668, Los Angeles, CA, USA, June 2010. 87

Mark Johnson, Stuart Geman, Stephen Canon, Zhiyi Chi, and Stefan Riezler. Estimators for stochastic unification-based grammars. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, ACL'99, pages 535–541, College Park, MD, USA, June 1999. 75

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. Adaptor Grammars: A framework for specifying compositional nonparametric Bayesian models. In *Advances in Neural Information Processing Systems 19: Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*, pages 641–648, Vancouver, Canada, 2007. 118, 119

Michael Allan Jones. *Foundations of French Syntax*. Cambridge University Press, Cambridge, UK, 1996. 41

Eric Kellerman. Crosslinguistic influence: Transfer to nowhere? *Annual Review of Applied Linguistics*, 15(1):125–150, 1995. 35

Vlado Keselj, Fuchun Peng, Nick Cercone, and Calvin Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics (PACLING)*, pages 255–264, Halifax, Canada, 2003. 17, 20

Dmitri V. Khmelev and Fiona J. Tweedie. Using Markov chains for identification of writers. *Literary and Linguistic Computing*, 16(3):299–307, 2001. 24

Bradley Kjell. Authorship determination using letter pair frequency features with neural network classifiers. *Literary and Linguistic Computing*, 9(2):119–124, 1994. 17

Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, July 2003. 70, 72, 73, 88

Howard H. Kleinmann. Avoidance behavior in adult second language acquisition. *Language Learning*, 27(1):93–107, 1977. 36

Ekaterina Kochmar. Identification of a writer's native language by error analysis. Master's thesis, Computer Laboratory, St. John's College, University of Cambridge, 2011. 25, 27, 65, 139

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, MT Summit X, pages 79–86, Phuket, Thailand, 2005. 29

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412, 2002. 14

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. Automatically determining an anonymous author's native language. In *Intelligence and Security Informatics*, volume 3495 of *Lecture Notes in Computer Science*, pages 209–217. Springer-Verlag, 2005. vii, xi, 4, 5, 9, 10, 11, 20, 25, 26, 39, 41, 42, 57, 58, 62, 63, 64, 65, 86, 92, 104, 123, 139

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1): 9–26, 2009. 1, 15, 21, 29

Moshe Koppel, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. Unsupervised decomposition of a document into authorial components. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 1356–1364, Portland, OR, USA, June 2011a. 14

Moshe Koppel, Shlomo Argamon, and Jonathan Schler. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94, 2011b. 14

Robert Lado. *Linguistics Across Cultures: Applied Linguistics for Language Teachers*. University of Michigan Press, Ann Arbor, MI, USA, 1957. 5, 25, 29

James P. Lantolf, editor. *Sociocultural Theory and Second Language Learning*. Oxford University Press, Oxford, UK, 2000. 37, 38

James P. Lantolf and Steven L. Thorne. *Sociocultural Theory and the Genesis of Second Language Development*. Oxford University Press, Oxford, UK, 2006. 37

Robert Layton and Paul Watters. Determining provenance in phishing websites using automated conceptual analysis. In *eCrime Researchers Summit (eCrime) 2009*, Tacoma, WA, USA, October 2009. 8

Robert Layton, Paul Watters, and Richard Dazeley. Automatically determining phishing campaigns using the USCAP methodology. In *eCrime Researchers Summit (eCrime) 2010*, Dallas, TX, USA, October 2010. 8

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. *Automated Grammatical Error Detection for Language Learners*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool, 2010. 6, 68

Gerald R. Ledger and Thomas V. N. Merriam. Shakespeare, Fletcher, and the two noble Kinsmen. *Literary and Linguistic Computing*, 9(3):235–248, 1994. 14, 22

Michael Levy. *Computer-Assisted Language Learning: Context and Conceptualization*. Oxford University Press, Oxford, UK, 1997. 6

Jiexun Li, Rong Zheng, and Hsinchun Chen. From fingerprint to writeprint. *Communications of the ACM*, 49(4):76–82, 2006. 20, 24

