DOCTOR OF PHILOSOPHY

Linguistic identifiers of L1 Persian speakers writing in English

NLID for authorship analysis

Ria Perkins

2014

Aston University

Linguistic Identifiers of L1 Persian speakers writing in English. NLID for Authorship Analysis.

Ria Charlotte Perkins, M.A.

Centre for Forensic Linguistics
School of Languages and Social Sciences, Aston University

A thesis submitted for the fulfilment of the degree of Doctor of Philosophy

December, 2012

Thesis Summary

Institution: Aston University

Title: Linguistic Identifiers of L1 Persian speakers writing in English. NLID for Authorship Analysis.

Name: Ria Charlotte Perkins

Degree: Doctor of Philosophy

Year of Submission: 2012

Synopsis:

This research focuses on Native Language Identification (NLID), and in particular, on the linguistic identifiers of L1 Persian speakers writing in English. This project comprises three sub-studies; the first study devises a coding system to account for interlingual features present in a corpus of L1 Persian speakers blogging in English, and a corpus of L1 English blogs. Study One then demonstrates that it is possible to use interlingual identifiers to distinguish authorship by L1 Persian speakers. Study Two examines the coding system in relation to the L1 Persian corpus and a corpus of L1 Azeri and L1 Pashto speakers. The findings of this section indicate that the NLID method and features designed are able to discriminate between L1 influences from different languages. Study Three focuses on elicited data, in which participants were tasked with disguising their language to appear as L1 Persian speakers writing in English. This study indicated that there was a significant difference between the features in the L1 Persian corpus, and the corpus of disguise texts.

The findings of this research indicate that NLID and the coding system devised have a very strong potential to aid forensic authorship analysis in investigative situations. Unlike existing research, this project focuses predominantly on blogs, as opposed to student data, making the findings more appropriate to forensic casework data.

Keywords: Native Language Identification (NLID), Authorship Analysis, Forensic Linguistics, Persian, Interlanguage.

Acknowledgements

I have always believed in the power of words, but they fail me when I try to express my gratitude to everyone who has helped me along the path to completing this thesis. My parents have given me so much; they taught me the value of education, and everything I have achieved is thanks to them. My Mum showed me what it is to be strong, and my Dad helped me to strategise rather than panic and to savour the moment. Dr Tim Grant, my supervisor has inspired me since I first met him and is one of the nicest people I would ever hope to meet. He has guided me through one of the greatest challenges of my life, and has a true gift for making the most difficult things seem completely feasible. My friends and running buddies have been a constant source of support and encouragement, especially: Baiba, Elie, Jenn, Amina, Margarita, Nat, Andrea, Klaus, Jon, Ene, Dan, Sofia, Tom, and Alec – thank you. Rob has helped me relax and smile in these final months. Special thanks also go to my best friend Maria Coker, who has been there for me for so long and through so much, with a genuine belief in me, countless phone calls and frequently; a cup of courage.

My gratitude also goes to my teachers who have inspired me along the way, particularly Dr Abdi Raifee whose love of Persian is truly infectious. The department of Languages and Social Sciences at Aston University is filled with incredibly supporting colleagues, especially within the Centre for Forensic Linguistics. Prof. Malcolm Coulthard and Dr. Krzysztof Kredens have always been there to answer questions. Virtually everyone I have met within the field of forensic linguistics has been inspiring, encouraging and kind. Special thanks go to Larry Solan and Rob Leonard for their advice and encouragement, and for making me feel like I had something valuable to say. Ann Maguire has been so much more than a dyslexia support tutor to me, she's been a friend. I'd also like to thank the people who shared and completed the questionnaire for Study Three. I asked for quite a lot of time from them and was unable to offer anything in return, except my heartfelt gratitude.

Finally I am grateful to Bailey and Marple. Bailey's constant reminders that nothing is as important as the state of his food bowl help me keep life in perspective.

Table of Contents

List of Tables

List of Figures

Part 1 – Introduction

This thesis introduces research into Native Language Identification (NLID) in relation to the field of forensic linguistic authorship analysis. It focused predominantly on linguistic identifiers of L1 (native) Persian speakers writing in English. The thesis is divided into three parts, each containing three chapters. This first part sets out the introduction and background to the wider research project, with the first chapter introducing the topic and the motivations, and the next two chapters considering relevant existing research. Part two details the principal study, Study One, focusing initially on the methodology in Chapter 4, then the analysis and findings and Chapters 4 and 6. Part three considers the applications of the research, the first two chapters containing two further sub-studies; Chapter 7 detailing Study Two, and Chapter 8, Study Three. The final chapter draws together the findings of the different studies and discusses the conclusions of the research as a whole.

Chapter 1. Introduction to NLID

> ***Uttering a word is like striking a note on the keyboard of the imagination (Wittgenstein, 1953)***

This chapter introduces the topics of this thesis, and lays out exactly what will be discussed as well as setting out the structure of this thesis. The first section, 1.1, introduces the topic of Native Language Identification (NLID), the second section sets out the relevant casework background, discussing several key cases, and demonstrating the casework motivation for this research. Section 1.3 discusses the legal context of NLID and the fourth section, 1.4 sets out the structure of this thesis.

- Chapter 1 – Introduction
  - 1.1 NLID – Introduction
  - 1.2 Relevant case history
  - 1.3 Legal and forensic linguistic context
  - 1.4 Thesis content outline

1.1 NLID – Introduction

Native Language Identification is a specific area of authorship profiling that focuses on identifying an anonymous author's native language. This research aims to investigate what

the features of a native L1 Persian speaker writing in English are. This thesis discusses the development of a system that can be used by forensic authorship analysts to determine whether an anonymous author is likely to be a native Persian speaker. The approach taken is firmly grounded within the field of forensic linguistics, and more specifically within the area of authorship analysis. Authorship analysis can be divided into two main sub-areas; profiling and authorship attribution. Authorship attribution determines which out of a selection of authors is more likely to have written a disputed or anonymous text, where as profiling seeks to answer the question "what kind of Person wrote this text?" (Grant, 2008, p. 222). Authorship profiling can seek to answer a wide range of questions, such as age, gender regional background, or in this case, what their linguistic background or L1 is. Figure 1-1 demonstrates how this research considers authorship profiling to be oriented in relation to forensic linguistics.

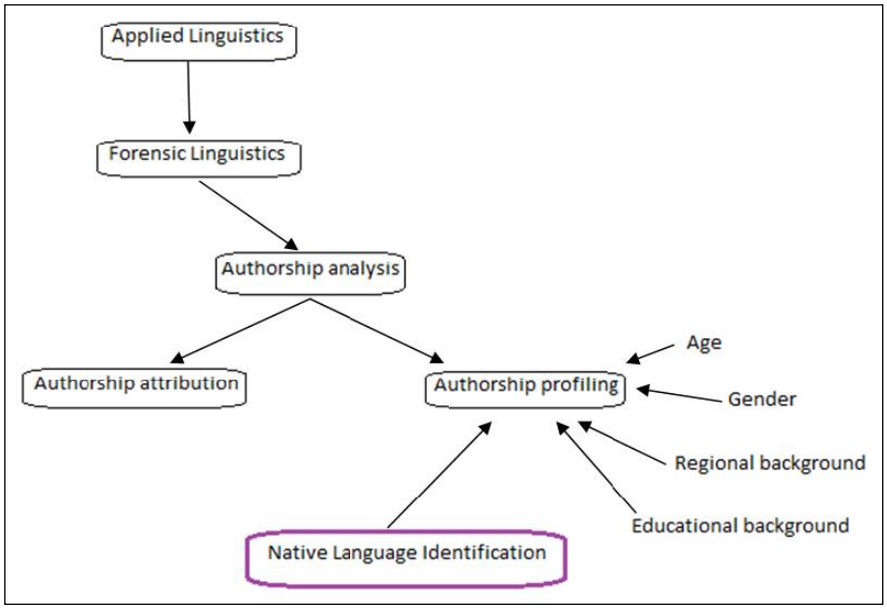


Figure 1-1 Authorship Profiling

The following section, Section 1.2, demonstrates that the belief one can identify someone's native language (L1) from the way they use a second language (L2) is not a new one. Simply described, Native Language Identification (NLID) seeks to indicate an author's native tongue (or L1) from the way they write in a second language (or L2). It is an understudied area of forensic linguistic authorship analysis, yet an area that holds considerable practical potential (Koppel, Schler, & Zigdon, 2005). The potential usefulness is even more significant when we consider the large number of people that operate in a second language on a regular basis. Due to practical constraints, this research focuses primarily on authors writing in English, and more specifically on native Persian speakers writing in English.

The majority of the world's population is bilingual (Thomason, 2001) and English is one of the widest spoken second languages. Exact numbers are virtually impossible to measure, predominantly due to difficulties with defining a second language speaker, but it is estimated that L2 speakers of English could outnumber L1 speakers, with up a quarter of all people having some degree of competency at speaking English (Bhatia & Ritchie, 2004, p. 519). In 2011 it was estimated that there were 565,004,126 internet users using English (Internet World Stats. Usage and Population statistics, 2012), a number even higher than the estimated number of native English speakers. Considering the prevalence of L2 English speakers, it is logical to conclude that a significant percentage of English language forensic texts must also be produced by L2 English speakers. Any text can be a forensic text, it is only the context that determines if it is *forensic*, meaning any text (no matter how mundane) has the potential to later become a forensic text (Coulthard, Grant, & Kredens, 2010; Olsson, 2004). For example, a shopping list could become a forensic text, if it later becomes relevant to a criminal or legal case. This indicates that there is a large area of casework that could benefit from specific NLID research, which to date is limited (relevant literature will be discussed further in Chapter 2).

The term Native Language Identification (NLID) could be seen as misleading, in that it is an oversimplification of the concepts surrounding being a native speaker and language influence, and it seems to indicate that there is a certain outcome. However, the term NLID still has benefits, as what we are trying to do is identify what language(s) an author has a native-like influence from. Perhaps we could term it Native language Indication, but that does not seem to as accurately convey the intention of the analysis. There is also a precedent (albeit limited) behind the term Native Language Identification, and this research is intending to build on the research that exists in this field (S. J. Wong & Dras, 2011; S.-M. J. Wong & Dras, 2009; Koppel et al., 2005; Perkins & Grant, n.d.). Therefore, utilising the same term would serve to add clarity.

1.2 Relevant case history

This research has very clear applications in the field of forensic authorship analysis. It is therefore important to consider some key relevant cases in order to understand what NLID casework questions are being raised, and how they are currently being answered. This section outlines some of the real-life and fictional cases that have provided motivation for this research, and discusses the implications.

The belief that one can identify someone's native language (L1) from the way they use a second language (L2) is not a new one, neither is the inevitable link to the potential forensic applications. In the 1930's case of Bruno Hauptmann, handwriting experts drew on orthographic and linguistic information in the ransom notes to hypothesis that the texts were most likely authored by a native German speaker. On March 1, 1932 the 20 month year old son of famous aviator, Charles Lindbergh was kidnapped. A ransom note was found on the nursery window sill (see Figure 1-2), and later more notes were received. Evidence led the police and FBI to suspect and arrest Bruno Richard Hauptmann, he was later tried in 1935 and found guilty of kidnapping and later killing the Lindbergs' baby. He received capital punishment and was electrocuted in April 1936. One of the key areas of evidence was handwriting analysis of the letters. Although this focused primarily on the orthography of the notes, there was also reference to the linguistic constructions within the texts. The International Herald Tribune wrote in 1935 of Hauptmann's appearance in court that he "spoke slowly, with a strong German accent and frequent Germanic grammatical constructions — such, incidentally, as appeared in the ransom notes he is alleged to have written." (The International Herald Tribune, 2010). Although the requirements for expert evidence have changed significantly since 1935, and hence the testimony given then may not fulfil the Daubert criteria (or equivalent for non-US countries) which regulates the admissibility of expert evidence in court, this case clearly shows that analysis similar to NLID has a place within the legal system. Whether or not NLID can ever hope to achieve the standard required to be admissible in court is discussed later in this thesis (in section 1.3 and in relation to the future in Chapter 9) , but it is important to note that the legal system is not restricted to court room settings and encompasses areas such as helping with investigations.

Figure 1-2 Lindbergh Kidnapping Ransom Note (University of Missouri-Kansas City, 2000)

A fictional example from Sherlock Holmes serves to demonstrate the potential for NLID. In Sir Arthur Conan Doyle's case; A Scandal in Bohemia, the famous fictional detective Holmes receives a letter at his Baker Street residences stating:

> ***There will call upon you to-night, at a quarter to eight o'clock," it said, "a gentleman who desires to consult you upon a matter of the very deepest moment. Your recent services to one of the royal houses of Europe have shown that you are one who may safely be trusted with matters which are of an importance which can hardly be exaggerated. This account of you we have from all quarters received. Be in your chamber then at that hour, and do not take it amiss if your visitor wear a mask." (Conan-Doyle, 1892, p. 8)***

Upon reading the note Holmes initially makes comments on its likely origin, due to physical forensic evidence. He then proceeds to draw conclusions about the anonymous author, stating; "And the man who wrote the note is a German. Do you note the peculiar construction of the sentence--'This account of you we have from all quarters received.' A Frenchman or Russian could not have written that. It is the German who is so uncourteous to his verbs." (Doyle, 1892, p. 8). Essentially he is performing a form of native language identification, based on a linguistic construction in the text, his knowledge of German grammar, and the understanding that structures in a person's L1 can influence target language production. Although his observations are designed to be novel revelations to both Holmes' assistant Watson and the reader, they are intended to be believable and 'elementary' when the evidence is presented to us. This indicates that there is (and has long been) a common public conception that one can identify a person's L1 from choices and constructions evident in a second language.

There are also many genuine cases documented, Hannes Kniffka has cited a case (Kniffka, 1996) in which he has used an anonymous author's use of language to indicate information about their mother tongue. Kniffka was asked to consult on a case in which threatening letters had been sent within a German company. The content indicated that they were likely by an employee of the company. Kniffka analysed the notes and noted that there were some unique linguistic constructions within the German notes, most notably; unusual spelling errors with umlauts, awkward lexical collocations and non-idiomatic use of German proverbs. He determined that these features provided a strong indication that despite the very high level of fluency in the texts, it was most likely that the letters were written by a

non-native (L2) German speaker. This indicated that the initial chief suspect may not be responsible. The company contained two L2 German speakers, one French and one American. They were both put under observation by the police and the American was later caught in the process of writing another threatening letter.

This is a succinct example of when Native language Identification could aid in a forensic context. It demonstrates the possibilities for when it could be of use. For example, if law enforcement officers considered that there was a possibility of a forensic text was authored by a non-native English speaker, we could analyse it to see if their suspicions were right, and what language could be the author's actual native language. Kniffka (1996) noticed that there were certain indicators of authorship by a native English speaker writing in German, this was probably influenced by his knowledge of L1 English students learning German as an L2.

Hubbard (1996) also details a case involving forensic authorship analysis involving a degree of NLID. In 1988 a South African chain store received ten extortion letters at their Johannesburg headquarters. The letters contained threats to poison food on the store's shelves and to alert the media unless a ransom of approximately $500,000 was paid. A fake payment was organised following the letters' demands, however police lost track of the payment. The man who lived directly next to the drop site became the main suspect. He was a medical doctor (and also claimed to hold an engineering doctorate) with German-English parentage, who had spent most of his life in Romania, but claimed Polish as his L1, even requesting a Polish court interpreter. During the trial the defence advocate hired a professor of English, who performed various analyses; most notably a form of error analysis, to determine it was unlikely that the doctor had written the extortion notes. The prosecution then approached Hubbard, whose evidence entailed three main parts: "(a) a critique of the evidence led by the defence witness; (b) a stylometic analysis; and (c) an error analysis" (Hubbard, 1996, p. 125). The accused was found guilty of the extortion charges, which led Hubbard to observe that "error analysis can have forensic value" (Hubbard, 1996, p. 137). Although NLID is distinct from error analysis, in that it describes all of the language, not just errors, there are still parallels that can be drawn.

NLID does not specifically answer all the questions that are raised by these cases, but we can see the potential of NLID as a tool for forensic authorship analysis. There are documented cases that would benefit from greater research into NLID, indicating that research into

Native Language Identification would be a beneficial for future cases. The exact application capabilities of this research will be discussed further throughout this thesis.

1.3 Legal and forensic linguistic context

The two main situations linguists are consulted about in relation to authorship analysis are to provide investigative support, or courtroom evidence. The main difference between these is the degree of certainty that is required from the expert. In order for a linguist to qualify as an expert witness in court they have to satisfy certain criteria: in the United States this is termed the Daubert Standard or Daubert Criteria. The Daubert Criteria superseded the Frye test in the United States for determining if a witness called by any of the parties during federal legal proceedings could be considered an expert witness. The statute states:

> ***RULE 702. TESTIMONY BY EXPERT WITNESSES***
>
> ***A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:***
>
> ***(a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;***
>
> ***(b) the testimony is based on sufficient facts or data;***
>
> ***(c) the testimony is the product of reliable principles and methods; and***
>
> ***(d) the expert has reliably applied the principles and methods to the facts of the case (Federal Committee, 2011)***

In the United Kingdom the legal standard for experts is found in the Court Practice Directions, and in particular in Practice Direction 35 – Experts and Assessors (Civil Procedure Rules, 2010). There are also a series of admissibility tests regarding what constitutes expert evidence. A recent review by the Law Commission recommended the following criteria for expert evidence:

> ***(2) The opinion evidence of an expert witness is sufficiently reliable to be admitted if:–***
>
> ***(a) the evidence is predicated on sound principles, techniques and assumptions;36***
>
> ***(b) those principles, techniques and assumptions have been properly applied to the facts of the case; and***
>
> ***(c) the evidence is supported by [that is, logically in keeping with] those principles, techniques and assumptions as applied to the facts of the case. (The Law Commission, 2011)***

There are clear similarities between the two, they both seek to ensure that the analysis is relevant to the case, has a reliable theoretical and methodological base, and that the analysis being presented utilises the established methods and principles reliably.

Grant (2008) discusses the fact that current methods in sociolinguistic profiling (as in psychological profiling) are seldom accepted as evidence in the UK judicial and court system (Grant, 2008; Perkins & Grant, 2013), instead they have much more use as investigative tools. He also suggests that linguistic profiling, like psychological profiling, will never be able to meet the criteria of evidence. He proposes that while "[u]nderstanding that different sorts of linguistic evidence may play different roles within the investigative and judicial process can be key in pursuing forensic practice A sociolinguistic profile might assist a police investigation but have no evidential value" (Grant 2008 p 224). The unpredictable nature of language and the number of variables involved means that error rates have limited reliability. This research is not intended to be submitted as evidence, however, the criteria for acceptance as legal evidence form a good framework; later research may evolve to become of use in a courtroom context, but we have not attained that level of reliability or validity yet. However, that does not mean we should not strive for it. There are numerous documented cases of forensic or sociolinguistic profiling being of great benefit to investigations and helping achieve justice.

NLID has a great potential to benefit law enforcement agencies or organisations (LEOs) with certain cases, when used within the framework of authorship analysis. However, areas of law enforcement that would likely benefit most significantly from research into NLID are those that deal especially with non-native speakers, particularly intelligence agencies. In fact

this research project related to part of a wider research project at the Centre of Forensic Linguistics at Aston University, which has already garnered some interest and funding from British Intelligence Agencies. Due to the nature of intelligence work, casework in this field is rarely published; however it is still possible to allow intelligence applications to shape research into NLID through collaboration with relevant and interested agencies and departments, this can be seen with reports such as Grant, Kredens, & Perkins (2010).

1.4 Thesis Outline

This thesis is divided into three main overarching sections: Part One - Background, Part Two - Features of L1 Persian speakers blogging in English, and Part Three - Application. Firstly the thesis considers the key background areas relating to this research; this includes introducing the topics and the literature reviews. Chapter One - the Introduction - sets out the motivation and basic background for this research, as well as introducing the purpose and structure of this research. Chapter Two is the first of two literature review chapters. It focuses on the sociolinguistics and pedagogical background of interlanguage and cross-linguistic influence, and provides a description of the Persian language past and present. The second literature review, Chapter Three, considers existing literature relating to the specific forensic linguistic context of this research, more specifically forensic authorship analysis and the implications of internet language as a data source in forensic linguistics. It also considers the specific context on Persian blog authors and introduces the phenomenon of Weblogistan.