Percy Liang, Slav Petrov, Michael I. Jordan, and Dan Klein. The infinite PCFG using hierarchical Dirichlet processes. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 688–697, Prague, Czech Republic, June 2007. 135

Richard L. Light and Diane Warshawsky. Preliminary error analysis: Russians using English. Technical report, National Institute of Education, USA, 1974. 32, 33

Dilin Liu and Johanna L. Gleason. Acquisition of the article *the* by non-native speakers of English: An analysis of four non-generic uses. *Studies in Second Language Acquisition*, 24:1–26, 2002. 116

Lisa M. Lix, Joanne C. Keselman, and H. J. Keselman. Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research*, 66(4):579–619, 1996. 55

Jianbin Ma, Ying Li, and Guifa Teng. Identifying Chinese e-mail documents' authorship for the purpose of computer forensic. In *Intelligence and Security Informatics*, volume 5075 of *Lecture Notes in Computer Science*. Springer-Verlag, 2008. 24

Liping Ma, John Yearwood, and Paul Watters. Establishing phishing provenance using orthographic features. In *eCrime Researchers Summit (eCrime) 2009*, Tacoma, WA, USA, October 2009. 8

Brian MacWhinney. The competition model. In Brian MacWhinney, editor, *Mechanisms of language acquisition*, pages 249–308. Lawrence Erlbaum, 1987. 37

Brian MacWhinney. A unified model. In Peter Robinson and Nick C. Ellis, editors, *Handbook of Cognitive Linguistics and Second Language Acquisition*, chapter 14, pages 341–371. Routledge, 2008. 37

Brian MacWhinney and Elizabeth Bates. *The Crosslinguistic Study of Sentence Processing*. Cambridge University Press, New York, NY, USA, 1989. 37

Shervin Malmasi, Sze-Meng Jojo Wong, and Mark Dras. NLI Shared Task 2013: MQ Submission. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133, Atlanta, Georgia, June 2013. 142

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. 72

John Maxwell and Ronald Kaplan. Unification-based parsers that automatically take advantage of context freeness. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the First International Conference on Lexical Functional Grammar*, Grenoble, France, August 1996. 69

John H. McDonald. *Handbook of Biological Statistics*. Sparky House Publishing, Baltimore, MD, USA, 2nd edition, 2009. 55

Thomas C. Mendenhall. The characteristic curves of composition. *Science*, 9(214):237–246, 1887. 1, 13, 20, 21

Thomas V. N. Merriam and Robert A. J. Matthews. Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. *Literary and Linguistic Computing*, 9 (1):1–6, 1994. 16, 23

George A. Miller, E.B. Newman, and Elizabeth A. Friedman. Length-frequency statistics for written English. *Information and Control*, 1(4):370–389, 1958. 58, 151

Thomas Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 352–359, Alberta, Canada, 2002. 107

Abdul Rashid Mohamed, Li Lian Goh, and Eliza Wan-Rose. English errors and Chinese learners. *Sunway College Journal*, 1(1):83–97, 2004. 32, 33

Andrew Q. Morton and Sidney Michaelson. The Qsum plot. Internal report CSR-3-90, Department of Computer Science, University of Edinburgh, 1990. 21

Frederick Mosteller and David L. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA, USA, 1964. vii, 1, 2, 14, 16, 22, 58, 151

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic, June 2007. 68, 69

Steven Myers. Introduction to phishing. In Markus Jakobsson and Steven Myers, editors, *Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft*, chapter 1, pages 1–29. John Wiley & Sons, Inc., 2007. 7

Dominick Ng, Matthew Honnibal, and James R. Curran. Reranking a wide-coverage CCG parser. In *Proceedings of Australasian Language Technology Association Workshop (ALTA'10)*, pages 90–98, Melbourne, Australia, December 2010. 87

Kamal Nigam, John Lafferty, and Andrew McCallum. Using maximum entropy for text classification. *Science*, 1:61–67, 1999. 75

Tomomi Niimura and Hayashi Brenda. Contrastive analysis of English and Japanese demonstratives from the prespective of L1 and L2 acquisition. *Language Sciences*, 18(3–4):811–834, 1996. 137

Terence Odlin. *Language Transfer: Cross-linguistic Influence in Language Learning*. Cambridge University Press, Cambridge, UK, 1989. 5, 30, 35, 36

Timothy O'Donnell. *Productivity and reuse in language*. PhD thesis, Harvard University, 2011. 119

Daisuke Okanohara and Jun'ichi Tsujii. A discriminative language model with pseudo-negative samples. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 73–80, Prague, Czech Republic, June 2007. 69

John W. Oller and Seid M. Ziahosseiny. The contrastive analysis hypothesis and spelling errors. *Language Learning*, 20(2):183–189, 1970. 32, 35, 57

Nelleke Oostdijk. *Corpus Linguistics and the Automatic Analysis of English*. Rodopi, Amsterdam, The Netherlands, 1991. 17

Fuchun Peng, Dale Schuurmans, Vlado Keselj, and Shaojun Wang. Language independent authorship attribution using character level language models. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics (EACL)*, pages 267–274, Budapest, Hungary, April 2003. 17, 20

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, COLING/ACL'06, pages 433–440, Sydney, Australia, July 2006. 87, 88

Manfred Pienemann. *Language Processing and Second Language Development: Processability Theory*. John Benjamins, Amsterdam, The Netherlands, 1998. 37, 38

Manfred Pienemann, editor. *Cross-linguistic Aspects of Processability Theory*. John Benjamins, Amsterdam, The Netherlands, 2005. 38

Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25(2):855–900, 1997. 119

Matt Post. Judging grammaticality with tree substitution grammar derivations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies — Short Papers*, ACLShort '11, pages 217–222, Portland, Oregon, June 2011. 70, 80

Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 38–42, Uppsala, Sweden, July 2010. 18

Arun Rajkumar, Saradha Ravi, Venkatasubramanian Suresh, M. Narasimha Murty, and C. E. Veni Madhavan. Stopwords and stylometry: A latent Dirichlet allocation approach. In *Proceedings of the NIPS 2009 Workshop on Applications for Topic Models: Text and Beyond (Poster Session)*, Whistler, Canada, December 2009. 106, 113

Delip Rao and David Yarowsky. Detecting latent user properties in social media. In *Proceedings of the NIPS 2010 Workshop on Machine Learning for Social Computing*, Whistler, Canada, December 2010. 14

Delip Rao, Michael Paul, Clayton Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. Hierarchical Bayesian models for latent attribute detection in social media. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, ICWSM, pages 598–601, Barcelona, Catalonia, Spain, July 2011. 14

Payam Refaeilzadeh, Lei Tang, and Huan Liu. Cross-validation. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 532–538. Springer, US, 2009. 89

Jack C. Richards. A non-contrastive approach to error analysis. *ELT Journal*, 25(3):204–219, 1971. 30, 31, 33, 34

Daniel Robertson. Variability in the use of the English article system by Chinese learners of English. *Second Language Research*, 16(2):135–172, 2000. 41, 137

Peter Robinson, editor. *The Routledge Encyclopedia of Second Language Acquasition*. Routledge, New York, NY, USA, 2013. 7

Michal Rosen-Zvi, Chaitanya Chemudugunta, Thomas Griffiths, Padhraic Smyth, and Mark Steyvers. Learning author-topic models from text corpora. *ACM Transactions of Information Systems*, 28(1):4:1–4:38, 2004. 106

Alla Rozovskaya and Dan Roth. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 924–933, Portland, OR, USA, June 2011. 6

Laura Sabourin, Laurie A. Stowe, and Ger J. de Haan. Transfer effects in learning a second language grammatical gender system. *Second Language Research*, 22(1):1–29, 2006. 32