The middle section of this thesis, Part Two – Features of L1 Persian Speakers Blogging in English is comprised of Chapters Four, Five and Six, and details the main body of research, including the methods and the results. Chapter Four sets out the outline for the methodology of the research study as a whole, it considers the aims, relevant existing theories and schools of research, and then it draws on these to outline how these manifests in the structure of this research. Chapter Four also considers the ethics of this research, the data chosen, and details the measures taken to ensure this research is as ethical as possible. Chapter Five and Six together detail the main study: Study One, Internet Data. Chapter Five discusses the creation of the coding system and a replicable method, as well as the results and observations from the coding of the data and the initial analyses. Chapter Six builds on the initial analyses in Chapter Five, and through statistical analyses garners greater insight into the data, including determining which features are best at discriminating native Persian blog authors from native English ones, and considers the accuracy of the analysis.

The third and final section Part Three - Application (Chapters Seven, Eight, and Nine) considers the applications of this research; this comprises two further sub-studies and a detailed discussion of the results, findings and implications. Chapter Seven details the second study of this research project, this study seeks to test whether we can distinguish influence from L1 Persian speakers as opposed to L1 speakers of other languages. Chapter Seven discusses the motivations for the languages focused on, as well as the methodology employed, before discussing the findings and the implications of this study. Chapter Eight discusses the third sub-study, which focuses on disguise data. The casework motivation for this study is described, as well as the methodologies employed, followed by the findings and the implications. Chapter Nine draws together all the areas of the wider research project. The first section summarises the findings with direct relation to the aims set out in Chapter Four section One. The practical limitations of these findings are then discussed, as well as the casework applications. Finally the future for this research and the wider field of Native Language Identification are considered.

Chapter 2. Literature Review Interlanguage and Persian language.

The language I have lern'd these forty years,

My native English, now I must forego;

And now my tongue's use is to me no more

Than an unstringed viol or a harp

(Shakespeare, 1825, Richard II Act 1 Scene. 3)

This chapter is divided into two main sections. The first main section (Section 2.1) looks at the sociolinguistic setting of this research; more specifically, at interlanguage. Within the section there are three further subdivisions. Section 2.1.1 considers interlanguage; its background and relevance, Section 2.1.2 considers the wider sociolinguistic setting of language contact, globalisation and world Englishes. Finally Section 2.1.3 ties the literature from the first two sections together and discusses whether we should consider cross-linguistic influence as an individual or social phenomenon. It also considers the implications of the literature discussed in this research. The second main section (2.2) focuses specifically on the Persian language context. The first subsection (2.2.1) considers a brief background to the Persian language. This sets the background for the second section (2.2.2), which describes the Persian language as it is today. The third section (2.2.3) then describes all relevant existing literature which describes potential Persian-English interlingual features that we may expect to find in the data.

- Chapter 2 - Literature Review Interlanguage and Persian Language
    - 2.1 Interlanguage / cross-linguistic influence
        - 2.1.1 Interlanguage background
        - 2.1.2 Language Contact and Globalisation
        - 2.1.3 Social versus individual influence
    - 2.2 Persian language
        - 2.2.1 Persian language background
        - 2.2.2 Modern Persian
        - 2.2.3 Persian-English Interlanguage

2.1 Interlanguage

2.1.1 Interlanguage background

The term *interlanguage* was originally coined by Selinker (1972). However, the concept it referred to had existed long before then and can be seen in early research into contact linguistics and most notably the research of Weinreich (1953) who discussed "Interlingual identifications", as well as interference, a term he described as implying "the introduction of foreign elements" (Weinreich, 1953, p. 1). Selinker introduced the interlanguage hypothesis as "the existence of a separate linguistic system based on the observable output which results from a learner's attempted production of a TL [Target Language] norm." (Selinker, 1974, p. 35). Target Language is comparable to L2 or Second language though emphasises that the language is learnt rather than acquired.

There are two main approaches to interlanguage, one which views it as a failed attempt at a target language or L2, and one which accepts it as a code in its own right. The first approach can be exemplified by Nemser's description of an 'approximate system' (the term he uses instead of interlanguage) as the "deviant linguistic system actually employed by the learner attempting to utilise the target language" (Nemser, 1974, p. 55). This definition has a clear focus on error, or the failure of a learner to successfully re-create the target language. It is not clear, however, what a successful reproduction of the target language would entail; whether that is native-like language or language that is 'grammatically perfect'. This is clarified by Kachru and Nelson (1996) who explain that in this approach the *"inter*-prefix refers to the notion that the linguistic system any given learner or community of learners or users, had at any particular moment is quantitatively and conceptually somewhere between the first language and the target." (Kachru & Nelson, 1996, p. 80). They proceed to highlight the difficulties of what the 'target' language exactly is. Although their perspective relates directly to classroom teaching and inner- or outer-circle variants of English, the questions they raise are also valid for this research, highlighting the limitations of this theory. Figure 2-1 is a graphical representation of this early, restricted approach.

Figure 2-1 Basic Interlanguage Representation

This prevailing sense that interlanguage is nothing more than an unsuccessful attempt at another language, influenced only by the learner's L1 and TL, is now more frequently seen as not fully accounting for the entire process. The alternative approach, which views interlanguage as a linguistic system in its own right (which is inevitable and acceptable) and which has influence not just from the L1 and TL (target language) but also from language learning strategies (Appel & Muysken, 1987; Corder, 1974a) now seems more prevalent. Recent studies (in particular Richards, 1971) have also indicated the regular occurrence in interlanguages of features that cannot be explained by L1 or TL influence. This second approach demonstrates that if research assumes that the only influences on language production are the L1 and TL, then there is the risk of limiting the potential findings according to preconceived notions. While this does not account for what exactly the TL is, it demonstrates that theory cannot be certain in accounting for all the influences on language production. This indicates that the most appropriate theory to follow for this research is a theory that allows for more potential influences, yet allowing for the possibility that existing research is not yet able to account for all the influences. This can be represented as Figure 2-2

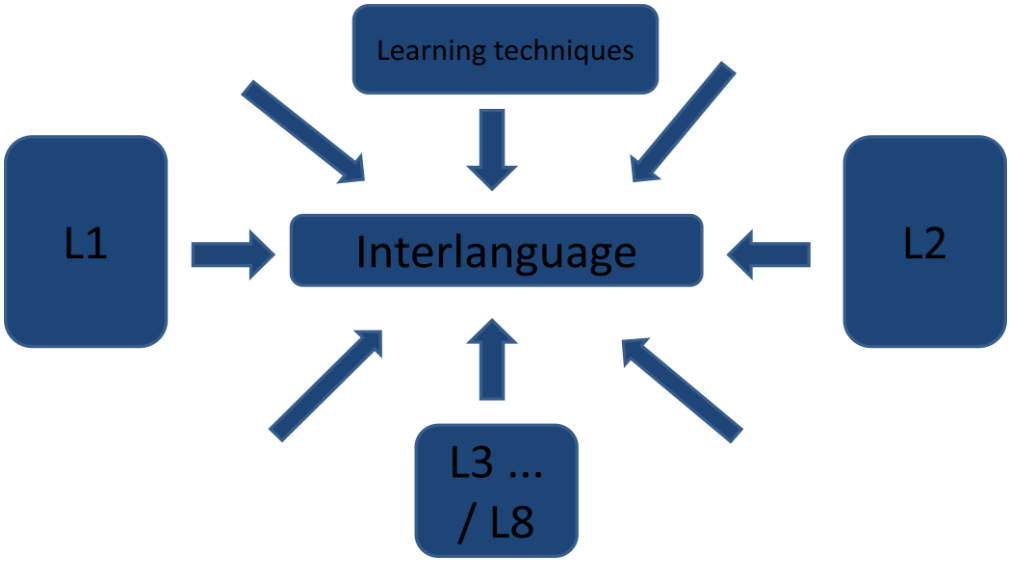


Figure 2-2 - Interlanguage model

De Angelis (2005) noted that "our understanding of what can and cannot be transferred from one non-native language to another in the process of acquisition remains quite limited to date" (Angelis, 2005, p. 380). Although she was referring specifically to transfer *within* L2, it serves to highlight the lack of ability to account for all influences on interlanguage or L2 production. Although this research is not focused on the cause of interlanguage features, it will work on the assumption that there are numerous potential influences on interlanguage so as not to exclude any potential interlanguage features.

Existing literature on interlanguage can be divided predominantly into two main types; purely theoretical and research driven. Early literature on interlanguage, interference and approximative systems, such as studies by Weinreich (1953), and Selinker (1972) relied very little (if at all) on actual data. Weinreich gave an example of German-English interlanguage "he comes tomorrow home" (Weinreich, 1953, p. 30), he explains that this is due to grammatical interference of the L1 German word order, but does not state where this example is from, insinuating that it is his own creation. He later asserts with reference to this example, that "such interference in the domain of grammatical relations is extremely common in the speech of bilinguals" (Weinreich, 1953, p. 37), though this observation is not supported and quantified to today's standards. Despite this Weinreich's work was the catalyst for much of the later research in interlanguage, his theorising informed later

research such as Haugen, 1966; and Heine & Kuteva, 2005, which took a more evidentiary approach. The theorising work has raised many interesting theories that are relevant for this research. However, it is the research driven approach that is of most use to the research presented in this thesis, as the field of forensic linguistics is focused on how language is actually used, rather than assumptions (however well founded) or 'rules' about how language should be used.

The majority of research on interlanguage is situated within the field of second language acquisition (SLA). In Hopkin's article entitled *Contrastive Analysis, Interlanguage, and the Learner* (1982) he wrote that "CA would be able to predict the errors of the [foreign language] learner (cf. Wardhaugh's "strong hypothesis" [1970]) and provide an integrated and scientifically motivated basis for error therapy, textbook construction, etc." (Hopkins, 1982, p. 32). It should be noted that my own research is not focused on eliminating interlanguage features. Instead the aim is to identify features common to L1 Persian speakers writing in English, which can then form the basis of a methodology indicating that an author of an anonymous forensic text has a native knowledge of Persian. There is a very limited field of research into interlanguage from a forensic linguistic perspective. Chapter 1 section 2 introduced a few documented cases that involve a form of NLID, most notably; Hubbard (1996) and Kniffka (1996). Both Hubbard and Kniffka predominantly set out the methodology they employed in the cases, with limited focus on research to build tools to aid with wider NLID cases. There are limited publications documenting research into NLID in the field of forensic authorship attribution, most of which take a computational approach (Koppel et al., 2005; Tomokiyo & Jones, 2001; Tsur & Rappoport, 2007; S. J. Wong, 2012; S. J. Wong & Dras, 2011; S. J. Wong, Dras, & Johnson, 2011; S.-M. J. Wong & Dras, 2009). Koppel, Schler and Zigdon (2005) are credited as being one of the founding initial works in this field (S. J. Wong et al., 2011b). Koppel et al. used text mining and error analysis to determine authors' native languages. Using the International Corpus of Learner English, they looked at Czech, French, Bulgarian, Russian and Spanish speakers writing in English, and created a fully automated system based on function words, letter n-grams and errors and idiosyncrasies. Wong, Dras and Johnson (2001) also use the ICLE corpus for their research, this mirrors the problem with interlanguage research from a pedagogical perspective, in that learner essays are not generally comparable with forensic data.

Fully automated approaches also risk ignoring the capacity for complexity that abounds forensic data; text classification of student essays has considerably fewer variables (most significantly the lack of disguise or threatening topics), than authorship analysis of real-life forensic data. The definition of interlanguage in relation to this research is not necessarily the same as in existing pedagogical research, due to the different aims and focuses. For the purposes of the research reported in this thesis, interlanguage is considered to be the language produced by an individual when they are speaking in any language that is not their L1.

2.1.2 Language contact and Globalisation

The concept of interlanguage evolved from the broader area of language contact. A succinct yet simple definition of language contact is given by Thomason (2001) who explains that it is "the use of more than one language in the same place at the same time" (Thomason, 2001, p. 1). Thomason also asserted that "language contact is everywhere" (Thomason, 2001, p. 8) and that there is no language that has not been subjected to contact with another language. It is therefore unsurprising that there are such a wide range of effects of language contact; from slight lexical borrowing to the formation of new languages (Winford, 2003). Multilingualism and second language acquisition (SLA) are also normally classed as forms of language contact. These are of particular interest to this project, as by definition an L1 Persian speaker writing in English must have experience of multilingualism or second language acquisition. There are also links between forensic authorship analysis and language contact. Foster (2001) sets out an approach to authorship analysis on the basis of the linguistic and literary environment indicating that 'you are what you read', and although his approach has been much criticised in the field it does draw on aspects of language contact.

Winford (2003) focuses on "bilinguals who exploit the resources of the languages they command in various ways, for social and stylistic purposes" (Winford, 2003, p. 101) (Winford, 2003, p. 101). He has previously he commented in the same publication that "[b]orrowing may also provide speakers with stylistic choices" (Winford, 2003, p. 39). He demonstrates that increased language contact, or knowledge of more languages gives people the opportunity to make more stylistic voices in relation to their language. This is reminiscent of David Crystal's view that the increase in 'text speak' or 'Netspeak' allows people to be more creative with their language (Crystal, 2005). While there will be

undeniable social motivations behind the stylistic choices a person makes; there is still a high degree of individuality as each individual will use the languages or codes in discreetly different ways.

> ***"While travel and migration are certainly responsible for much linguistic contact beyond geographical confines, they are by no means exclusive. In times of technical advancement and globalization, language contact has long ceased to be restricted to personalized face-to-face encounters or to the original locus of a language." (Muehleisen, 2002, p. 175)***

Globalisation has frequently been discussed as a cause of language contact (Clyne, 2003; Muehleisen, 2002). "The term globalization is most commonly used as shorthand for the intensified flows of capital, goods, people, images and discourses around the globe, driven by technological innovations, mainly in the field of media and information and communication technology, and resulting in new patterns of global activity, community organization and culture" (Blommaert, 2010, p. 13). Although the physical distance remains the same, it is dramatically easier and quicker to access people, information and ideas on the other side of the planet. This leads to increased interaction between people from different countries; people with different codes of communication. One can understand the tendency to view phenomena such as *globalisation* and *mass media* as social; the words themselves indicate collectivity rather than individuality and presuppose that we should focus on them as social entities. The next section in this chapter (section 2.1.3) considers the interplay between the social and the individual in relation to interlanguage and this research.

The internet plays a key role not only in globalisation, but also more specifically in language contact. Blommaert labels it the "defining technology of globalization" (Blommaert, 2010, loc. 1537). Winford explains that "[a]nother type of "distant" contact leading to lexical borrowing can be found in the spread of global avenues of communications such as radio, television, and the internet." (Winford, 2003, p. 31). Weblogistan (see section 3.1) can therefore be seen as an instigator of language contact and language change. This is one of the reasons why it is a contemporary and valuable data source. A concept that overlaps with interlanguage as discussed in this thesis, is that of world Englishes. Bhatt defines world

Englishes as “varieties of English used in diverse sociolinguistic contexts” (Bhatt, 2012, p. 527). Warschauer, Black and Chou, (2010) examine the impact of the internet on the concept of world Englishes. It could be considered that the English used in weblogistan (see section 3.1) constitutes a specific world English. This distinction is, however, predominantly a terminological one in relation to my own research, and would have no impact on the analysis or findings of this research, as the language of weblogistan is undeniably influenced by both Persian and the online context (as discussed in section 3.4). The focus of this research is on interlingual features present in the language, rather than how the variety is defined.

2.1.3 Social versus individual influence

> ***Language is primarily a cultural or social product and must be understood as such – (Sapir, 1949, p. 166)***

Language is an undeniably social activity, the very concept of communication means that there must be a degree of consistency across individual people’s languages, in order for people to be able to communicate with each other. Without the mutual understanding that is enabled through the similarity of individuals’ languages, there could be no communication. Yet the language each person chooses to express an idea or concept will vary considerably from person to person, indicating that language is also a distinctly individual entity, with each person having a unique phraseology, lexicon and grammar. This duality within language itself indicates that it is not simple to determine whether language contact should be examined as an individual or social phenomenon.

In her 1996 book “The Linguistic Individual”, Barbara Johnstone demonstrates the importance of analysing linguistics in relation to the individual, rather than only focusing on societal trends, and that by focusing on the individual one can learn significantly more about linguistics as a whole. If we consider the earlier quote from Thomason that “Language contact is everywhere” (Thomason, 2001, p. 8) in relation to Winford’s assertion that “Every outcome of language contact has associated with it a particular kind of social setting and circumstances that shape its unique character” (Winford, 2003, p. 10), then it becomes clear

that we must treat each individual's interlanguage as unique. Each person has a unique linguistic history and a unique history of language contact. The concept that all the language a person encounters has the potential to influence their language links to Foster's approach to forensic linguistics. Foster (2001) documents some of the cases he has worked on, and argues that many and varied elements that can influence our language. He posits that an individual's linguistic choices are affected by any linguistic input they have received throughout their lives. His work has drawn criticism from linguists, as the majority of his assertions are based on literature and have little empirical support, Chaski notes that his "approach is purely speculative and relies completely on the literary critic's subjective response to the words, it is impossible to test scientifically" (Chaski, 2001). While Chaski's cautions are valid, Donald Foster's work can still be seen as demonstrating the potential any language we encounter has to influence the language we use, even though it is impossible to quantify this, and as Chaski notes, it is very subjective.

The concept of an each individual having a distinct pattern and style of speech is one that has been discussed extensively in relation to idiolect (Chaski, 2001; Grant, 2007; Kredens, 2000; McMenamin, 2001; Olsson, 2004). There are numerous approaches to idiolect. Grant (2010) separates these into two main groups; cognitivist and stylistic theories, but ably demonstrates that they need not be seen as two opposing approaches. The first group of theories is most commonly associated with a cognitive model of language production. Research that uses this approach has demonstrated a uniqueness that is consistent across an individual's language. Kredens' doctoral research (Kredens, 2000) indicates that a person's idiolect does not change significantly over time, and elements of 'uniqueness' can be seen over long periods of time. While idiolect clearly investigates language and being unique to each individual, the majority of researchers in this field allow for the influence of social factors. As Blommaert states; "Our real 'language' is very much a biographical given, the structure of which reflects our own histories and those of the communities in which we spent our lives." (Blommaert, 2010, p. 103). This is supported by Winford (2003) who posits that the results of language contact (such as code-switching, or borrowing) dramatically increases the number of stylistic choices available to an individual. However, this increased individuality still bears witness to the influence of social factors, as demonstrated by Winford's (2003) observation that code-switching often follows a distinct pattern that is prevalent throughout the society. The data being focused on in this research is predominantly formed of online weblogs, which can be considered a society in their own right, particularly when considering the entity of Weblogistan. There are certain conventions

within online communities (discussed further in section 3.4) which will affect the language produced by each individual author, and the impact of this should not be disregarded. Politics also plays a considerable role in language contact (Winford, 2003); this role of politics from a Persian language perspective, and particularly with reference to weblogistan is discussed more thoroughly in Section 3.1.

There is no simple answer as to whether language contact should be viewed as an individual or a social phenomena. Due to the nature of language, both the individual and the social are completely intertwined. One major factor in how a researcher will view language contact is their research question or their aim. This research aims to find out how L1 Persian speakers write in English, in order to identify and document which features exist, are standard in this social group, and are the best for aiding in determining the L1. While this group is made up of individuals, each with a personal writing style or idiolect, the predominant focus is on the similarities across this society, rather than the authors' unique features. However, the intended application is for the analysis of individual authors. The aim of the research will of course determine whether one favours the individual or the social, but it would be impossible to truly separate forms of language contact into individual and social – it would also not be productive to do this for this research. For this study in particular, there will be times when it is useful to view a form of language contact from a societal perspective, and times when the same form of language contact could be viewed from the perspective of its effects on the individual. There are also other factors, such as politics and creativity, which have an impact on language. One must therefore conclude that it is not useful to examine different forms of language contact as either individual or social phenomena, instead they should be examined as a mixture of individual, social and political phenomena.

2.2 Persian Language

Persian (also called Farsi) is the national language of Iran. It is the most frequently used modern Iranian language (Comrie, 2001; Mahootian & Gebhardt, 2007), but is only the native language for approximately half of the population of Iran (Mahootian & Gebhardt, 2007). The Persian word *Farsi* stems from the word *Pars* which was an area of Persia during the Archaemenid Empire. This was changed to *Farsi* due to the influence of Arabic, in which there is no *p* phoneme. Farsi is the Persian word for the Persian language, (this is comparable to the difference between German and Deutsch) written فارسی. However, recently people have started borrowing the term Farsi into regular English usage. This has provoked some anger on internet discussion boards, as well as provoking much debate (Suren-pahlav, 2007). This thesis will predominantly use the term *Persian*, unless it is

necessary to emphasise or clarify a certain point, some secondary literature favours the term *Farsi* instead of *Persian*. Persian is an Indo-European language. Within the Indo-European group of languages there are many different branches, including; Baltic, Celtic, Germanic, Indo-Iranian and Italic. Persian belongs to the Indo-Iranian group, which has two subsections; Indo-Aryan and Indo-Iranian. Persian belongs to the south-western group of the Indo-Iranian branch. Within Persian there are several dialectal groups; the Farsi of Iran (the language of Tehran is commonly seen as being the official version), Dari (which is prevalent in Afghanistan), and Tajik (which is a variant spoken in Tajikistan). Dari and Tajik are occasionally treated as separate languages rather than dialects of Persian, but as there is very little lexical, grammatical or phonological variation, this research will adhere to the more prevalent view that they are variants of Persian.

2.2.1 Background of the Persian language

Persian belongs to the south-western group of Iranian languages which is a subcategory of the Indo-Iranian language group. Modern Persian is descended from Middle Persian, which in turn descended from Old Persian. The wider languages of Iran can be divided into three main sections: Old Iranian, Middle Iranian and Modern Iranian. These roughly correspond to the Achaemenid period (558-330 BCE), the Sasanid era (224-651 CE) and the modern Islamic period (Skjaervo, 2009; Windfuhr, 2009). The geographical area covered by the languages within these three groups has stayed relatively consistent, though as Kent notes "at all periods there have been islands of non-Iranian speech within it [the approximate area of Iranian speakers], and islands of Iranian speech outside it." (Kent, 1953, p.6) The area covered by the Iranian languages of all periods, relates approximately to the Iranian Plateau.

There are two main languages within the group of languages under the heading *Old Iranian*: Old Persian and Avestan. There are also numerous other languages belonging to the same group, the most significant of which was Median (R. G. Kent, 1936, p. 7). The other languages in the group include: Sogdian, Sakan (Scythian), Carduchi, and Parthian. Old Persian (525BCE – 330 BCE) refers to the language of Southwest Persia (in particular Persis) and relates specifically to the language found in the cuneiform inscriptions. Kent referred to Old Persian as "the vernacular speech of the Archaemenian rulers." (R. G. Kent, 1936, p. 6). Old Persian is believed to be the main root of Middle Persian, which later led onto Modern Persian. It was written in *Old Persian Cuneiform* script which was predominantly written on clay using a reed. This creates the characteristic wedge-shaped impressions which gave cuneiform its name. The cuneiform script evolved from pictographic script of the Sumerians of southern Mesopotamia into a form of semi-alphabetic syllabary, this required far fewer

lines. The German philologist Georg Friedrich Grotefend (1775 - 1853) contributed significantly to the translating of Old Persian cuneiform. He made significant early steps to understanding and hence deciphering the script. His work was expanded upon by Christian Lassen and George Rawlinson among others. Old Persian was the first cuneiform script to be translated and as Grotefend predicted, it held the key to the deciphering of other languages written in cuneiform. There are some surviving examples of the language, mostly from the time of Darius and his son Xerxes (522-465 B.C). As can be seen in the artefacts like the Tablet of Persepolis or the Daiva inspription of Xerxes the inscriptions were often accompanied by translations in both Elamite and Akkadian or Babylonian, which used a different form of cuneiform script. Old Persian was also, occasionally, accompanied by translations in the hieroglyphic languages of Egyptian and Aramaic. The language shows features of the south-west Iranian dialect, for example a change in palatal stops more congruent with south-west Iranian dialects than other Iranian languages. Morphologically the number of cases was reduced from eight to six, as the dative form was merged into the genitive, and the instrumental and ablative forms also joined together. There were three genders; masculine, feminine and neuter, and nouns, adjectives and pronouns distinguished between three levels of plurality; singular, dual or plural.

Avestan is known purely through Zoroastrian scripture. Zoroastrianism is a religion which follows the teaching of the prophet Zarathustra, also known as Zoroaster. The main collection of religious Zoroastrian texts is the Avesta, which demonstrates various varieties of Avestan. These different varieties can be split into *Old Avestan* and *Young Avestan*; Young Avestan can in turn be split into further subsections *Old Young Avestan* and *Late Young Avestan*. It is believed that the Avestan texts represent only a small fragment of the Avestan language, which was based predominantly on oral traditions with texts being remembered by specially trained priests and some of these traditions being transferred to writing during the Sasanian Period. Avestan was linguistically similar to Old Persian, indicating their shared roots. Khanlari relates this to them having been "the dialects of the same Old Iranian language with minor variations" (Khanlari, 1979, p. 163). However, Middle Persian generally acknowledged as descending most directly from Old Persian.

The term Middle Persian covers the Iranian languages from approximately 300 B.C. to 900 A.D., including the official languages of the Arsacid and Sasanian dynasties: Arsacid Pahlavi and Southwest or Sasnian Pahlavi respectively. Middle Persian is also referred to as Pahlavi which is a direct derivative of the Old Persian word *Parθava* or 'Parthian' (Kent, 1953). It is

written in the Aramaic-based Pahlavi script. Arsacid Pahlavi survived longer than the Arsacid dynasty, though later versions were often referred to as Parthian, or Chaldeao-Pahlavi. Pahlavi is also known as the language of the Sasnians (Williams, 2009) as the Sasnian dynasty made it the official language of Iran. There are strong indicators that Sasnian Pahlavi was derived from Old Persian. Williams (2009) notes, that one of the most significant features of Pahlavi is that its orthography incorporates both Iranian and Semitic words, even though the phonetic background is predominantly Iranian. This led to Book Pahlavi (a form of written Pahlavi most commonly associated with Zoroastrian religious texts) comprising almost equal portions of Aramaic heterograms and Iranian eteograms (Williams, 2009, p. 827).

2.2.2 Modern Persian

Like Arabic, Persian is written horizontally, right to left. Persian uses the Arabic alphabet, with a few modifications to account for phonemes that are not present within the Arabic language. *Perso-arabic script* is the term used to refer to this type of writing system that has derived from the Arabic script, though not all scripts that have derived from the Arabic script are Perso-arabic script. Other languages that use the Perso-arabic script include Ottoman Turkish, Pashto, Urdu, Saraiki, Kurdish, Tatar and Azeri. Perso-arabic script is a cursive script which comes from the Latin *cursivus* or *flowing*, relating to the fact that the individual symbols are joined up when written to enable quicker writing; unlike in English this also applies when writing on a computer. The shape of the letter depends on its position within the word. Persian is sometimes written with a form of Roman script for purposes of conveying the original text when the Perso-arabic script is not appropriate (e.g. in scholarly books, or second language literature); this is known as transliteration. While this is relatively standardised there are still differences in the symbols used to represent certain sounds, for example the Persian letter ش can be transliterated as *ŝ* or *sh* (both representing exactly the same sound) depending on which transliteration system the writer prefers. Tajik, which is commonly accepted as a variant of Persian, sometimes uses a form of the Cyrillic alphabet, which was instigated in the 1930s in Tajik Soviet Socialist Republic.

Windfuhr summarises the Persian phonological system as “29 segmental phonemes consist[ing] of four pairs of stops and four pairs of fricatives, two nasals, liquid and trill, three glides, and three pairs of vowels.” (Windfuhr, 2009, p. 448). This is a relatively symmetric system, the systematic pattern is mirrored in the stress pattern of the language which Windfuhr describes as “predictable” (Windfuhr, 2009, p.451). Unlike in Middle Persian there are no consonant clusters in Modern Persian at the beginning of words. The consonant clusters that had existed before were separated out with extra vowels, either at the start of

the word, or between the existing consonants. There are six syllable structures that occur in Persian:

- Vowel
- Vowel Consonant
- Vowel Consonant Consonant
- Consonant Vowel
- Consonant Vowel Consonant
- Consonant Vowel Consonant Consonant (Mahootian & Gebhardt, 2007)

Persian orthography is very phonetic, though short vowel sounds are not normally transcribed (Wilson and Wilson, 2001). Persian syntax has a canonical SOV, subject object verb, word order (Mahootian & Gebhardt, 2007). It is a pro-drop language, meaning that pronouns are not always explicitly referenced when they can be easily inferred from the context. The verbs agree with subjects with regards to person and number, verbs also contain information regarding tense and aspect. Persian is predominantly a head-initial language, with the exception of verbs, which tend to be phrase final. Morphologically, personal endings and pronouns are used to distinguish between the three singular and three plural classes, but beyond this Modern Persian has hardly any synthetic inflection: nominal or verbal. These elements can be seen in other Proto-Iranian languages, but within the different forms of Persian the decline of these elements can be traced back to late in the Old Persian era.

Persian texts tend to have less punctuation than English texts, with more clauses per sentence (Wilson & Wilson, 2001)(this can also be seen in Arabic). The lexis of modern Persian has been heavily influenced not only by Old and Middle Persian, but also by languages from other branches of the linguistic family tree, including; Arabic, French, English and Russian. This is particularly the case with technical vocabulary and is rooted (among other causes) in both the political history of Iran itself and the wider political history of the world. Though it is not technical vocabulary alone which sees an abundance of loan words, the Persian word for *thank you* in its transliterated form is *mersi*. A more precise account of the impact of other languages on Modern Persian (and where relevant Old and Middle Persian) will be given later in relation to language contact.

The next section, 2.2.3, considers the linguistic structure of Persian in greater depth, focusing on this might affect Persian-English Interlanguage. It will discuss Persian in relation

to the implications for the data examined in this research, and the features that might be discovered.

2.2.3 Persian-English Interlanguage

This section will consider the limited research surrounding Persian-English interlanguage and contact linguistics. It should be noted that this overview of existing research in this area, was used to compare to the findings from this research, as opposed to influencing the analytical process. As is discussed in Chapter 4, this research takes a 'bottom-up' perspective, focusing on what is in the data, rather than looking for what existing literature expects to find in the data. This is in part because most research on Persian-English Interlanguage focuses on specific areas or features, rather than providing an overview. The only key exception to this is a book chapter, based on comparative linguistics, by Wilson and Wilson (2001). This key text which summarises the features we would expect for Persian-English Interlanguage, provided by Wilson and Wilson (2001) in an edited volume aimed at teachers. It is clear from the title that this is firmly seated in a pedagogical perspective, and the aim is to identify problems - predominantly errors- which may be encountered by L1 Persian speakers learning English. Although it is not explicitly stated, we can assume that the aim of identifying these potential difficulties it to eliminate them in learners' production of L2 English. Despite the error/problem centred focus of this book, it gives a very good summary of what we might expect in Persian-English Interlanguage, and is therefore worth considering in more detail. We cannot accept the proposed interlanguage as definite, as there is no explicit reference to the proposed difficulties being discovered through empirical research. It is very likely that Wilson and Wilson based their proposed interlingual features on years of experience as teachers and linguists, however, for a forensic authorship analysis perspective we need a more empirical research driven corpus approach. The proposed interlingual features are nonetheless very interesting and useful to this research, as they can be compared to the features identified here, and may add an extra dimension to research. Table 2.2-1 below is a table developed predominantly from Wilson and Wilson's chapter (2001), and reviews the main areas that are likely to feature in the Persian-English language in the data[1]:

[1] Persian words or morphemes are transliterated into Roman script for ease. It should be borne in mind that there are varying conventions for transliteration, so differences may be seen in comparison to other texts.

Table 2.2-1 - Proposed Interlingual features

	Explanation, difference to English	**Possible Manifestation for this research**
Phonology	Persian has fewer vowels, diphthongs and consonants than English.	Confusion over similar sounds
Phonology	Consonant clusters do not exist in single syllables in Persian	Likely addition of vowels in clusters e.g. *perice* for *price* or *promptes* for *prompts*
Phonology	Initial consonant clusters do not appear at the beginning of words in Persian, borrowed words containing, often have vowels inserted	"Persian speakers learning a language such as English which allows consonant clusters tend to break up the clusters with epenthetic vowels." (Mahootian & Gebhardt, 2007, loc. 5292) E.g. *Faeranse* for *France*, or *Esport* for *sport*
Phonology	Persian spelling is very phonetic	When unsure, phonetic spelling may be relied upon
Orthography	Short vowels are normally represented by accents, but can be left out completely	Short vowels may be missed in spelling, or present confusion
Orthography	Persian does not have capital letters	Confusion about when to capitalise or not
Punctuation	Historically very little punctuation was used, past century has seen the introduction of a system similar to English, however with a more liberal approach.	There is unlikely to be over punctuation, and it may be used liberally or in a marked way
Punctuation	Most punctuation marks are similar to English, though the question mark is reversed, and the comma is inverted	Possible use of the Persian variant
Punctuation	"quotation marks are rarely used, and then not in a set way." (Mahootian & Gebhardt, 2007, loc. 4311)	Quotation marks may be used in a marked way
Punctuation	Paragraphing is a recent addition to Persian, indentation and separation mainly only appear in newspaper language.	Lack of paragraphing
Grammar	Word order differs from English. Adjective always follows the corresponding noun, verbs usually occur at the end of the sentence	Word order may mimic Persian e.g. *Yesterday girl beautiful (I) saw*
Grammar	Yes/no questions are signalled by the word *aya* rather than auxiliaries	Questions marked only by intonation/question mark Missing auxiliaries
Grammar	The pre-fix *na-* can be used to make past and present verbs negative	Not may be used in a similar way e.g. *she not eat supper*
Grammar	Auxiliaries not always used in their full form, but can be added to the noun as a suffix	E.g. *She(a) teacher* for *she is a teacher*
Grammar	Persian only has one main tag question *na* (meaning *no*)	There is unlikely to be a wide range of tag questions

	Explanation, difference to English	**Possible Manifestation for this research**
Grammar	Simple past and present perfect both represented as one tense in Persian	Confusion between simple past and present perfect tenses
Grammar	Past progressive tense in Persian is formed with the equivalent of the verb to have Present progressive can also be formed in the same way in some cases	The auxiliary *have* may be used in the place of *was* e.g. *he had eating*
Grammar	Reported speech uses the original verb tense	Tense change in reported speech. E.g. *he said I am feeling ill*
Grammar	Present tense is used for a greater variety of functions than in English: Present is used for present progressive If an action that is happening now, started in the past, it still takes the present. i.e*. I live in Birmingham for four years* It can be used for the future e.g. *he comes next week*	Use of the present tense where not expected in English, particularly in the place of the present progressive, present perfect, and in some cases the future tense.
Grammar	The majority of English modals are represented by one word in Persian *bayad* When the verb *can* is used, both the verb can and the associated main verb part are inflected	Confusion between modals (also possible use of *must* in past tenses rather than *had to*) E.g. *I could went*
Grammar	Gerund form does not exist in Persian, instead the infinitive is used	Potential confusion when distinction is required in English e.g. *Instead to fight, they danced*
Word order	Standard order is Subject, object, verb	This may lead to word order confusion
Word order	Adjectives usually follow the noun they relate to	e.g. *I cat grey saw*
Grammar	Passive forms are not used as widely in Persian. Passive is mainly formed by adding the verb *become* or omitting the third person plural pronoun	Less passives may be used May see *become* used in passive constructions
Articles	There is no equivalent for the definite or indefinite articles in Persian, instead suffixes are used to mark if a noun is definite or indefinite.	Article use is general is likely to cause confusion. They are likely to be omitted when normally required in English, and added where not necessary
Articles	The same word is frequently used as both an adjective and an adverb. When an adjective cannot be used as an adverb, then is must be turned into an adverbial phrase e.g. the adjective *dangerous* in Persian cannot be used as an adverb, instead it would be converted to *in a dangerous way*	Likely confusion between adjectives and adverbs Likely occurrence of awkward adverbial phrases

	Explanation, difference to English	**Possible Manifestation for this research**
Articles	Comparative ad superlatives are both always formed using a suffix The preposition used with comparatives is the equivalent of *from*	The two different methods of forming comparatives and suffixes is likely to cause confusion Likely to see an overuse of the preposition *from* in comparative phrases
Articles	There is no distinction between gender in the third person pronoun for *he*/*she* in Persian	Likely confusion between *he/she* pronouns
Articles	If a plural noun is used with a number, then the noun itself does not need to be written as plural e.g. *I saw two man*	Plural nouns may be written as single when used with a number
Prepositions	Preposition often follow the verbs, like in English, but the preposition used do not frequently correspond to the equivalents in English	Likely confusions include; *He climbed <u>from</u> the hill* (instead of *up the hill)* *She threw it out <u>from</u> the window* *I travelled there <u>with</u> car*
Phrasal verbs	Phrasal verbs do not exist in Persian	This can lead to comprehension issues and a difficulty in understanding that a particle can change the meaning of a verb. However for language production the more significant issue is difficulty when the particle is separated from the verb by the object, this may lead to the particle being omitted e.g. *I put my coat*.
Subordinate clauses	There is only one relative pronoun in Persian which is used for both animate and inanimate objects, subject and object and when a possessive is needed.	Knowing which relative pronoun to select may be difficult.
	Object pronouns in a relative clause are included in Persian (but omitted in English)	e.g. *The man, which I saw him...* *T book, which I gave it to you...*
Subordinate clauses	Prepositions always appear at the beginning of a relative clause	Constructions such as *who did you buy it for?* may be difficult. The more formal construction such as *For whom did you buy it?* may be favoured.
Conditional	Present and future conditionals are formed using a similar construction to English. The past conditional (type 3) contains the past tense in the main clause (unlike English).	e.g. *If I had finished my work, I was going to the party* (for *If I had finished my work, I would have gone to the*

	Explanation, difference to English	**Possible Manifestation for this research**
		party).
Concession	In Persian the equivalents of *although* and *but* can be used in the same sentence e.g. *Although he had no money, but he travelled to America*	This structure could be carried over
Conjunctions	Conjunctions are used considerably more frequently in Persian than English (particularly at the start of sentences) *And* in particular is used a lot, and is frequently used to string clauses together	There may be a higher occurrences of conjunctions (especially at the start of sentences) Likely to see lots of clauses joined together with *and*
Vocabulary	Few lexical similarities between English and Persian Some high frequency words show the shared indo-european history (e.g. *mother* and *brother*)	Should be limited interference
Vocabulary	Some technical terms are borrowed from English into Persian. Some have acquired distinct Persian pronunciations and more recently some terms have been replaced with Persian equivalents	There are very few false friends. There are some confusions over words whose meaning have changed after they have been borrowed e.g. *machin* = *car* not *machine*
Vocabulary	Persian uses descriptive compound nouns e.g. *ketab khane* lit. Book house = library	Compound nouns may sometimes be directly translated
Culture	Heavy emphasis is placed on formal literary language	The tendency is for student to avoid mastering colloquial spoken forms or writing in a simple style.
Culture	Iranian society places great importance on the art of conversation	

Wilson and Wilson's chapter (2001) contains a pertinent insight into potential interlingual features that will be experienced in L1 Persian speaker's English production, however, there is an absence of explicit reference to empirical research. It is to be assumed that their predictions are based on observations of L1 Persian learners of English, as well as a contrastive analysis of linguistic structures in Persian and English. Forensic linguistic research requires an understanding of how language is actually used, rather than how we would predict it is used. For this reason the potential features indicated are to be compared to the features identified in the data after the analysis, rather than informing the features that are being looked for within the data. This literature has mainly been focused on after the empirical research, as a comparison, rather than a template.