Geoffrey Sampson and Anna Babarczy. A test of the leaf-ancestor metric for parse accuracy. *Natural Language Engineering*, 9(4):365–380, 2003. 69

Jacquelyn Schachter. An error in error analysis. *Language Learning*, 27:205–214, 1974. vii, 30, 35, 36, 97, 115

Jacquelyn Schachter and William Rutherford. Discourse function and language transfer. *Working Papers on Bilingualism*, 19:3–12, 1979. 37

Deborah Schaffer. The language of scam spams: Linguistic features of 'Nigerian fraud' e-mails. *ETC: A Review of General Semantics*, 69(2):157–179, 2012. 8

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. Effects of age and gender on blogging. In *Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205, Stanford, CA, USA, March 2006. 14

Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. Authorship attribution with latent Dirichlet allocation. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 181–189, Portland, OR, USA, June 2011. 106, 111

Yanir Seroussi, Fabian Bohnert, and Ingrid Zukerman. Authorship attribution with author-aware topic models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 264–269, Jeju Island, Korea, July 2012. 106

Roumyana Slabakova. L1 transfer revisited: the L2 acquisition of telicity marking in English by Spanish and Bulgarian native speakers. *Linguistics*, 38(4):739–770, 2000. 32, 41

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009. 1, 13, 15

Efstathios Stamatatos, George Kokkinakis, and Nikos Fakotakis. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495, 2000. 17

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214, 2001. 15, 17

Mark Steyvers and Thomas L. Griffiths. Probabilistic topic models. In T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*, chapter 21, pages 427–448. Lawrence Erlbaum, 2007. 107, 112, 113

Guihua Sun, Gao Cong, Xiaohua Liu, Chin-Yew Lin, and Ming Zhou. Mining sequential patterns and tree patterns to detect erroneous sentences. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, pages 925–930, Vancouver, Canada, July 2007a. 68, 69

Guihua Sun, Xiaohua Liu, Gao Cong, Ming Zhou, Zhongyang Xiong, John Lee, and Chin-Yew Lin. Detecting erroneous sentences using automatically mined sequential patterns. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 81–88, Prague, Czech Republic, June 2007b. 68, 69

Marlene Kendra Sun-Alperin and Min Wang. Spanish-speaking children's spelling errors with English vowel sounds that are represented by different graphemes in English and Spanish words. *Contemporary Educational Psychology*, 33(4):932–948, 2008. 32, 35, 57

Michael Swan and Bernard Smith, editors. *Learner English: A Teacher's Guide to Interference and Other Problems*. Cambridge University Press, 2nd edition, 2001. 3, 94, 98

Benjamin Swanson and Eugene Charniak. Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACLShort '12, pages 193–197, Jeju Island, Korea, July 2012. 25, 28, 100, 101, 139

Matt Tearle, Kye Taylor, and Howard Demuth. An algorithm for automated authorship attribution using neural networks. *Literary and Linguistic Computing*, 23(4):425–442, 2008. 23

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. 142

Guifa Teng, Maosheng Lai, Jianbin Ma, and Ying Li. E-mail authorship mining based on SVM for computer forensic. In *Proceedings of Machine Learning and Cybernetics 2004*, pages 26–29, Shanghai, China, 2004. 23

Joel Tetreault and Martin Chodorow. Examining the use of region web counts for ESL error detection. In *Proceedings of the Web as Corpus Workshop (WAC-5)*, San Sebastian, Spain, September 2009. 6

Joel Tetreault, Jill Burstein, and Claudia Leacock, editors. *Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Montréal, Canada, June 2012. 6

Ivan Titov and Ryan McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, OH, USA, June 2008. 105, 111

Laura Mayfield Tomokiyo and Rosie Jones. You're not from round here, are you? Naive Bayes detection of non-native utterance text. In *Proceedings of the Second North American Chapter of the Association for Computational Linguistics*, NAACL '01, pages 239–246, Pittsburgh, PA, USA, 2001. 4