Chapter 3. Literature Review: Internet language, Authorship Analysis and Weblogistan

که در آفرینش ز یک گوهرند
چو عضوی به درد آورد روزگار
دگر عضوها را نماند قرار
تو کز محنت دیگران بی غمی
نشاید که نامت نهند آدمی

Of One Essence is the Human Race,
Thusly has Creation put the Base.
One Limb impacted is sufficient,
For all Others to feel the Mace.
The Unconcern'd with Others' Plight,
Are but Brutes with Human Face. [5]
Saadi – poem in front hall of United Nations building

This chapter discusses literature relating to key aspects and fields not discussed in the previous literature review chapter. Section 3.1 introduces the phenomenon of Weblogistan, and discusses the roles blogs play for Persian speakers. The second section, 3.2, considers the key literature relating to authorship analysis; the intended application of this research. The third section 3.3 discusses the related field of LADO; Language Analysis for the Determination of Origin, and the significance of this research for native language identification (NLID). The fourth section introduces literature surrounding using the internet and blogs as a linguistic resource; it also discusses the impact of computer mediated communication (CMC) research.

- Chapter 3 - Literature Review: Authorship Analysis, Internet language and Weblogistan
  - 3.1 Weblogistan
  - 3.2 Authorship Analysis
  - 3.3 LADO
  - 3.4 Blogs as a linguistics resource and CMC

3.1 Weblogistan

The data for this research comprises online blogs, the blogs by native Persian speakers are part of a concept called *weblogistan*. The term *weblogistan* is a combination of *weblog* and the suffix *–stan*. The root term of *blog* or *weblog* came into popular usage during the 1990's.

There are many different definitions of what a weblog entails exactly, but essentially the term *weblog* is a compound of two other words *web* and *log*. Some definitions state that for an online webpage to constitute a *blog* it must be a chronological list of entries, or a log, with the most recent occurring at the top. Definitions seldom limit it to one particular form, but say it may include a mixture of material, from writing, links, photos, video, to a mixture of all of these (Androutsopoulos, 2010; Baron, 2000). Peter Merholz is attributed with humorously splitting this into *we blog* (Economist, 2006), the word is frequently abbreviated to *blog* which has more recently become a verb, and a person who blogs is called a blogger. While weblog and blog may be relatively interchangeable it is interesting to notice that the derivatives appear to have only grown out of the shortened form *blog*. The suffix *-stan* means place in Persian, this can frequently be seen in names such as Pakistan or Kurdistan, normally referring to a place in the physical sense rather than the metaphorical or virtual sense we have here. It is difficult to ascertain the exact origins of the term *weblogistan*, but the phenomenon can be traced back to three Iranian students who started blogs in 2001 (Hendelman-baavur, 2007).

> ***"These blogs constitute what has been popularly named in Persian "Weblogistan," a distinct sort of public space wherein Iranians (particularly the young) assemble, express, and rearticulate themselves through cyber exchange and interaction." (Rahimi, 2008, p. 48)***

The colloquial term *Weblogistan,* refers to "the Iranian cyber-sphere of online self-publishing journals"(Hendelman-Baavur 2007). It should be noted here that there is no set definition of what constitutes a blog belonging to *weblogistan*; language is not a significant factor as there are many blogs in Persian, in English, or multiple languages. Location of the author would also seem to be irrelevant, as many famous Persian bloggers are either abroad for studies, or are in exile. One example is the blogger *Hossein Derakhshan*, also known as Hoder, who started blogging while studying in Canada. He has duel Iranian-Canadian citizenship, but has become famous as the Iranian blogfather' (BBC, 2010; Kiss, 2010)[2]. The key requirement would appear to be a concept of Persian or Iranian heritage, which is a feature of the blog itself. Due to the constantly changing nature of the internet, it is hard to say exactly where the Iranian blogosphere ranks in size, compared to other communities.

[2] Hossein Derakhshan (Hoder) is currently in prison in Iran, accused with "propagating against the regime" (BBC, 2010) and for travelling to Israel.

However, most estimates place it well inside the top ten (Kelly & Etling, 2008). There are also a disproportionately high number of blogs in the Persian language online; disproportionate if one compares it to the number of native Persian speakers. Mina (2007) notes that public access to the internet in Iran did not occur at the same time as other countries, the high occurrence of Persian blogs is even more astonishing. This would seem to indicate that blogging plays a particularly important role within the Persian community.

> ***"On the one hand weblogs have gained recognition as an important forum for debate and a valuable source of information. Yet on the other hand, they have become targets of government efforts to limit freedom of expression." (Hendelman-baavur, 2007, p. 8)***

In Iran (the country most associated with the Persian language), as in many other countries, the link between politics and the media is a complex one. The media worldwide has played a role in many revolutions, protests, elections, and governmental decisions (Friedland, 1992; Sreberny-Mohammadi, 1990; Mina, 2007). President Khatami's second term in office heralded an extensive crackdown on the media, despite having promised greater press freedom before his election. (Hendelman-Baavur, 2007; and Khiabany and Sreberny, 2001). This crack down on the wider media is viewed as being a major reason for the growth of Iranian bloggers; many journalists found themselves increasingly out of work, as the newspapers they worked for were closed. Many of these out of work journalists started blogs. Mina illustrates this with the example of Hanouz (p7 Mina, 2007), a group blog that was founded by three young journalists who were steered towards blogging after they lost their jobs due to the media restrictions following President Ahmadinejad's election in June 2005.

> ***Free and anonymous expression [on the internet and particularly weblogs] mediated by computers and practiced in the privacy of one's own home (or an isolated computer station) has also enabled the dismantling of social and physical restraints. The unedited, and informal nature of weblogs has turned them into a source of empowerment for Iranian youth and especially for Iranian women. (Hendelmann-Baavur, 2007)***

Reporters Without Borders (Reporters Sans Frontiers), an international organisation dedicated to campaigning for the freedom of press, refers to Iran as "one of cyber-

censorship's record-holding countries" (Reporters Sans Frontiers, 2010, p. 2), listing it as one of the worst offending "internet enemies". The technology the Iranian government employs for filtering is widely reported as not being particularly advanced[3], and can be easily bypassed through the use of proxies[4] (Kelly & Etling, 2008; Mina, 2007). One of the major deterrents against people openly criticising Islam, the Iranian government, or regime is the prosecution of numerous bloggers, most frequently on charges against Islam or against Iranian national security. Reporters Sans Frontiers reports that were 18 Netizens, or internet citizens, imprisoned in Iran in 2012. Arguably the most renowned of these is Hossein Derakhshan, also known as Hoder. Hoder is recognised as one of the instigators of the Persian blogging sensation that was to become known as weblogistan (Hendelmann-Baavur, 2007). The title of Hoder's blog in part explains the appeal of the internet; "Editor: myself". In a country which has experienced increasing state control of the media; the internet offers a rare opportunity to express oneself without being edited, or being dependant on state approval (to avoid being closed down).

Social media such as the internet have played a strong role in disseminating information internationally, even when governments have sought to prevent the spread of information. One example of this is the video of Neda Agha Soltan, who was shot in June 2009 during the protests in Tehran surrounding the presidential election. The video was allegedly filmed on a bystander's mobile phone, then later posted on YouTube, where it received thousands of views and was also picked up on by many news stations worldwide. While the events in the video cannot be checked for authenticity, due to the surrounding secrecy, the video reached out to many people, with Neda being hailed as a martyr and the face of the revolution (Reporters Sans Frontiers, 2010), and many protesters started carrying placards with her picture. However, social media, and the greater freedom for expression does not guarantee validity. An example of this from the Arab spring is the blog; *A Gay Girl in Damascus*. This blog was active between February and June 2011, which included the beginning of the Syrian uprising during the Arab Spring. The blog was purportedly by a young gay female called Amina Arraf, who was an American-Syrian living in Damascus. In June 2011 a post claiming to be written by Amina's cousin said that she had been taken into custody by the government.

---

[3] It has been reported that the Government uses a pirated version of a commercial programme desgined for parents to set-up parental restrictions, limiting and protecting their children's interent use (Mina, 2007)

[4] Reporters Sans Frontiers even has an advice booklet online with advice for using proxies to circumnavigate government restrictions.

It was later discovered that the real author was a 40 year old, male American student studying at Edinburgh University (Addley, 2011; BBC News, 2011). The problem of ascertaining identities of blog authors for this research will be discussed further in Chapter 4.

The significance of weblogistan is perhaps most noticeable through the content of the blogs, which frequently focus on a sense of Iranian or Persian identity. It is important to understand and be aware of the surrounding context of L1 Persian blogs, as the impact may spread beyond the content and affect later analysis.

3.2 Forensic Authorship Analysis

Forensic authorship analysis is a branch of forensic linguistics, it has two main sub-branches; comparative authorship analysis (also known as authorship attribution), and authorship profiling (also referred to as sociolinguistic profiling). Comparative authorship analysis involves multiple texts which are compared for either common or distinct authorship. The intended application of this research is firmly seated within the field of sociolinguistic profiling, as the intended application of this research is to determine information about an anonymous author. Authorship analysis (as described in Section 1.1) is a wide area attracting interest from a wide range of academic fields, most notably including; forensic computer science, psychology, and literature studies. Grant (2008) separated the aim of forensic authorship analysis into answering questions in four main areas:

1. How was the text produced?
2. How many people wrote the text?
3. What kind of person wrote the text?
4. What is the relationship of a text with comparison texts?

This research develops a method for indicating the native language of an anonymous author, which fits firmly into the third question of what kind of person wrote the text. In this research the analytical method devised seeks to determine whether a particular author belongs to the set social group of native Persian speakers. Although the main intended application for this research is authorship profiling, it may also be relevant to some authorship attribution cases, as the two are not completely distinct from one another. A linguist, who was initially asked to perform a profile, may later be consulted to perform authorship analysis when a suspect has been identified. The features that are identified in profiling may then feed into the authorship analysis. This interplay is discussed further in

Section 9.3, along with a Canadian case that is more akin to an authorship attribution, than authorship profiling case (but contains elements of both).

Linguists sometimes consult as investigative linguists and provision of evidence linguists; providing either evidence in court, or expert advice to aid investigations. Authorship attribution is used as both an evidentiary and investigative tool, whereas authorship profiling is predominantly of use in an investigative context. This is discussed more fully in Chapter 1 Section 3. Authorship profiling evolved directly from sociolinguistics, a fundamental precept of which is that a person's linguistic output is influenced by a number of social factors. These typically include gender, linguistic and geographical background, age, and educational status. Authorship analysis and sociolinguistic profiling aims to determine information about the author(s) of the text, this does not include psychological observations, or inferences about intent (including threat assessment), as these are beyond the remit of a linguist.

A widely referenced case (Leonard, 2005; Perkins & Grant, 2013; Solan & Tiersma, P, 2005) that provides a good example of the authorship profiling process is one in which linguist Roger Shuy was consulted by American police to provide help narrowing the suspect field when the following ransom note was found at the home of a kidnapped child;

> ***Do you ever want to see your precious little girl again? Put $10,000 cash in a diaper bag. Put it in the green trash kan on the devil strip at corner 18th and Carlson.***
>
> ***Don't bring anybody along.***
>
> ***No kops!! Come alone! I'll be watching you all the time. Anyone with you, deal is off and dautter is dead!!!***

Shuy's analysis suggested that the author was likely an educated man; this was inferred due to the misspellings and their patterning. The author misspelt simple words such as dautter (daughter), kops (cops) and kan (can), yet produced the correct spelling for more complex words, most notably; precious. The grammar is fully coherent and well structured, indicating a high level of education which the author was attempting to disguise through deliberate misspellings. Shuy also suggested that the author might originate from, or have strong links with the small town of Akron, Ohio. This geographical indication was derived from the use of

the term *devil strip*, which refers to the strip of grass that separates the pavement from the road. This is a unique term, with usage restricted predominantly to Akron, Ohio, indicating that the author had strong links to this town. The profile as suggested by Shuy corresponded to one of the people the police were considering as suspects; the man was arrested and later confessed to kidnapping the child and authoring the ransom note. Shuy's profile was not intended to be submitted as evidence in court, but demonstrates how profiling can greatly aid investigations.

As is demonstrated by the ransom note above, the texts that occur within the forensic context are often very short. Coulthard (1994) estimates that the majority of forensic texts are between 400 and 700 words long. The brevity of texts is even more prominent with the increase in computer mediated texts, such as text messages or tweets, occurring in forensic contexts (Silva & Laboreiro, 2011). There is a great degree of variability in the texts that occur in forensic authorship analysis situations, from the Unabomber Manifesto which was over 100 pages long, to text messages in cases such as Malcolm Coulthard's Jenny Nicholl no-body murder case. More recently cases are appearing revolving around tweets that are only 140 characters long. It is not only the length that varies considerably in these cases, but also the genres. This variability means that it is very difficult (and potentially damaging) to recommend one standard analytical technique for authorship analysis (Grant, 2008). However we can research and develop tools that linguists can apply in authorship analysis situations.

3.3 LADO

Language analysis for the determination of origin, frequently shortened to LADO, is a form of applied linguistics "used by governments to assess asylum seekers applying for refugee status" (Fraser, 2012, p. 9). Given the nature of being a refugee, very few asylum seekers are able to supply official documentation to support their application, meaning "experiential narration is basically the only tool for explaining and supporting the application" (Maryns, 2004, p. 243). Language analysis is one way for examining the experiential narration for veracity. The procedure differs from country to country, though it is now common that the host government will utilise a commercial company which specialises in language analysis (Patrick, 2010) such as the Swedish company Sprakab, or the Dutch company De Taalstudio (founded by linguist Maaike Verrips). An interview is then conducted with the refugee

applicant, whose language is analysed to help determine either regional or ethnic origin. There has been great variation in how LADO is performed, particularly the analysis. There has been a significant effort from linguists to regulate procedure, most notably a report published by Australian linguists Eades, Fraser, Siegel, McNamara, & Baker (2003), which highlighted several key failings in the way LADO was being performed in Australia. Although this and the Guidelines for the Use of Language Analysis in Relation to Questions of National Origin in Refugee Cases (Language and National Origin Group, 2004) have had an impact on how LADO is conducted in many cases, it is still standard that an asylum seeker undergoing LADO performs an interview, which is then analysed to determine information about the speaker's origin. As is discussed later in this section, the focus of the 2003 report was on who performed the analysis and the method for analysis. It is easy to draw parallels between LADO and the research being discussed in this thesis, as the essential aim of both is to determine information about a particular person from their language. They both work on the principle that a person's L2 will contain features that can be used to determine information about a person's linguistic origin or L1. There are however some key differences:

- LADO relies on spoken interviews, rather than written text.
- LADO relies on elicited data, as opposed to collected data.
- LADO reports play a key role in deciding whether an asylum seeker is granted refuge, and as such is taken as evidence in decision making or at a hearing or tribunal. In contrast, NLID is intended to help guide investigations, not act as evidence.

Two further differences that were highlighted by the 2003 report that LADO is not always performed by qualified linguists, and the methodology replied is not always consistent or based in theory. It should be stressed that these issues have been greatly diminished and are not present in the work of individual consultants or companies that are striving to improve standards in LADO. This research is more aligned with the linguistically grounded approach to LADO that uses trained linguists and consistent methods, yet is still a distinct field. LADO and NLID both have predominantly forensic applications, and relate to sociolinguistics and more specifically language contact. However, the different contexts, different data collection methods and data production methods require different analytical methods and approaches. Despite being separate fields, they should not be completely isolated from one another, as research in one field may provoke advancements in another.

Eades et al. (2003) summarised that language analysis or LADO as it was being used in Australia before their report was neither valid nor reliable, and was "based on "folk views" about the relationship between language and nationality and ethnicity, rather than sound linguistic principles." (Eades et al., 2003, p179). They listed the key problems as follows:

> ***i) a person's nationality cannot always be determined by the language he or she speaks, ii) a few key words and their pronunciation normally cannot reveal a person's nationality or ethnicity, iii) common perceptions about pronunciation differences among groups of people cannot be relied upon, iv) any analysis of pronunciation must be based on thorough knowledge of the language and region in question and must involve detailed phonetic analysis. (Eades et al., 2003, p. 179)***

All these problems should also be addressed when building a reliable NLID methodology, which like the newer approach to LADO, seeks to ground itself within linguistic theory. The first point; that nationality and language are not linked, is particularly valid with reference to Persian. As discussed in Section 2.2, the languages of the Iranian Plateau do not mirror political boundaries. Persian is spoken in Iran, Afghanistan, and Tajikistan (as well as much further afield), and there are many other languages spoken in the same regions as Persian. The second observation of Eades et al. (2003) is that key words and pronunciation are not particularly significant of a person's ethnic or national origin. The research reported on in this thesis does not focus on spoken language, and the aim is to determine an author's L1 rather than their origins, however, this research still avoids placing too much significance on key words or common perceptions (which the report also demonstrates are unreliable), instead focusing on underlying linguistic structure. The concept of common perceptions is also expanded further in Study Three (as discussed in Chapter 8), which looks at disguise data, and asks participants to pretend to be native Persian speakers, in order to understand what the common perceptions about L1 Persian speakers writing in English are. The fourth point of the report relates predominantly to spoken data, which this research does not employ, however it also relates to the analysts knowledge of the region and language. The importance of the analyst's knowledge of Persian is discussed further in Chapter 4 Section 2.

There is considerable research surrounding LADO, language analysis, and the relating ethical issues (McNamara & Shohamy, 2008; Patrick, 2010; Patrick, 2010). The ethical issues that have arisen from debates surrounding LADO and are relevant to NLID are discussed in the next Chapter, in Section 4.4.

3.4 Blogs as a linguistic resource

Hunston writes that corpora and corpus linguistics "have revolutionised the study of language, and of the applications of language over the last few decades." (Hunston, 2002, p. 1). There can be little doubt that corpus studies have enabled many new insights into language, and that they have helped linguists to correct many previously held conceptions. As Stubbs writes:

> ***Studies of large corpora provide two main contributions to linguistics. First, they provide many new and surprising facts about language use. [...]. Second, by looking at language from a new point of view, corpus studies can help to solve paradoxes which have plagued linguistics for at least a hundred years. (Stubbs 2007)***

Stubbs (1996) notes that corpus linguistics is a complete contrast to the previously popular Chomskyan linguistics. Chaski (2001) concludes that forensic stylistic approaches that rely on prescriptive grammar are unreliable on many accounts:

> ***These techniques do not quantify linguistic patterns; they are not amenable to statistical testing nor the calculation of error rates. Further, these prescriptive techniques rest upon factually incorrect ideas about individuality in language performance and violate theoretical principles of modern linguistics. Finally, these techniques fail to differentiate between documents authored by different writers and/or fail to cluster documents authored by the same writer with a high level of accuracy. (Chaski 2001)***

As Stubbs (2007) states above, the corpus approach has many positive benefits. Rather than creating linguistic hypotheses then looking for data that support these, a corpus based approach can encourage an unbiased analysis of the way language is actually being used. The internet provides considerable advantages for corpus building (significantly for this project) including speed of collection and the ability to collect large volumes of data (Hundt, Nesselhauf, & Biewer, 2007). It is constantly being updated, and contains language as it is being used in a natural environment (Leech, 2007). There is however, a difficulty with constructed identities, there is a perceived anonymity online and it is difficult as a researcher

to verify a person's linguistic history (Crystal, 2001) (this is discussed more thoroughly in Chapter 4 in relation to the methodology for this research).

There are limitations to a corpus approach. As Biber et al. write; "It is important to realize up front is that representing a language – or even part of a language – is a problematic task." (Biber, Conrad & Reppen, 1998, p.246). The predominant aim is to understand how L1 Persian speakers write in English, this is done through a representation of the wider social group in the form of a corpus of blogs. Biber et al. (1998) also refer to the constraints and compromises that are an integral part of corpus building, they recognise the importance of being realistic, stating that "Every corpus will have limitations, but a well-designed corpus will still be useful for investigating a variety of linguistic issues." (Biber, Conrad, & Reppen, 1998, p. 246). Chapter 4 Section 3 discusses the practicalities of representing L1 Persian speakers, while dealing with the inevitable limitations.

Coulthard (1994) discussed the importance of corpora for forensic linguistics, stating that the future of forensic linguistics and "any improved methodology must depend, to a large extent, on the setting up and analysing of corpora." (Coulthard, 1994, p. 40). The aim of this research is not to highlight mistakes that are made by L1 Persian speakers writing in English. Instead it is to focus on how English is normally used by Persian speakers as opposed to native English speakers. The key word here is 'normally'. The best way of understanding how a written language is normally used by a collection of people, is to build a corpus of texts that represents the writing of the people being studied. Here the main issue is what constitutes a representative collection of texts. This study focuses on the forensic application, therefore texts that are likely to arise in a forensic context would be the most useful. The majority of existing corpora of L1 Persian speakers writing in English are comprised of elicited student texts. Olsson (2004) stated that any text had the potential to become a forensic text depending on the context, however none of the relevant cases discussed so far relate to student data. The growth of internet and computer mediated communication has lead to an increase of computer-mediated crime, and hence there are now established and continually growing fields such as *computer forensics* and *internet forensics* (Berghel, 2003). It is therefore appropriate that this research uses internet language as a data source, as it will not only provide language that is natural and not elicited, but also of a genre that is relevant to forensic investigation.

Text from the internet has the benefit that the medium of the internet elicits a more relaxed style, congruent with the 'real-life' data necessary for this research. David Crystal has

discussed the much publicised informality of internet language in many works, he illustrates the situation by writing “The electronic medium [...] presents us with a channel which facilitates and constrains our ability to communicate in ways that are fundamentally different from those found in other semiotic situations.” (Crystal, 2006, p. 5) This builds on Labov’s work on field methods for sociolinguistic research, in which he ascertained that style is related to the amount of attention a speaker pays to their speech and in turn that the less attention is paid to speech, the more systematic data it provides for linguistic analysis (Labov, 1984, p. 29). As this research is intended to create a model for analysis in an investigative or authorship analysis scenario, the use of internet language means the research is built on the type of language that is more likely to require analysis in a real-life situation.

Part Two – Features of L1 Persian speakers blogging in English

Part Two details the main study of this research. The first chapter, Chapter 4, sets out the methodology for the wider research project. Chapter 5 focuses on Study One, the main study focusing on L1 Persian and L1 English blog authors, and describes the development and implementation of the NLID coding system. The third chapter of this part, Chapter 6, also focuses on Study One and presents the statistical findings.

Chapter 4. Overall Methodology

Flowers which have blossomed in the garden today; If you pluck them tomorrow they will be of no use. – Ferdowsi

This chapter sets out the overall structure for this research. Firstly it discusses the aims of this project as a whole (in Section 4.1), next it considers the effect of relevant existing methodological theories and the implications these have for the methodologies of this research (Section 4.2). The first two sections have then laid the foundations for the third section, Section 4.3, which discusses the practical aspects and how they manifest in the structure and plan for this research project. More specific details for each study are discussed in the relevant chapters but section 4.3; considers how the three sub-studies fit within the overall methodology for this project as a whole. Finally, in Section 4.4, issues relating to ethical considerations for each area of this research are discussed, along with the measures that have been put in place to ensure that this research is as ethically considerate as possible.

- Chapter 4 – Methodology
  - 4.1 Aims of research
  - 4.2 Methodological theories and fields
  - 4.3 Structure and plan of overall research
  - 4.4 Ethical considerations

4.1 Aims of research.

The main aim of this research is to investigate how interlingual features present in an author's non-native written English can aid forensic authorship analysis. More specifically, it

investigates how features present in an author's L2 English production may indicate information about the author's native language or L1. This is termed NLID or Native Language Identification. In particular, this project focuses on native Persian speakers writing in English. Therefore we are asking if features in an author's L2 (non-native[5]) written English indicate that the author in question might have a native-like influence from Persian. The next question is whether these features distinguish between an L1 influence from Persian, and an L1 influence from other languages that are linguistically close, with similar grammar and structures. It could be the situation that the features that are identified as being indicators of authorship by an L1 Persian speaker could be general to the language family. An analyst would therefore struggle to distinguish between similar languages, or languages from the same branch of the linguistic tree. It could be that features that are significant for distinguishing between authorship by an L1 English speaker and an L1 Persian speaker are not the same features that distinguish between linguistically similar languages. Therefore, a subsequent aim is to see how the features distinguish between linguistically close languages, and if different features are significant for answering this question.

There are numerous documented cases in forensic linguistics in which authors have tried to disguise their language (examples are discussed in Chapter 4 section 1); this may be expected due the forensic context. The prevalence of attempted disguise raises the question of whether the features which are identified as indicating an L1 Persian influence would indicate the same thing if an author was merely pretending to be an L1 Persian speaker. The question then, is whether we can we distinguish between someone pretending to be an L1 Persian speaker and an actual L1 speaker, or not – whether a lay-author disguising their language would incorporate the key distinguishing features this research identifies.

When considering questions like this from a forensic linguistic perspective, we need to understand the degree of accuracy. It is generally considered that profiling is not sufficiently accurate to be relied upon as evidence in a court situation (Grant, 2008). However the benefits of profiling in an investigative context are exceedingly well documented (Grant, 2008; Kniffka, 1996; Shuy, 2001). As forensic scientists, it is important to understand how much confidence we can place in findings from certain analyses, as well as striving towards more accurate methods. It is also necessary to understand the extent to which results are replicable. In order to do both of these things, we need to understand the accuracy of any

[5] In this thesis the term L2 is used to represent any language which is not an L1 or native language, it does not mean that it is the second language as opposed to a third or fourth language.

methodological system that has been developed. In this study the accuracy will be tested at various stages through statistical analyses. This empirical marking of the error rates and accuracy is exceedingly useful, but there is also need for a practical and theoretical discussion of issues affecting the accuracy of the findings.

While considering the need for knowing the degree of accuracy so that it can be improved upon in future research, it is also necessary to make it clear that this research is not intended to posit a complete method for Native Language Identification. Rather this research is intended to be a starting point from which more research into NLID can build upon (or contradict and dispute). It could therefore be said that the main underpinning aim of this research is to stoke an academic discussion into Native Language Identification in forensic authorship analysis. It should also be noted that these aims are not all encompassing. With more time and more resources there are many other related areas that it would be of benefit to research. It was necessary to recognise the constraints on this research, and focus on several key questions as have been detailed in this section. Calls for further research are detailed at relevant points throughout this thesis, but most specifically in Part 3, Chapter 9 Section 4. The structure of this thesis approximately mirrors these aims, with the final two studies being designed specifically to correspond with one key aim each. In order to get a better overview, we can summarise the aims as follows:

1. To determine if interlingual features in L2 writing can be used to indicate an author's native tongue (Study One, and throughout other studies)
2. To develop a methodology of NLID (Native Language Identification) (Study One)
3. To determine what features indicate authorship by a native Persian speaker (Study One)
4. To determine if we can attribute influence to being from a specific language, rather than a language family and to determine if we can distinguish between two languages from the similar geographical area (Study Two).
5. To determine if it is possible to distinguish between a genuine native Persian speaker writing in English and someone is trying to disguise their language, to give the false impression that they have an L1 influence from Persian (Study Three)
6. To understand with what degree of accuracy we can draw conclusions based on the analysis involved (throughout all studies)

4.2 Methodological fields and theories.

So far this thesis has been discussing the concept of an L1, native language, or mother tongue, as a straight forward issue; however, it is not as simple as it might initially appear. Due to a variety of factors (such as globalisation), very few people have a simple linguistic background. Multilingualism is increasingly prevalent, as is the influence from other languages, which we might not necessarily speak, through the media. If we consider the history of Iran and the Middle East in general, we can see that the language maps do not match the political borders, far from it in fact (for a more detailed discussion of Persian and languages in the Middle East, see Chapter 2). This has methodological implications as well as theoretical ramifications for selecting blogs for the L1 Persian corpus. Rampton (1990) discusses the implications of the term native-speaker, and how each implication is now widely contested. In a later paper, Leung, Harris and Rampton (1997) discuss the term *language affiliation* which "refers to the attachment or identification they [any person] feel for a language whether or not they nominally belong to the social group customarily associated with it." (Leung et al., 1997, p. 555). Rampton (1997) recommends speaking of language experts or having language expertise, rather than using terms such as native speaker or mother-tongue. In an educational context, this makes sense. In the context of this present research however, we are considering how bloggers describe their own linguistic history, rather than applying labels as researchers. It is only reasonable to expect different people to describe their linguistic histories in different ways. For this research an author's L1 or native language, will be considered to be the language they themselves identify as being their native language, or mother-tongue. Of course the exact term an author may use can vary considerably, the terms used by participants in this research most often were 'native language' or 'mother tongue', but other variants were also accepted. If an author states more than one mother-tongue or native language, then they will not be considered as a native speaker of that particular language, however, if they profess an ability in other languages, but do not give these the status of 'native language' or 'mother tongue' (or any combination of these terms), this will not be considered as contradicting the information about L1. Ultimately there is an element of subjectivity when determining which authors are suitable to be included during data collection. While this cannot be excluded completely, irregularities are aimed to be diminished through applying consistent criteria through data gathering and also when information pertaining to linguistic background is encountered during closer analysis of the texts. One difficulty that was only experienced once data collection had begun, is that a native English speaker, blogging in their native tongue, is very unlikely to have cause to discuss what their L1 is, unless there is a heavy

influence from another language (for example they live abroad, or other languages have a significant role in their life). The same criteria as the L1 Persian corpus were applied, so the participants were required to state that English was their L1 (or native/mother-tongue), and not obviously contradict themselves. Persistent searching for data meant that eventually the required amount of data could be collected.

When researching cross-linguistic influence, it is important to consider the linguistic history of the analyst(s).

> ***"A realistic methodology would not insist that he [the linguist] abandon all description until he can trade insults with the man in the street or dispute local theology in the full flood of an enlightened scepticism. But if the linguist recognizes the existence of these higher levels of competence, he can use his developing grasp of the language on second or third trips to the field to locate the differences between norms and behaviour; by doing so, he would deepen the value of his original observations for an increasingly dynamic and secular linguistic theory." – p103 (Labov 1984)***

Although I am not fluent, my growing knowledge of Persian is of great importance to this study. One could even say that not being a native speaker enables a greater understanding of the differences between English and Persian, as native speakers are often not consciously aware of certain structures or anomalies in the language that they use intuitively. (Labov, 1984 p. 103). In this research I am the sole analyst. I am a native English speaker who has learnt varying languages to varying levels, most notably German, in which I completed an undergraduate degree, as well as gaining relative fluency when I lived and worked in Germany. Due to the practical intelligence implications of this project, it was suggested that a language which may be of more political interest, may have greater impact, as well as generating more interest and support. As I have a passion for learning languages, I undertook Persian evening classes based at Westminster University. In order to gain as much competency in Persian as possible, as quickly as possible, I did two courses a year for the first two years (Grades 1 and 2, followed by Grades 3 and 4). The intention from the start was to 'front-load' the language learning, enabling me to gain a higher level of understanding, before focusing more intensely on the analytical sections of this research. There are both benefits and drawbacks to being a non-native Persian speaker, however through gaining a competency and understanding in Persian, and consulting with native

Persian speakers, it should be possible to make maximum use of the benefits of my time spent learning, and enable the greatest understanding of the texts and their features.

Another issue is the perceived anonymity of the internet. This means that people are not always forthcoming about their real identities; there is nothing to prevent people from creating new profiles, and lying completely about their linguistic backgrounds (Baron, 2000). This is even more possible when one considers the political concerns prevalent around weblogistan as discussed in Chapter 3 Section 1. There is also the issue of whether we consider the influence of other languages the authors may speak (for example in my situation, the influence of German). Both of these potential issues are intended to be diminished through choosing a corpus approach, however, they should be borne in mind during the analysis. It is unlikely that a significant number of the participant authors will have decided to create a consistent and fake linguistic background, or that the data would focus on a large number of authors with a linguistic influence from an unrelated language. While strenuous efforts will be made to select the 'best' blogs for inclusion in the data (see section 4.3 below), it is also believable that some erroneous ones will slip through, but the effect of these will be diminished due to the volume of accurate ones. Siemund wrote that "We know that languages can influence one another in a situation of contact, but predicting the outcome of a language contact situation remains an immensely challenging task." (Siemund, 2008, p. 3). The pedagogical perspective of interlanguage is predominantly focused on predicting errors in order to eliminate them, conversely this research is based in forensic linguistics and is focused on documenting features that actually occur through unbiased observation, rather than prediction.

4.3 Structure and plan of overall research

The structure of this research is designed to answer the aims as set out in section 4.1. The overall research is divided into three separate studies. Study One, the main section of analysis, focuses on two corpora of blogs; one by native Persian speakers writing in English, and one by native English speakers. Focusing on internet blogs has many benefits for research in forensic linguistics, the most significant for this research being that it is collected data, as opposed to elicited. Conversely most existing research investigating cross-linguistic influence looks at student data, which is elicited by teachers, and is also written with the purpose of being critically read, whereas forensic texts have a predominantly communicative purpose. Using internet blogs as a data source means that the data more closely matches the forensic texts which may later benefit from the application of NLID analysis.

In order to better identify and understand the features present in the writing of L1 Persian speakers writing blogs in English, it is important to have a control corpus that is as similar as possible to the L1 Persian corpus. A control corpus also prevents features of the genre being falsely identified as features of L1 Persian speakers. Therefore the data for Study One comprises two corpora, the study corpus of L1 Persian blog authors writing in English, and the control corpus of L1 English blog authors writing in English. Authors for both were collected using the same methods. The aim was to collect approximately 2,000 words of text per author, only complete blogs were included, and as many blogs as necessary were collected to form an author's contribution. It was impossible to collect such a large volume of text for some authors, as they had not posted enough blogs. Blogs were selected for inclusion based on information provided by authors within their blog or site about their linguistic background (see section 4.2 above for more specific information regarding criteria for determining if an author constituted an L1 speaker of a certain language). All blogs were identified through Google searches, using keywords to retrieve results which included blogs where authors identified themselves as being an L1 speaker of the relevant language. A typical search entry was: Blog "I am a native Persian speaker". The collection method meant that all data self-identified as being a blog; either through information contained within the text, or due to the web-address. Effort was also made during collection to ensure that the data conformed to the wider definitions of blogs, as discussed in Section 3.1, as well as ensuring that it met the criteria for this research data (such as being by a single author and without ostensibly being edited).

During the analysis of the internet blog data it became apparent that, as expected, there had been some limited discrepancies during data collection. Some blogs were duplicated and some appeared to be incomplete. A close reading of some of the blogs showed inconsistencies in the claims of some of the author's L1's. This was a possibility that had been planned for, therefore blogs with major incongruity as to the stated L1 were discounted from the study and replaced with another one that had been collected. The initial plan had been to focus on 30 blogs in each group, however during the coding stage, it became apparent that the amount of time that it would take to code all 30 outweighed the benefit of having 30 blogs, therefore it was decided that 25 would be a more appropriate number. The evolution of the coding is discussed in Chapter 5. An outline for the distribution of the data to be collected can be seen in Table 4-1 later in this section, Appendix B shows the precise information for the data as it was collected per author and for each corpus.

Study Two seeks to answer the aim of determining whether it is possible to distinguish an influence due to L1 Persian from influence(s) due to different, yet similar language(s). In order to test this, two different languages were chosen; Azeri (also called Azerbaijani) and Pashto. The motivations for choosing these languages, and the implications are discussed more thoroughly in Chapter 7, specifically Section 7.1. The methodology for data collection remained consistent with the data collection for Study One, as outlined above, the only significant different was the volume of data that was collected. Due to time constraints it was decided that there would be fewer authors in each corpora, and a smaller volume of text for each author. This was also appropriate, as it was more difficult to find authors that identified themselves as native speakers of these two languages, in part because they are much smaller languages than Persian and English. It was decided that a total of 10 authors (five for each language), contributing approximately 1,000 words each, would yield enough feature rich data for analysis. The smaller text size also allows us to see how NLID analysis copes with a reduced volume of text, this is important as forensic texts are normally only a few hundred words in length (Coulthard, 1994). The analytical methodologies for Study Two will be explored in greater detail within Chapter 7.

Study Three (discussed in Chapter 8) examines NLID in relation to authors' attempts to disguise their language (as is frequently seen in documented forensic authorship analysis casework), specifically, if an author is attempting to disguise their language by falsely indicating an L1 Persian influence. Due to the element of deception required within the text, this data had to be elicited rather than collected. Participants were asked to write a section of text, in the style of a blog, pretending to be a native Persian speaker. It did not matter what their native language was, or what other languages they spoke, so long as they did not consider themselves to be native Persian speakers. In practice it quickly became apparent that participants would need some knowledge of the Persian language, or at least the acquaintance of L1 Persian speakers, in order to undertake disguising their language as L1 speakers. Participants were asked to answer a series of questions in order to understand their linguistic histories, and how much exposure they had had to the Persian language and L1 Persian speakers. Due to the difficulty of the task, and in order to better understand participants' decisions, it was decided that there would be benefit from questions relating to the linguistic decisions they would make when disguising their language to appear as L1 Persian speakers writing in English. In practice, participants were much more willing to answer the questionnaire than undertake the writing task. Potential participants were asked to complete information in four main areas: ethical consent, linguistic background, linguistic

decisions and the writing task. All areas were distributed through an online questionnaire[6]. The intention was to create a corpus of disguise data that would be comparable to the previous corpora discussed; however, in practice it was difficult to collect the volume of data that would have been required. Nevertheless a small corpus of disguise data, together with the accompanying questionnaires, provides invaluable data and is discussed more thoroughly in Chapter 8. The questionnaire is discussed more thoroughly in Chapter 8 (Section 2) and can be seen in Appendix O. The intended distribution of data collected is shown in Table 4-1 below.

Table 4-1 Data collection overview

Corpus	**L1 language**	**Number of authors**	**Words per author (approximate aim)**	**Studies**
L1 English Corpus	English	25	2,000	1
L1 Persian corpus	Persian	25	2,000	1,2 &3
Other languages corpus				
	Azeri	5	1,000	2
	Pashto	5	1,000	2
Total:	Azeri & Pashto	10	1,000	2
Disguise corpus	English (or other)	9	As much as possible, but at least over 100	3

For ease, a research assistant was used to help with the data collection. This was particularly useful as finding blog authors who consistently identified themselves as L1 Persian or English, involved trawling through a lot of internet searches, and was considerably time consuming. The aims, and the search criteria were fully explained to the research assistant, and as much assistance as possible was given throughout the process to ensure consistency. A record was also kept of who collected the data for each author. Blogs for the first two studies were collected between January 2010 and July 2011, the data for Study Three were collected between May 2011 and May 2012. Although the data for each author normally comprised of multiple blogs, once the data were collected, the blogs for each author were considered as a single unit. This enables ease of discussion and analysis.

[6] The website used was surveymonkey. www.surveymonkey.com

4.4 Ethical Considerations

This section discusses the ethical considerations surrounding this research. It starts by considering how to collect and handle the data for this research in an ethical manner, as well as protecting participants, the researcher and any other related parties. Next this section deals with the wider ethical implications of Native Language Identification (NLID) research. The Linguistic Society of America breaks down the ethical responsibility into five main areas, responsibility to individual research participants, to communities, to students and colleagues, to scholarship and finally to the public (Linguistic Society of America, 2009). This research seeks to honour its responsibility to each and every one of these areas.

Ethical considerations will be discussed with reference to the blog corpora used for Study One and Study Two, and the Questionnaire data used for Study Three. There are some over-arching approaches that will be applied throughout this research: all data have been collected and handled in accordance with Aston University School of Languages and Social Sciences Policy on Research Ethics (Aston University Ethics Committee, 2007) and under guidance from my supervisor. The specific ethical considerations for the disguise data are discussed in Chapter 8. The overarching ethical principles remain the same as outlined here, however, the practical aspects of this are discussed more thoroughly in Section 8.2.

Ethical considerations are considerably more complicated in relation to online data, as opposed to the offline world (Gao, Kong, & Sar, 2010); the increased use of the internet in research has made this an even greater area of concern (Bassett & Riordan, 2002). All of the data from the internet that has been used in this project has been taken from public forums. While some other forums online require registration to access them, the data for this project has only been taken from blogs that were open access at the time of collection. In other words, all the data is taken from a totally public sphere and could have been accessed by any one, without registering in any way. It was decided that the requirement to register for a site, to any extent, could indicate a joining of the community at some level and texts published within the community are intended for members of the community. Registering in order to access data could be construed by some members as breaking the code of the social group, therefore it would be appropriate to seek permission and informed consent, which would considerably complicate the methodology of data collection.