Oren Tsur and Ari Rappoport. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of ACL 2007 Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16, Prague, Czech Republic, June 2007. 4, 9, 25, 26, 41, 65

Fiona J. Tweedie and R. Harald Baayen. How variable may a constant be? Measure of lexical richness in perspective. *Computers and the Humanities*, 32:323–352, 1998. 15, 16

Fiona J. Tweedie, Sameer Singh, and David I. Holmes. Neural network applications in stylometry: The Federalist Papers. *Computers and the Humanities*, 30(1):1–10, 1996. 16, 23

Hans van Halteren. Source language markers in EUROPARL translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 937–944, Manchester, UK, August 2008. 28, 29

Irena Vassileva. Who am I/how are we in academic writing? A contrastive analysis of authorial presence in English, German, French, Russian and Bulgarian. *International Journal of Applied Linguistics*, 8(2):163–185, 1998. 32

Garbriella Vigliocco, Brian Butterworth, and Merrill F. Garrett. Subject-verb agreement in Spanish and English: Differences in the role of conceptual constraints. *Cognition*, 61(3):261–298, 1996. 34

Joachim Wagner, Jennifer Foster, and Josef van Genabith. A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 112–121, Prague, Czech Republic, June 2007. 69

Joachim Wagner, Jennifer Foster, and Josef van Genabith. Judging grammaticality: Experiments in sentence classification. *CALICO Journal*, 26(3):474–490, 2009. 68, 69, 70

Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26 th International Conference on Machine Learning*, pages 1105–1112, Montreal, Canada, 2009. 111

Chong Wang, David Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1903–1910, Miami, FL, USA, June 2009. 105, 111

Xuerui Wang, Andrew McCallum, and Xing Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pages 697–702, Omaha, NE, USA, 2007. 120, 126

Richard Wardhaugh. The Contrastive Analysis Hypothesis. *TESOL Quarterly*, 4(2):123–130, 1970. 25, 30

Xing Wei and W. Bruce Croft. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference (SIGIR'06)*, pages 178–185, Seattle, WA, USA, 2006. 105

Sze-Meng Jojo Wong and Mark Dras. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61, Sydney, Australia, December 2009. 27, 39, 65, 68, 86, 88, 92, 139

Sze-Meng Jojo Wong and Mark Dras. Parser features for sentence grammaticality classification. In *Proceedings of the Australasian Language Technology Association Workshop 2010*, pages 67–75, Melbourne, Australia, December 2010. 69, 70, 73, 74, 75, 77, 78, 87

Sze-Meng Jojo Wong and Mark Dras. Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1600–1610, Edinburgh, Scotland, July 2011. 27, 28, 67, 90, 100, 101, 139

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. Topic modeling for native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 115–124, Canberra, Australia, December 2011. 103

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. Exploring adaptor grammars for native language identification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 699–709, Jeju Island, Korea, July 2012. 103

Suying Yang and Yue Yuan Huang. The impact of the absence of grammatical tense in L1 on the acquisition of the tense-aspect system in L2. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 42(1):49–70, 2004. 32

Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*, pages 412–420, Nashville, TN, USA, 1997. 74, 87

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, OR, USA, June 2011. 27

George U. Yule. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge, UK, 1944. 15, 21

Karen Zagona. *The Syntax of Spanish*. Cambridge University Press, Cambridge, UK, 2002. 41

Rong Zheng, Yi Qin, Zan Huang, and Hsinchun Chen. Authorship analysis in cybercrime investigation. 2665:59–73, 2003. 7, 20, 24

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393, 2006. 13, 14, 15, 16, 19, 21, 24

Shenghuo Zhu, Xiang Ji, Wei Xu, and Yihong Gong. Multi-labelled classification using maximum entropy method. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 274–281, Salvador, Brazil, 2005. 76

Sven Meyer zu Eissen and Benno Stein. Intrinsic plagiarism detection. In *Proceedings of the 28th European conference on Advances in Information Retrieval*, ECIR'06, pages 565–569, London, UK, 2006. 15