The Association of Internet Researchers (AoIR) state that “the greater the vulnerability of the author – the greater the obligation of the research to protect the author.” (AoIR ethic working committee & Ess, 2002, p. 5). They also raise the debate of webpages created by

minors, and how this affects informed consent. In relation to this research, it is difficult to know the age of the authors, as they rarely state it. The content of the blogs suggest that the majority of the participants are adults. One could question whether in an online context, vulnerability could include a limited technological awareness i.e. a lack of understanding that something is publically accessible. However, it could be argued that a degree of technological awareness is required to set up a blog, and the concept of a blog being a form of publication is more transparent than it is in a mixed mode context such as Facebook, Myspace or Twitter.

Basset and Riordan (2002) proposed an alternative view of the ethics of internet research, arguing that the commonly employed human subjects research model is not completely applicable in the online sphere and that we should respect the cultural phenomenon of the online texts. They highlight that there "are issues and rights at stake in these debates other than those of privacy and safety. The internet user is also entitled to a degree of representation and publication in the public domain" (Bassett & Riordan, 2002, p. 244). This is very relevant to the current data, as it is all publically accessible at the time of collection. The authors have chosen to create blogs that are publically available, and while it is important to protect the ethical considerations of research participants perhaps that includes respecting the authors' desire to be represented publically. It should be noted that this research does not intend to comment on the authors' views, beliefs or the opinions expressed within the text. The only focus of this research is the language employed by the authors and their linguistic backgrounds.

The forensic perspective of this research means that there are ethical considerations not just in the realm of collection of data, but also with consideration to the application of this research. Sorell (2011) warns that "university based researchers interested in radicalization may appear to some members of some communities to belong as much to the establishment apparatus as judges" (Sorell, 2011). This is particularly relevant, due to the potential intelligence applications of Native Language Identification. NLID could potentially be misinterpreted as a tool for persecution or judgement, recognising and acknowledging the limitations of this research as well as strenuous focus on unbiased "objective scientific evidence" (Linguistic Society of America Executive Committee, 2011, p. 1), should mitigate this. Conley and Peterson highlighted that a social scientist and expert consultant should be aware "that expertise may have grave consequences for one or more of the litigants, and may also have a significant effect on society itself" (Conley & Peterson, 1996). The

methodology for NLID set out within this research is intended to be considered in relation to the potential consequences and effects. All texts and future reports are (and will be) treated with as much ethical consideration as possible, all stages of the project adhere to the Aston University code of ethical conduct, guidance from my supervisor, the Linguistic Society of America's Code of Ethics for Linguistics in Forensic Linguistics Consulting, and my own moral and ethical code. It is impossible to control completely how the findings of this research will be used. It is feasible that a person or organisation might try and use methodology or findings discussed in this research to persecute individuals or minorities, however, this would entail a wilful misinterpretation of the methods, potential conclusions and limitations of this study.

Chapter 5. Study One. Internet Data : Analysis

Der Grenzen meine Sprache sind die Grenzen meiner Welt

The limits of my language means the limits of my world –

(Wittgenstein, 1922)

This chapter and the next one (Chapter 6), set out the analysis and finding for the main body of data for this research project; two blog corpora. One corpus is of blogs written by authors who identify themselves as native (L1) Persian speakers, and the other is a control corpus by authors who identify themselves as native (L1) English speakers. Due to the amount of analysis and information to be discussed, this study will be documented over two chapters. The first chapter (this one, Chapter 5), sets out the methodology, its development and the initial findings. It also introduces and discusses the evolution of the analytical method that was developed and utilised. Findings demonstrate that it is indeed possible to use interlingual features to indicate the L1 of an anonymous author. The implications of these findings for the rest of this research will also be briefly discussed. The second chapter (Chapter 6) examines the initial findings in greater depth, it details the statistical analyses and the implications of the findings.

Structure of Chapter 5:

- 5.1 Creation of NLID method
- 5.2 NLID coding system
- 5.3 Results and findings
- 5.4 Discussion – comparison to existing studies

5.1 Evolution of NLID Method

As already discussed in Part One there is no set existing methodological approach for Native Language Identification (NLID). There are however many differing approaches to authorship analysis, as has been discussed in Chapter 3. One of the aims for this research is to design and evaluate a potential method of NLID. This section outlines the evolution of this method;

the coding system which evolved will be discussed in greater depth in the next section. The benefits and motivation for choosing a predominantly corpus based approach have already been demonstrated; now we can consider in greater detail the analysis of the corpora, and how this could form an 'NLID method'.

The analysis can be broken into several distinct phases;

- close reading to identify list of potential features
- creation of coding scheme
- coding all texts according to coding scheme criteria (which could also be applied to other languages and research)
- basic analyses (frequencies etc)
- statistical analyses

The first phase was a close read-through of a large sample of the texts (from both the L1 Persian and L1 English corpora), this is commonly an important early phase for authorship analysis, and allows us to better understand what features (in the form of anomalies or marked language use) were appearing. This yielded a raw list of features. The second phase of analysis involved organising this preliminary list of raw features into a systematic coding structure, which could be used to code all the texts and account for all potential markers within the texts. The coding of all the corpora is stage 3. Finally we can then consider the basic findings and results of the coding, this can be seen in section 5.3 (Chapter 6 considers more complex statistical analysis based on the results of the coding). The majority of this analysis was undertaken using the software programme *NVivo*[7], which is created for qualitative analysis and allows better visualisation of trends and more intricate analysis. A close reading of the text highlighted certain features which were given corresponding codes[8] that attempted to describe the feature.

Native Language Identification (NLID) is distinct from error analysis, as it is not concerned with the 'correct' or overly prescriptivist view of the English language, it is also not concerned with predicting the errors learners make. Instead it looks at anomalies, linguistic

[7] various versions were used initially but the NVivo 9 was used for the majority of the analysis.

[8] NVivo terminology is such that you code a feature at a node. For clarity this section of the thesis will refer to the names given to the features as codes unless specifically discussing them in respect to their function in NVivo. It should however be remembered that the terms code and node are relatively interchangeable.

features that are unique to the group of people that constitute L1 Persian speakers writing in English. It also observes how the language is actually used, rather than attempting to predict these features through a close examination of the L1, L2, and - depending on your school of thought - Universal Grammar (or other underlying influences). It is for this reason, and the lack of specifically applicable existing research, that a 'bottom-up' approach was decided best to describe the language of the corpora. This approach allows the method and coding system to develop in a completely data lead manner, rather than imposing methodologies and coding systems upon it. This means that the analytical system produced is not constrained or limited by existing research (and the results are not biased by false predictions) which does not translate well to this context. The preliminary phase is the epitome of the bottom-up approach. By describing every marked feature in the texts, we develop a long list showing the breadth of the potential features. A screen shot of some of the initial features can be seen below (Figure 5-1).

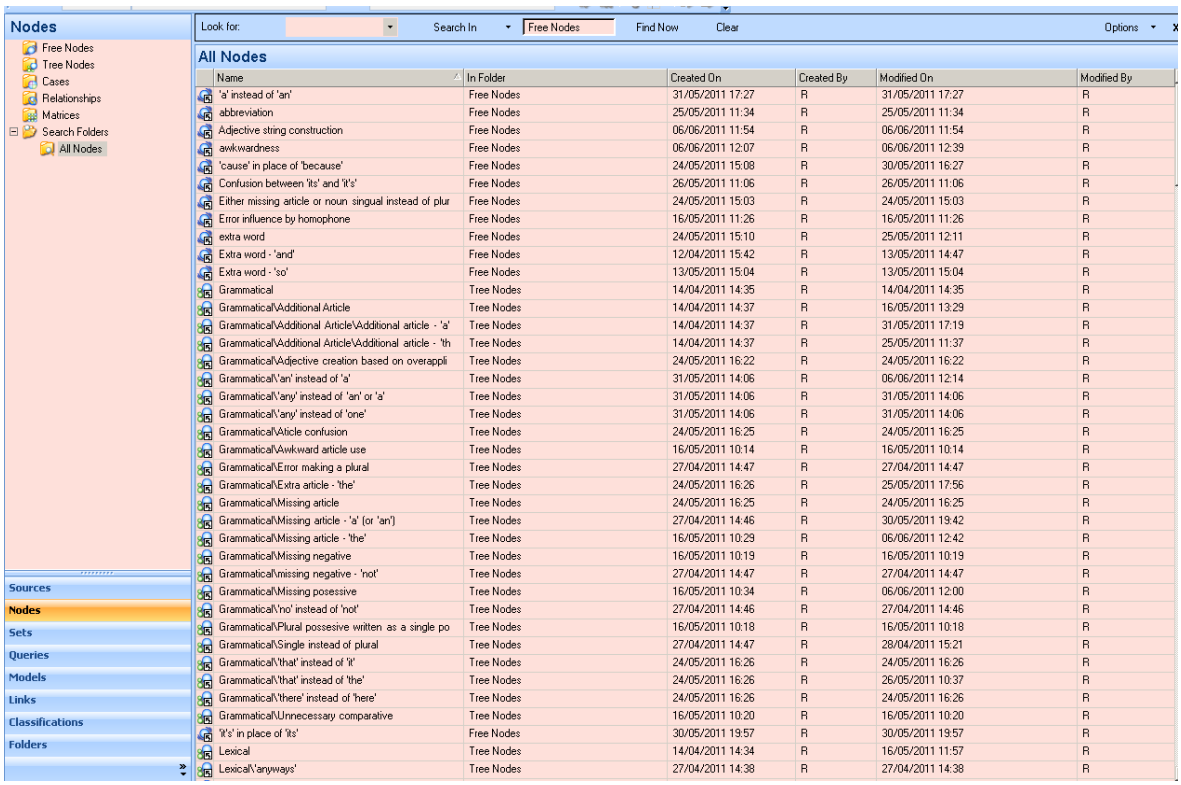


Figure 5-1 - Initial Features NVivo Screenshot

One thing that became clear during this phase is that the list of codes was organic. As more texts were read, many more features were discovered, and the more the definitions of existing codes were called into question.

Stage two of this procedure sought to account for all the features already found, as well as any potential new ones, while simultaneously providing a structure to the coding. It was noticed that there were certain trends in the existing codes. Certain codes describe the motivations or influence behind a marked section of texts (such as a spelling error that seemed most likely to be due to typing methods). The rest of the code described why the text seemed marked. It was noticed that these descriptive codes predominantly pertained to either the marked presence, marked absence, marked construction, or marked choice of certain elements. Furthermore the majority of features could be grouped into vague areas based on what text they were describing; Position and Ordering, Adjective, Adverb, Article, Capitalisation, Conjunction, Lexical, Preposition, Pronoun, Punctuation, and Verbal. A few of the categories did not break down completely into the sub-categories of marked presence, absence, choice or construction, due to the nature of the constructions they were describing. There were also other important categories that could not be avoided; problematic to classify, and unsure (as well as influence, as mentioned previously). These will be discussed more thoroughly in the next section (5.2). This gives us the following underlying structure for the descriptive element of the coding system:

Figure 5-2 - Background Coding Structure

- Article
  - Marked Presence
  - Marked Absence
  - Marked Choice
  - Marked Construction
- Adverb
  - Marked Presence
  - Marked Absence
  - Marked Choice
  - Marked Construction
- Preposition
  - Marked Presence
  - Marked Absence
  - Marked Choice
  - Marked Construction
- Pronoun
  - Marked Presence
  - Marked Absence
  - Marked Choice
  - Marked Construction
- Capitalisation
- Punctuation
  - Marked Presence
  - Marked Absence
  - Marked Choice
  - Marked Construction
  - Marked Position or Ordering
- Conjunction
  - Marked Presence
  - Marked Absence
  - Marked Choice
  - Marked Construction
  - Marked Ordering and Positioning
- Verbal
  - Marked Presence
  - Marked Absence
  - Marked Choice
  - Marked Construction
- Lexical
  - Marked Presence
  - Marked Absence
  - Marked Choice

    - Marked Presence
    - Marked Absence

- Position and Ordering
- Influence

    - Marked Construction
- Adjective
    - Marked Presence
    - Marked Absence
    - Marked Choice
    - Marked Construction

- Problematic to classify/unsure
- Collection Issue

It should be noted that this is only intended as the background structure. It is still important to code the text in as much detail as possible, this follows Gass's (Mackey & Gass, 2005) observations that: "if researchers code data using as finely grained a measurement as possible, the data can always be collapsed into a broader level of coding later if necessary, but finely grained categories are harder, if not impossible, to reconstruct after the data are coded." (Mackey & Gass, 2005, Loc. 63055). Therefore, the background coding listed above is only the framework into which the more finely defined codes fit.

The full coding system and the application criteria are discussed more fully in the next section (5.2). The background coding system as detailed above grew out of the preliminary features that had been identified within the texts. However, one key benefit of the coding system is that it would be applicable for looking at NLID in other corpora. It is expected that the features that fit into this background system will change to reflect the specific nature of the corpus being analysed, however, this background pattern should still be applicable. The coding process should not be considered a static one, this is discussed more fully in the proceeding section, and particularly it is represented in Figure 5-3.

5.2 NLID coding system

This research takes a 'bottom-up' approach, so the codes were determined according to the features that were apparent in the data. For ease of comprehension and analysis, these codes were organised into a tree, loosely based on their grammatical nature. This present section describes the codes, their significance, difficulties or decisions regarding their definition, and the analysis according to the key higher level codes.

As discussed in the previous section, the codes that were initially identified were arranged into a larger tree, which would represent the relationships between similar features. It quickly became apparent, that it is possible to design a tree which not only accounts for the

features already identified, but also allows for the possibility of new features, and could therefore theoretically be applied consistently to similar analysis for other languages. There are two main areas of the coding; the first detailing the linguistic features, and the other which attempts to explain certain features according to the likely influence. For example if 'word' were to appear in the text spelt 'wrod', it is most likely that this is a simple typing error, rather than the author genuinely thinking that this is the correct spelling for the word 'word'. While this is a rather obvious case of a typing error, there could be many less obvious examples, and it would be risky to attribute something solely to a spelling error, when it could actually be a feature of this group of authors. Therefore it should be marked as an error in lexical construction, as well as marked as possibly being influenced by a typing error. In this way the two aims of coding any section of the texts is; to describe a feature, and where relevant to explain it.

Section 5.1 discusses the forming of a template coding structure (which is demonstrated in Figure 5-2 - Background Coding Structure), into which further lower-level codes fitted, to create an extensive tree of features. This full tree can be seen in Appendix E. The finely detailed level of coding seeks to replicate the detailed descriptive element that was visible in the initial codes. Where one of the initial codes may have been 'missing word 'the'' this would be coded under *article – marked absence* and then within the group it would be coded as *–marked absence 'the'*. Within the finest level of description it was not always possible, or indeed beneficial, to create a new descriptive code with more details. For example, it may be apparent that there is a marked absence of an article, but it may not be appropriate to determine which article it is that is absent. It is not the aim of the coder to correct the text, or try to surmise what an author intended, or should have written. (It is for this reason that there is such a necessary large number of references coded as 'unsure' within the problematic to classify section of codes). Therefore in order to avoid 'over-fitting' or excessive supposition[9] the preliminary fine level descriptive code is 'unspecified'. For example, in the scenario above when it is not clear which particular article was absent we would code the section as; *article – marked absence – unspecified*. A second reason that a section of text may be coded as *unspecified* within a certain category, is when there is no specific fine detailed code to describe the feature. Rather than creating an unlimited number of such fine level codes, it is up to the analyst's discretion whether there was a potential

[9] It is recognised that some degree of supposition is unavoidable, and that each coder may have a different perspective. This is discussed more thoroughly in Chapter 9 section 2

trend developing that required such corresponding codes. For example, it would be possible to create a new code for every time a word was spelt wrong, this would result in such an excessive number of fine level codes that they would lose their usefulness as codes. Instead, they each section can be coded as *lexical – marked construction – unspecified* and if the analyst suspects a trend may be developing as the coding progresses, they can create corresponding codes, and where necessary re-code any text that has been coded as *unspecified*. In this way the coding is a constantly evolving circular process, this is represented in Figure 5-3 below. The creation of a data driven coding system requires that the process is organic and open to change, as some useful features and trends may not be apparent until later in the coding process.

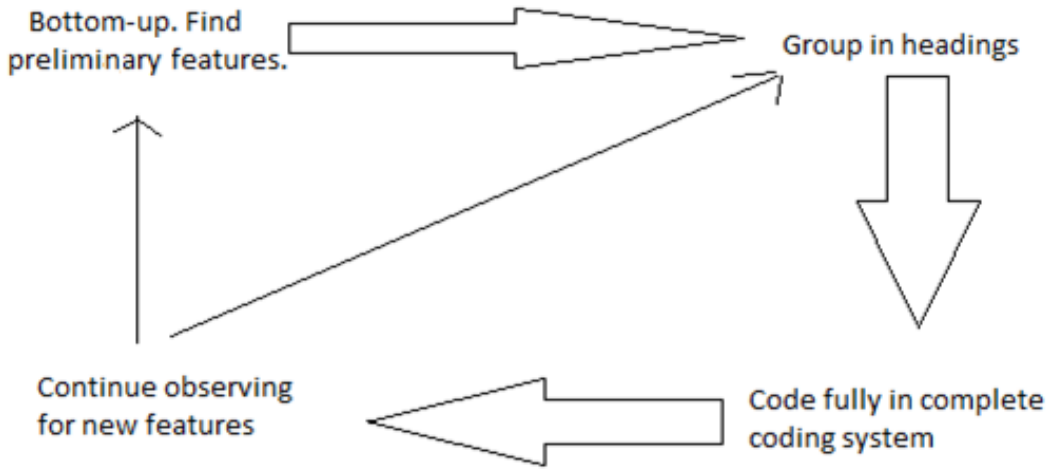


Figure 5-3 - Coding Process

It became apparent during the creating of the coding system, that there was no allowance for differentiating between a code that seemed slightly unnatural and one that seemed exceedingly unlikely to be produced by a native speaker. This could be discussed in terms of errors, however, this research is not focused on prescriptive grammar or how the authors 'should' write. The forensic linguistic aspect of this research means that it focuses on how rare usage of a feature is, rather than how 'correct' it is. Corder (1974b) supported this stating that "the idiosyncratic sentences of a learner" (Corder, 1974b, p. 163) should not be termed ungrammatical, as they are grammatical with reference to the learner's language, even if they are not grammatical according to the L1 or L2. This research will refer to these 'errors' or 'unnaturalness' as degrees of how *marked* a section of language seems. Markedness in this research can therefore be considered to be the difference between the

language of the data, and the language that would be expected from an L1 English speaker. One way to account for the variability of 'markedness' would be to grade each feature. There are various potential ways of doing this, but it was ultimately decided that each section of language that was coded in the grammatical categories mentioned above would be coded not just with a code describing the feature, but also a code describing the degree of 'markedness'. Rather than unnecessarily complicating the analysis further, it was decided that how marked an item was would be limited to either *2-Non-standard* or *1-Awkward*[10]. The following examples demonstrate this distinction as well as giving examples of how the coding system is implemented:

Example 1 (from L1 Persian corpus):

Mohammad Afshar – "After a fairly long summer vacation, I am back to uni now as second year student." In this example this underlined section is coded at the following lower level nodes: *Preposition – marked choice – unspecified*, and *Preposition – marked choice – 1 Awkward*. This is because it is not completely unusual to talk about going back to university, but would be less marked in this context to use the preposition *at*.

Example 2 (from L1 Persian corpus):

Emad – "I experienced problems with bathroom" In this example this underlined section is coded at the following lower level nodes: *Article – marked absence – 'the'*, and *Article – marked absence – 2 Non-standard*. This is quite significantly marked language, therefore is is coded as Non-standard rather than awkward. Naturally there is an element of subjectivity to this, as many codes could be coded at either *1-Awkward* or *2-Non-standard*. As there is only one analyst, decisions are primarily down to the linguist's judgement, with a strenuous attempt to maintain consistency throughout the analytical process. (Variability within the coding and the risk of analyst error is discussed more fully in section 9.2). Where unsure, Google was used as a corpus to aid understanding of how marked a feature is.

The coding system is designed so that the codes are independent and when a feature is coded at multiple codes, this does not cause complications for the statistical analyses. The texts were all coded using the qualitative analysis software NVivo. A range of versions was used, but NVivo 9 was the main version employed. A USB stick with the complete version of

[10] One could also consider that there is also a *0-native-like language*, and that any section that is not coded, is by default coded as this.

the NVivo project can be found attached, the NVivo project is Appendix F. The L1 Persian and L1 English texts were simultaneously analysed (rather than one group followed by the other) to minimise subconscious bias from the coder and to promote consistency in the application of coding. The use of NVivo introduces some disparity in terminology, what so far has been termed a *code* or *feature* is termed a *node* by NVivo. For clarity the NVivo term will be used when referring specifically to NVivo and the methodology therein. Once the texts were coded at the lower level nodes of description and markedness, all nodes were then merged with their higher level (parent) nodes. Merging means that each section of text that is coded, is counted once at a higher level node, even though each section of text is coded more than once (at a descriptive code, and a markedness code). This then demonstrates the distribution of features across the feature tree. The features were then exported from NVivo into a spreadsheet representing how many pieces of texts were coded at each node per author, this can be seen in Appendix H. The coding matrix was then adjusted according to the word count for each author, to give the normalised number of features per 2,000 words. In order to do this the number of items coded at each feature was divided by the total number of words collected for that particular author, then multiplied by 2,000. The normalised coding matrix can be seen in Appendix I and it is these normalised results that are used for all continuing analysis and graphical representations (unless expressly stated otherwise).

5.3 Results and findings

As well as coding all the blogs according to the system as set out above, there are several initial analyses that it is worth considering. A selection of the key analyses will be discussed here, it should be noted that with this volume of data there are many more different ways of viewing it, however, only the ways considered most pertinent have been discussed here due to time and space constrictions. First this section will discuss general observations from the coding, and the word frequencies (section 5.3.1). Next this section will show that the data appears to follow Zipfian distribution rules (section 5.3.2). Finally the feature frequencies will be discussed, along with the implications and indications for the rest of the research.

The data overview in Appendix B contains some basic statistical information about the blogs collected from each author. It is interesting to note that for the L1 Persian authors, the average number of paragraphs is higher, yet the average paragraph length is longer. This indicates that the L1 Persian authors write in chunks of texts that are divided up into shorter paragraphs. It is also worth noting that the average blog length for the L1 Persian authors is considerably lower, which will affect the average paragraph length, however, the number of

blogs per author are higher. This suggests that authors are posting short frequent blogs rather than occasional longer posts, which is in keeping with the concept of Weblogistan as a society. Perhaps the most interesting preliminary statistic is that the average word length is higher for the L1 Persian group than the L1 English writers. This seems counterintuitive, as the L1 Persian group are writing in their second language, whereas L1 English speakers are using their first. Within both Persian and Arabic history, there is a strong literary tradition, with writing being considered as a popular art form in a way that has been lost in modern Western society. This concept of writing as an art may encourage the use of longer, more elaborate words. It should be noted that these statistics were calculated using information gathered from Microsoft Word, this does not allow for many of the complexities within the data, meaning they may provide interesting observations but are unlikely to be useful for NLID.

Frequency lists are a good way to develop an initial understanding of the data. As Gass wrote: "Frequencies [...] are often presented in second language studies even when they do not relate directly to the research questions. This is because frequency measures provide a succinct summary of the basic characteristics of the data, allowing readers to understand the nature of the data with minimum space expenditure." (Mackey & Gass, 2005, Loc. 6871). This is supported by Barnbrook (1996) who recommended considering frequency lists as a preliminary analysis as, "Despite its apparent simplicity, it is a very powerful tool" (Barnbrook, 1996, p. 43) especially when it is possible to compare lists from similar corpora. Appendix G shows the word frequency lists for the L1 Persian and L1 English corpora, as compiled by NVivo. As expected there is a lot of similarity in the most frequently occurring words, with the top 12 most frequent words being identical for each corpus, with only a little variation in their exact ordering. These top 12 words are: *the*, *and*, *to*, *of*, *i*, *a*, *in*, *is*, *it*, *that*, *my*, and *for*, these are all function words. The frequency lists help highlight the difference between the corpora, this can be seen by the increased use of the words *Iran* (38th, 164 occurrences, 0.32%), *Iranian* (79th, 78 occurrences, 0.15%) and *Persian* (83rd, 74 occurrences, 0.15%) with the word *English* being used slightly less at 69th position at 0.14% of the Persian Corpus. This contrasts to the L1 English corpus in which the most frequently used term relating to nationality or language is *English* at 150th with 45 occurrences constituting 0.08% of the corpus. From this we can infer that language or nationality features more significantly in the content of the L1 Persian blogs. This could to some extent be expected as the authors are all writing in an L2, however, it could also in part be due to the nature of Weblogistan and the importance of language and nationality to the Weblogistan identity. The latter is

supported by the more frequent use of the terms *government* and *people* in the L1 Persian corpus, indicating that the Persian corpus has a greater focus on political issues. These observations are again of limited use from an NLID perspective, they are a simple way to get a basic understanding of the texts. The lack of accuracy can be seen through the high occurrence of the strong *3a00* which is ranked 37th in the L1 Persian corpus, with 164 uses at 0.32% and 93rd in the L1 English corpus, with 76 uses at 0.14%. This string is found in the URLs between punctuation marks, therefore NVivo assumes they are words.

"In order to visualize trends in the data, it is generally useful to plot the data even before carrying out statistical analysis." (Mackey & Gass, 2005, Loc. 673). Zipf's law states that linguistic features such as word and error distribution will often follow a log-normal distribution (Zipf, 1932). It is therefore interesting to note that if we plot the Frequencies of each feature according to their rank (Figure 5-4), and the number of occurrences of the word at each rank (Figure 5-5), then the graphs seem to follow a clear log normal distribution. This is reassuring as it demonstrates that the features are following the expected distribution patterns, and are reflecting well established trends within language and linguistics. This suggests that they are not an arbitrary set of random features, but are representative of linguistic structures in the data.

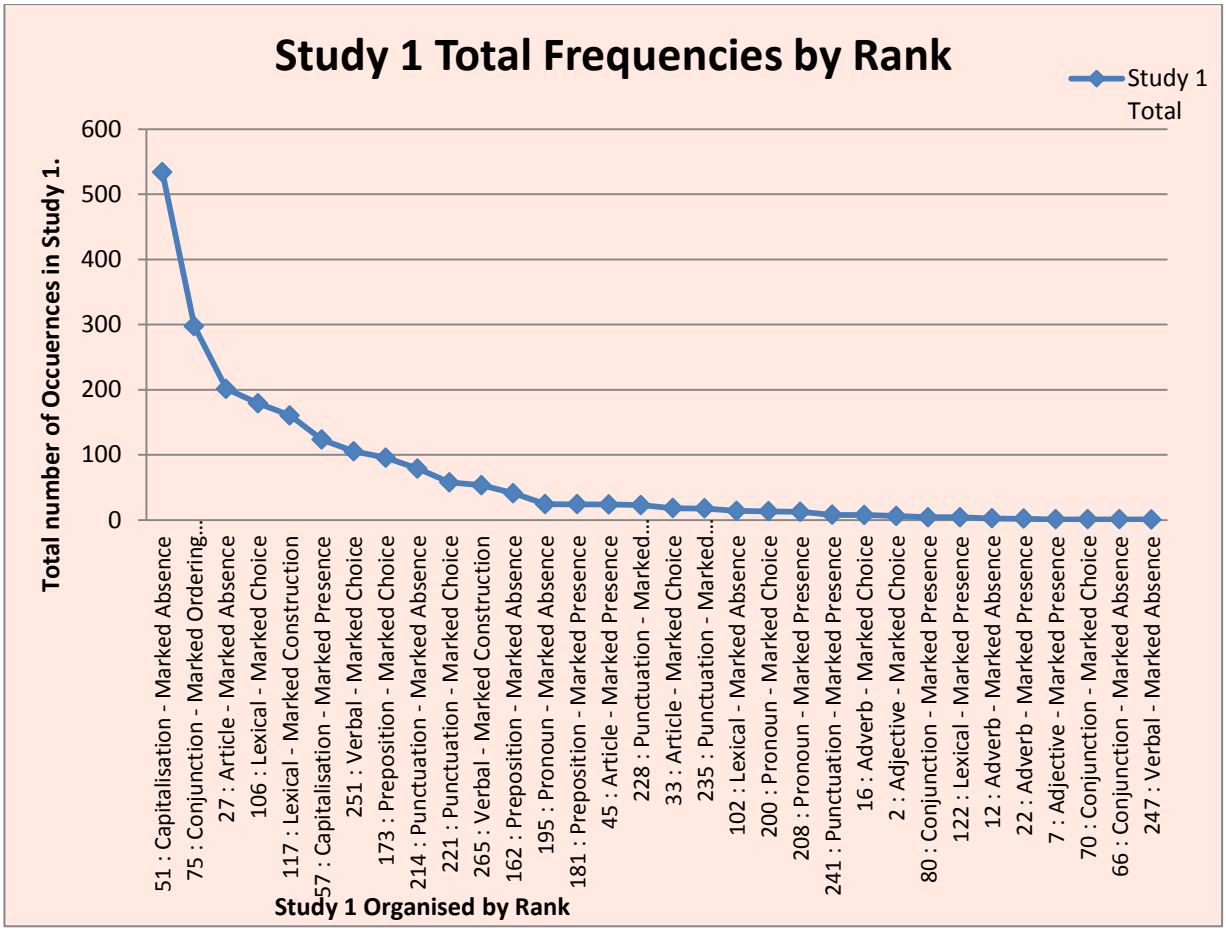


Figure 5-4 Study 1 Feature Frequencies

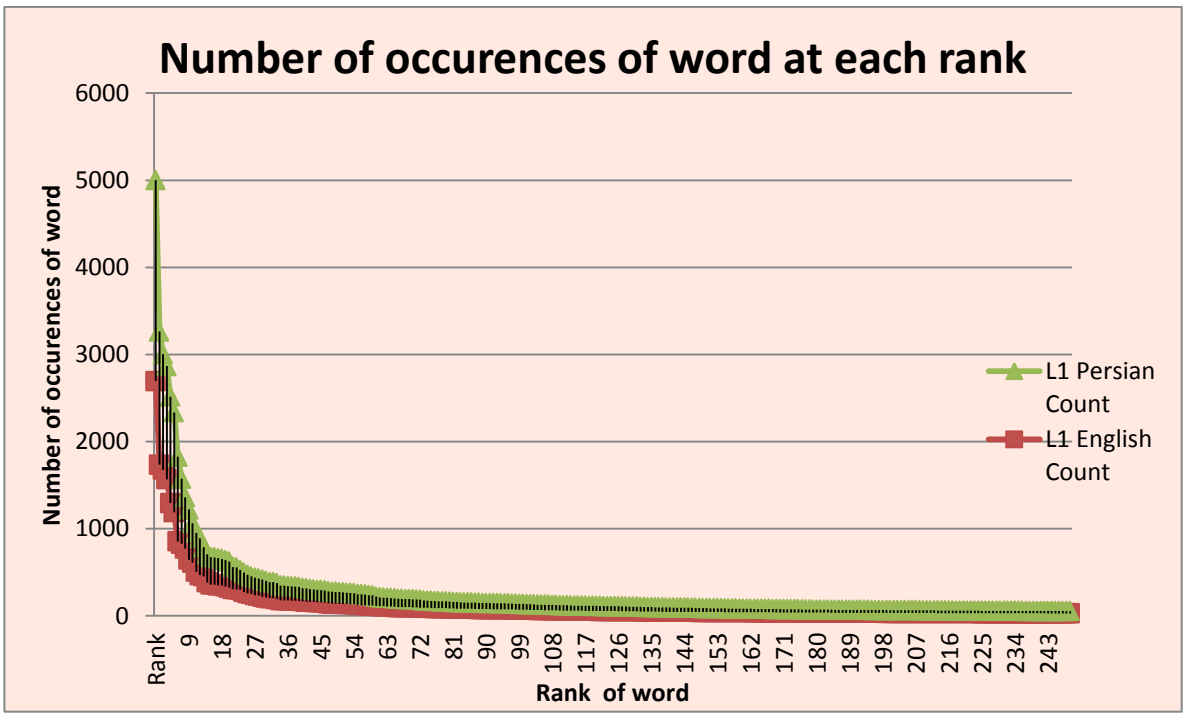
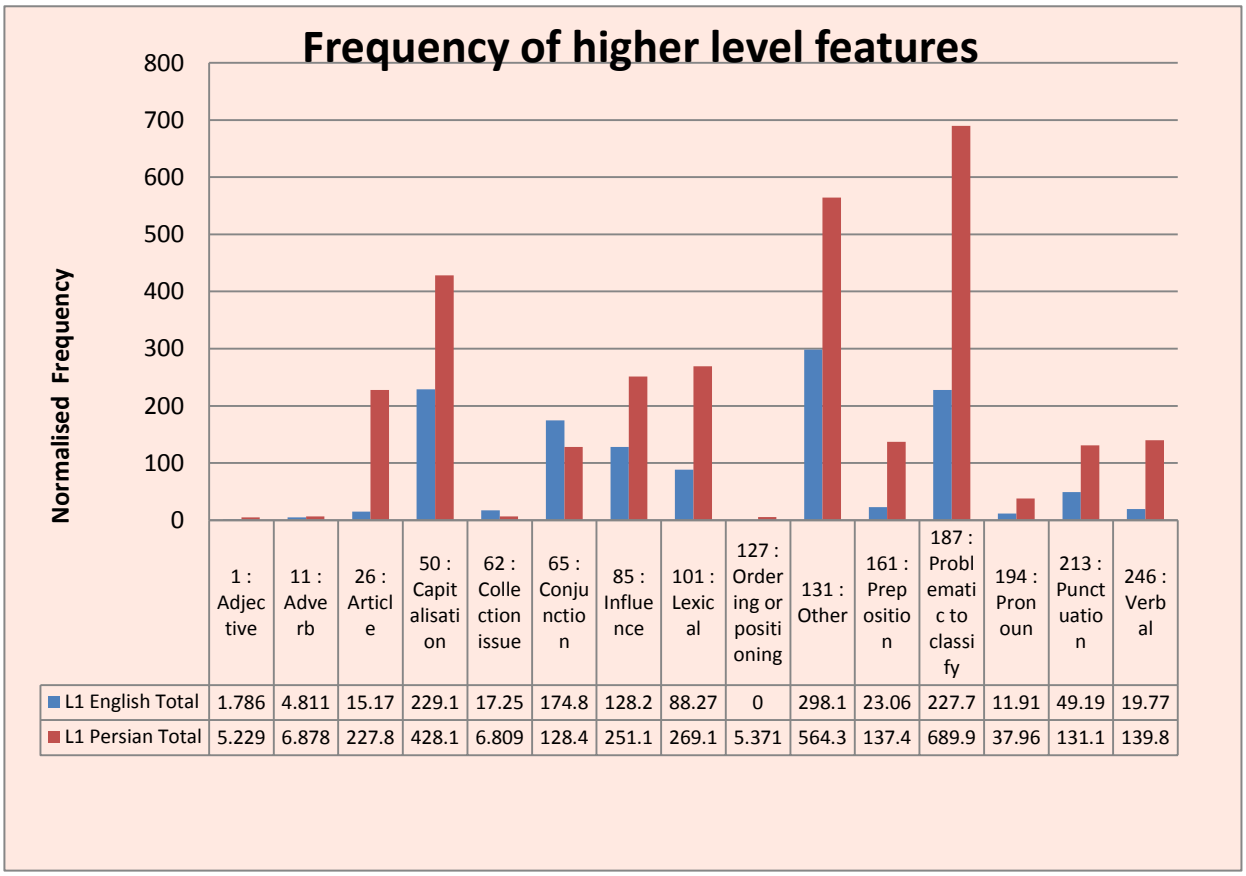


Figure 5-5 Study 1 Word Frequencies by Rank

The first observation we can draw about the frequency of the features, is that within both groups, *problematic to classify* and *other* are both the most frequent high level nodes. It is expected that *problematic to classify* would have a very high occurrence rate, as it is not always possible to code the text without making assumptions about the author's intended meaning. These assumptions are to be avoided as much as possible, therefore it is necessary to code such sections of text as *unsure* and *problematic to classify*.



	1 : Adjective	11 : Adverb	26 : Article	50 : Capitalisation	62 : Collection issue	65 : Conjunction	85 : Influence	101 : Lexical	127 : Ordering or positioning	131 : Other	161 : Preposition	187 : Problematic to classify	194 : Pronoun	213 : Punctuation	246 : Verbal
L1 English Total	1.786	4.811	15.17	229.1	17.25	174.8	128.2	88.27	0	298.1	23.06	227.7	11.91	49.19	19.77
L1 Persian Total	5.229	6.878	227.8	428.1	6.809	128.4	251.1	269.1	5.371	564.3	137.4	689.9	37.96	131.1	139.8

Figure 5-6 Study 1 Frequencies of Higher Level Features

Appendix J shows the frequencies of the lower level features. As stated in sections 5.2 and 5.3 this research is more than an error analysis, it focuses on features that are marked in any way (this can include hyper-correctness). These two graphs support this as there are several points in which the L1 English group contains a higher number of occurrences than the L1 Persian group, the most significant of which is *Conjunction – marked ordering and positioning* which had rate of 175 occurrences per 2,000 words in the L1 English corpus, and 122 in the L1 Persian. The most frequent feature in both texts was *Capitalisation – marked absence*, the fact that this is the most frequent feature in both corpora demonstrates the

need for further statistical testing, as the high frequency of one feature, does not mean it is unique to a corpus.

5.4 Discussion

There are clear differences between the feature distribution of the texts, these will be analysed further in the next chapter. The Zipf curves indicate that the features demonstrate the expected distribution in the data.

Inter-coder and Intra-coder reliability are two areas that link closely to the issue of subjectivity, and ensuring continuity throughout the coding phases. Inter-coder reliability refers to the tendency for two different coders or analysts to interpret the texts differently, and hence code them differently. Intra-coder reliability relates to how one particular coder may view a text differently due to any number of factors such as different times, or different context. There is an element of subjectivity, even in the best coding system. This is even more the case when we are considering texts that have a high volume of 'errors'. In the same way that each reader understands a text slightly differently, there are also different ways of interpreting what it is, an author is trying to say. This research aimed to overcome the impact of subjectivity in several ways all aimed to promote consistency; only having one coder, making extensive notes on decisions relating to definitions of the codes, and accepting that some features must be coded as *unsure*. Having only one coder, meant that it was easier to keep consistency across the decisions, both at a conscious and subconscious level, as "Individual coders can come to internally stable views" (Carletta et al., 1997, p. 29). However, this means a greater time investment from one analyst, and may not always be beneficial in a casework setting (as discussed in the next paragraph). Another method for diminishing the impact of subjectivity was to promote consistency through defining each code as well as possible and keeping detailed notes of any decisions made about difficult classifications. This ensured as much consistency as possible, and could be considered a vital part of the research, even more so if there were more than one analyst. The final, and arguably most important method, was for the analyst to recognise that they would not be able to code every section of text, without placing too much of their own interpretation on it. In a teaching situation it is natural to correct language by assuming what someone is trying to say, this is not a teaching situation. Therefore, the analyst has to accept that a large number of sections need to be coded as *unsure*, and that this does not constitute failure on

the analyst's part, rather it is the nature of the analysis. One way to test the extent of the subjectivity would be to study how multiple analysts would code the texts differently, and then analysis if this had a statistical impact on the overall findings. Due to the limitations of this research, it was not feasible to include it in this project, but it would be a useful extension and possible path for further research in the future.

Most existing research focuses predominantly on single structure, therefore it is hard to compare the frequency findings here for validation. However, it is interesting to note that of the most frequent features, Wilson and Wilson (2001) suggested that these areas might be areas of difficult for L1 Persian speakers learning English. We can see from the coding and initial analyses that there are some apparent differences between the two corpora. It is likely that these apparent differences in features within the text, will translate well to the statistical analyses, and that there will be significant statistical differences between the two corpora. Part of the statistical analysis will aim to see if we can predict group membership (i.e. which of the two corpora any given text belongs to) based on the features and codes identified, and with what degree of reliability such a predication has. The differences that can be seen now are not definitive indicators that the features will have a high predictive value, but it is indicative that this may be the case. The statistical analyses will also distinguish how predictive the features are, this is different from the frequency; as a feature that has a high number of occurrences, may not necessarily have a high predictive ability. Chapter 6 forms the second part of this sub-study. Based on the results and findings already discussed, it takes a more statistical look to evaluate the coding system and its predictive properties.

Chapter 6. Study One – Internet Data: findings

عقل که نیست جون در عنابه.

Lack of wisdom brings torment to the spirit (Persian Proverb)

The frequencies discussed in Chapter 5 are interesting and give us information as to which features we are likely to encounter in the case of an L1 Persian author writing in English. However, we cannot tell how significant or unique these features are to L1 Persian speakers. The ultimate practical aim is to use just the features, to determine which group a text should belong to. This chapter builds on the initial findings of the previous chapter with more complex statistical analyses. The first section, 6.1, discusses the statistical methodology. The next two sections 6.2 and 6.3 detail the progression of the statistical analyses. Finally, Section 6.4 reviews the implications of the findings from the statistical analyses.

- Structure of Chapter 6:
  - Statistics plan
  - Higher level features and all lower level
  - Refined feature set
  - Discussion, implications for rest of studies

6.1 Statistics plan - Logistic regression analysis

Logistic regression is a statistical analysis for predicting the outcome of a situation based on a series of variables, which in this case are the features that have been identified and coded as nodes. It is similar to linear regression, except that linear regression has a continuous outcome and continuous predictor variables, whereas logistic regression allows for non-continuous predictors and a dichotomous outcome. For example, if I wanted to predict how many fig biscuits I would eat in an afternoon based on the number of cups of coffee consumed, I could plot the data from previous afternoons and determine that I consume approximately twice as many biscuits as cups of coffee, plus an extra two biscuits directly after lunch (see Figure 6-1 below). This gives us a linear equation of y=2x+2 (where y= the number of fig biscuits eaten, and x=the number of cups of coffee consumed). If I then know that I will consume 3 cups of coffee tomorrow, I can predict that I will eat 8 biscuits. Should the number of coffees increase to 4, then the number of fig biscuits I eat will be likely to

increase correspondingly to 10, as we have determined that there is a direct linear correlation between the number of coffees drunk, and the number of fig biscuits eaten.

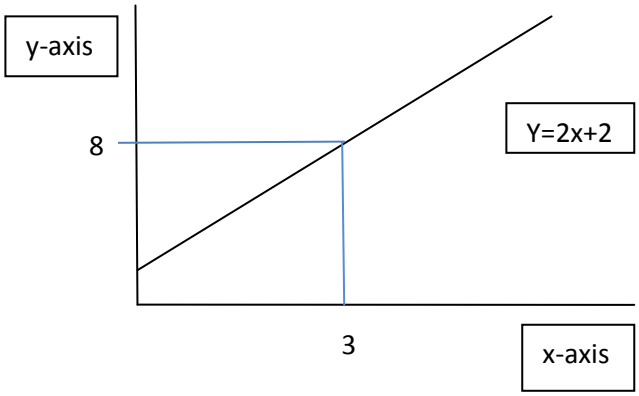


Figure 6-1- Linear Regression of Fig Biscuits to Coffee

In the context of this research, there are two discrete potential outcomes; that any given text could belong to the group of native Persian speakers, or the group of native English speakers. Therefore, rather than using the distribution of features (x) to predict a linear answer as above, instead the y-axis represents the linear probability that given the distribution of features, the author will belong to a certain group.

There are two main forms of logistic regression; binomial which has two potential outcomes, and multinomial (also called polychotomous logistic regression) which has many discrete potential outcomes. The analysis here has two potential outcomes; that an author is an L1 Persian speaker, or an L1 English speaker, therefore binomial logistic regression is the most appropriate (if there were more than two potential outcomes, then multi-nomial logistic regression would be the most appropriate analysis). Weisburd and Britt (2007) also recognised the potential of logistic regression as a tool for criminal justice.

The statistical programme SPSS[11] has the capacity to perform complex statistical analyses quickly and accurately, therefore the analyses discussed here were performed using SPSS. Using SPSS enables quicker and more accurate computation of the statistical analyses. The use of SPSS has necessitated some changes in the titles of the features, this is purely for ease of manipulation and has no bearing on the findings. When exporting the features, NVivo

[11] Specifically the majority of the analysis was performed using IBM SPSS Version 20

automatically assigned each feature a number before its name, this also aided recognition of features and avoided confusion while carrying out the statistical analyses.

Once the binomial logistic regression has been run, certain pertinent results in the output can be examined. The Wald-chi square statistic evaluates the significance of the features, and which are the best at aiding in determining group membership. This distinguishes between which features are simply the most frequent, and which are the best determiners. The other key test is the Hosmer-Lemeshow goodness of fit test. This is a statistical analysis of how good the model is at predicting the outcome of the group memberships. In this case it gives a measurable way of testing how good the selected features are at predicting whether any one of the given texts belongs to the group of L1 Persian speakers, or the L1 English speaker group. These tests, and other relevant statistics generated during the logistic regression analysis will be discussed where relevant in the following sections.

6.2 Initial analyses

6.2.1 Higher level features:

As discussed in Section 5.2, the codes are arranged in a tree formation. This means that not all features are equal or directly comparable when it comes to statistical analyses. In order to get an initial understanding of the data, it is interesting to focus purely on the higher level features (which represent the overarching areas e.g. *preposition*, *article*, *lexical*). This will enable us to see which areas might hold the most discriminatory features, as well as indicating how effective the higher level features alone might be at discriminating group membership. If one area in particular holds a particularly high significance then that might indicate further investigation into this particular area would be of interest.

The binomial logistic regression was run using SPSS, selecting just the higher level features to be variables. The variable names were altered slightly due to the process of exporting the data from NVivo, into Microsoft, then exporting it into an SPSS spreadsheet. As stated in the previous section, this does not affect the analysis, and the full feature name is still included after features' corresponding letters and number. Twelve features were used in this; *1Adjective*, *11Adverb*, *26Article*, *50Capitalisation*, *65Conjunction*, *101Lexical*, *127Orderingorpositioning*, *131other*, *161Preposition*, *194Pronoun*, *213Punctuation*, and *246Verbal*. Higher level features that did not relate to specific marked linguistic features were excluded; this includes the following higher features; *unsure*, *collection issue*, and

influence. The full output can be seen in Appendix K. The first key piece of information contained in the output, is the Classification Table (see Table 6-1 below), this shows an approximate accuracy of the results by visually displaying the distribution of the authors according to the features (predicted), versus the actual group membership (observed). Group 1 relates to the group of L1 English speakers, Group 2 is the group of L1 Persian speakers.

Table 6-1 - Study 1. Higher level features, Classification Table

Classification Table[a]

	Observed		Predicted		
			L1 Group (1=English, 2=Persian)		Percentage Correct
			1	2	
Step 1	L1 Group	1 (English)	25	0	100.0
		2 (Persian)	0	25	100.0
	Overall Percentage				100.0

a. The cut value is .500

Here we can see that there is complete separation in the data. This is indicative that the features may be overly fitted to the specific data, and not generalisable to other data. As this initial stage only focuses on the highest level features, which are not in themselves descriptive, we can expect that this will change later.

Also included in the output in Appendix K is a table tiled *Variables in the Equation* which among other information, details the Wald-chi score for each feature. The Wald-chi score shows the significance of each feature for determining group membership. The higher the Wald-chi score, the more determining power the individual feature has. If we rank the features (see Table 6-2 below) according to their Wald-chi score; we can see which features are more predictive.

Table 6-2 - Study 1. Higher features by Wald-Score

Feature	B	Wald	Sig.
@127Orderingorpositioning	-509.865	.0008865	.976
@11Adverb	-746.292	.0008284	.977
@213Punctuation	20.766	.0007507	.978
@50Capitalisation	-9.332	.0006286	.980
Constant	-197.524	.0005102	.982
@1Adjective	136.514	.0004141	.984
@26Article	53.203	.0003695	.985
@246Verbal	17.433	.0002920	.986
@101Lexical	24.585	.0001572	.990
@65Conjunction	5.819	.0000284	.996
@194Pronoun	-24.325	.0000153	.997
@131Other	2.587	.0000114	.997
@161Preposition	5.957	.0000098	.998

The importance of considering the statistical significance of the features, as opposed to focusing on the frequency counts, can be seen by the fact that the most significant features in the predictive algorithm are not the same as the most frequent. According to the Wald-chi square statistic, the most predictive higher level feature is *orderingorpositioning*, which has the lowest number of total occurrences of all the higher level features with only 5 occurrences in the Study One data.

The Wald chi square score is low for all of the features, less than 0.0009 in every case. This means that individually the higher level features are not particularly predictive. However, we can see from the classification table in the output, that through using only the higher level features it is possible to correctly predict which group each author belongs to. It is interesting to note that *Preposition* is the least predictive feature. This contradicts Wilson and Wilson's (2001) prediction, as discussed in Chapter 2 section 4, that prepositions are likely to be a difficult area for L1 Persian speakers using English. The high predictive ability of *Adverb* demonstrates the importance of considering more than the frequency counts, as this feature is among the least frequent for both corpora. On their own the higher level features give us very little information about the linguistic features actually present in the data. Another problem with purely focusing on the higher level features is that these groups are not clearly delineated (as discussed in section 5.2). Therefore, the next stage of analysis is to expand the linguistic nodes to see how predictive the lower level features are, and which are the most predictive.

6.2.2 Lower level features – linguistic

Rather than considering all the lower level features, it is more interesting to consider the features that are purely linguistic. As both *punctuation* and *capitalisation* could be considered more orthographic than linguistic, we will include both of these features at the higher levels, rather than replacing the higher level features with their lower level constituents. The list of lower level features is as follows:

1. Adjective – Marked Choice
2. Adjective - Marked Presence
3. Adverb – Marked Absence
4. Adverb – Marked Choice
5. Adverb – Marked Presence
6. Article – Marked Absence
7. Article – Marked Choice
8. Article – Marked Presence
9. Capitalisation**
10. Conjunction – Marked Absence
11. Conjunction – Marked Choice
12. Conjunction – Marked ordering or positioning
13. Conjunction – Marked Presence
14. Lexical – Marked Absence
15. Lexical – Marked Choice
16. Lexical – Marked Construction
17. Lexical – Marked Presence
18. Ordering or positioning *
19. Other *
20. Preposition – Marked Absence
21. Preposition – Marked Choice
22. Preposition – Marked Presence
23. Pronoun – Marked Absence
24. Pronoun – Marked Choice
25. Pronoun – Marked Presence
26. Punctuation**
27. Verbal – Marked Absence
28. Verbal – Marked Choice
29. Verbal – Marked Construction

There are a total of 29 features, despite the fact that the model frame work allowed for more. This is because the model allowed for more features than occurred in the data, and any features that did not occur in the data are redundant for the statistical analyses. A binomial logistic regression was run using SPSS and the same method as before, altering only the features that were selected as variables. The full output can be seen in Appendix K. As before (in Table 6-1 - Study 1. Higher level features, Classification Table) the classification table shows that using just the lower level linguistic features we get a theoretical accuracy of 100%, in that all of the authors were assigned to the correct groups according to their feature distribution.

While this may initially seem to indicate that the features and the model are exceedingly good, it actually indicates that there is 'over-fitting' of the model to the data. This means

that while the features and model are very good for this particular data, they could be considered too good, as they would not be applicable to any other cases. As the purpose of this research is to create a model that would be useful in casework situations, it is very important to avoid over-fitting. One way of doing this is to reduce the number of variables, (that is, features) that are used. The next section (6.3) will discuss the reducing of the variables. We can also see that the variables are overfitted through the Hosmer-Lemeshow Test (Table 6-3). The Hosmer-Lemeshow test and the Model Summary evaluate the fit of the model. A significance score lower than 0.05 indicates that the model does not fit the data well. A score of 1.000 as can be seen here, shows that the model is over fitted to the data, so would not be generalisable to other data. The ideal significance score is over 0.05 but less than 1.000. Reducing the number of features will alter this score and make the model more reliable.

Table 6-3 - Study 1 - Lower level features Hosmer-Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	8	1.000

Despite the fact that this model seems to be over fitted, there is still very useful information contained within the analysis, relating specifically to the significance of the individual features, and this is vital for the later stages of the analysis. As before we can arrange the features used according to their Wald Chi statistic (see Table 6-4 below). As the model was overfitted, SPSS was unable to calculate the Wald Chi square scores in Block 2, which is the more accurate version. However, SPSS always calculates the scores for the features in Block 1, which is represented in a table labelled Variables Not in Equation. Where possible the Wald chi scores for each feature will always be taken from the Block 2 table, as this is more accurate, however when the model is overfitted then the Block 1 figures will be used.

Table 6-4 Study 1 All lower level features arranged by Wald

Feature	B	S.E.	Wald	df	Sig.	Exp(B)
@122LexicalMarkedPresence	82.310	90154.703	.00000083354	1	.999	5.57978E+35
@66ConjunctionMarkedAbsence	299.983	502712.437	.00000035608	1	1.000	1.909E+130
@22AdverbMarkedPresence	-523.913	951154.098	.00000030340	1	1.000	2.9328E-228
@251VerbalMarkedChoice	22.558	49757.869	.00000020553	1	1.000	6262433352
Constant	-55.947	143757.874	.00000015146	1	1.000	5.04164E-

Feature	B	S.E.	Wald	df	Sig.	Exp(B)
						25
@102LexicalMarkedAbsence	29.838	78455.608	.00000014464	1	1.000	9.08861E+12
@265VerbalMarkedConstruction	-54.142	147242.032	.00000013521	1	1.000	3.06619E-24
@200PronounMarkedChoice	-85.467	235682.236	.00000013151	1	1.000	7.62358E-38
@208PronounMarkedPresence	-73.280	272708.586	.00000007221	1	1.000	1.49516E-32
@33ArticleMarkedChoice	56.047	222898.283	.00000006323	1	1.000	2.19336E+24
@45ArticleMarkedPresence	-30.768	136880.656	.00000005053	1	1.000	4.34212E-14
@106LexicalMarkedChoice	21.795	116643.038	.00000003491	1	1.000	2919447078
@131Other	-.505	4232.876	.00000001425	1	1.000	0.60337427
@75ConjunctionMarkedOrderingorpositioning	2.600	24641.688	.00000001114	1	1.000	13.469089997
@70ConjunctionMarkedChoice	-204.400	2110681.344	.00000000938	1	1.000	1.69849E-89
@80ConjunctionMarkedPresence	132.401	1489377.181	.00000000790	1	1.000	3.16838E+57
@195PronounMarkedAbsence	7.292	154401.421	.00000000223	1	1.000	1468.078516
@7AdjectiveMarkedPresence	-54.327	1293382.788	.00000000176	1	1.000	2.54805E-24
@117LexicalMarkedConstruction	3.241	79517.120	.00000000166	1	1.000	25.559139334
@127Orderingorpositioning	29.612	753214.777	.00000000155	1	1.000	7.2512E+12
@173PrepositionMarkedChoice	3.872	101758.120	.00000000145	1	1.000	48.058637774
@12AdverbMarkedAbsence	-26.145	736776.370	.00000000126	1	1.000	4.42167E-12
@181PrepositionMarkedPresence	-21.122	612654.077	.00000000119	1	1.000	6.70988E-10
@162PrepositionMarkedAbsence	1.433	43838.577	.00000000107	1	1.000	4.192442159
@2AdjectiveMarkedChoice	5.690	242384.888	.00000000055	1	1.000	295.7720283
@247VerbalMarkedAbsence	-90.970	4620452.412	.00000000039	1	1.000	3.10568E-40
@27ArticleMarkedAbsence	-1.030	69118.453	.00000000022	1	1.000	0.356879386
@50Capitalisation	-.228	23747.113	.00000000009	1	1.000	0.796039083
@213Punctuation	.353	60217.228	.00000000003	1	1.000	1.423251406

Feature	B	S.E.	Wald	df	Sig.	Exp(B)
@16AdverbMarkedChoice	3.274	1258322.290	.00000000001	1	1.000	26.40372499

Compared to just the higher level features we can see that all of the individual features have even lower Wald-Chi square scores. This means that each feature on its own does not have a high predictive ability, but the model is accurate due to the combination of all the features. The column labelled *Sig.* indicates the statistical significance score for each feature. A score less than 0.05 would indicate that the score is statistically different from zero. In this case however, all of the scores are considerably closer to 1 than 0.05. Together these show that although the features are exceedingly predictive when all 29 are used collectively, each individual feature is not particularly predictive alone. Due to the overfitting demonstrated in the Model Summary above, we need to reduce the number of features. This can be done by selecting the most significant features according to the Wald-chi statistic of each feature, where a higher score means the individual feature is more significant at predicting group membership. The method of reducing the variables and reaching a final model is discussed in the following section (section 6.3).

6.3 Progression of statistical analysis

The previous section detailed the method for ranking all 29 lower level features according to their individual significance for predicting group membership for author, between the L1 Persian group or the L1 English group. The Output demonstrated that using all 29 features meant that the model was overly specific and 'too good' at determining group membership, which indicated that the features would not be transferable to other data. Reducing the number of features will increase the generalisability of the model. When reducing the features, it is best to keep the features that are the most significant at determining group membership. The features were ranked according to their individual Wald-chi scores, as determined by performing binary logistic regression using all 29 lower-level features. This was detailed in the previous section, and ranked features can be seen in Table 6-4 above. The output from performing binary logistic regression with all 29 features demonstrated that the model was overfitted, as the Hosmer-Lemeshow test had a significance of 1.000. In order to find a more generalisable model, a series of binomial logistic regressions can be run, each time removing the least significant features, and watching how this affects the Hosmer-Lemeshow test scores, and the Model Summary statistics.

The Hosmer-Lemeshow test is described by SPSS as being the most reliable way to evaluate the fit of the model within SPSS (Pallant, 2010, loc. 3234). A significance score of less than 0.05 indicates that the model is a poor fit to the data. When the model contained all 29 features earlier, the significance score was 1.000, this indicates that the model is over fitted. Therefore, the ideal significance score is over 0.05 but less than 1.000. It is expected that as we reduce the number of features, the Hosmer-Lemeshow significance score will decrease. Over fitting was found until the number of features was reduced to the top 10 most significant according their Wald-chi scores. The 10 features were:

1. 122LexicalMarkedPresence,
2. 66ConjunctionMarkedAbsence
3. 22AdverbMarkedPresence
4. 251VerbalMarkedChoice
5. 102LexicalMarkedAbsence
6. 265VerbalMarkedConstruction
7. 200PronounMarkedChoice
8. 208PronounMarkedPresence
9. 33ArticleMarkedChoice
10. 45ArticleMarkedPresence

This gives us the following results for the Hosmer-Lemeshow Test:

Table 6-5 - Study 1. 10 Features Hosmer Lemeshow test.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.128	6	.793

The Chi-square value is 3.128 with a significance value of 0.793. This is considerably higher than 0.05 yet less the 1.000 indicating that using the ten selected features generates a model that is very reliable, but not over fitted to the data. This is also supported by the Model Summary (Table 6-6)

Table 6-6 - Study 1. 10 Features Model Summary

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	24.171[a]	.595	.793

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

The Cox and Snell R Square and Nagelkerke R Square values are pseudo R square statistics. They signify approximately how much of the variability is explained by the chosen features, which in this case is between 59.5 percent and 79.3 percent. The higher the percentage (or the closer the score is to one) the more accurate the model is.

Table 6-7 - Study 1 - Optimum Model Classification Table

Classification Table[a]

	Observed		Predicted		
			L1 Group (1=English, 2=Persian)		Percentage Correct
			1	2	
Step 1	L1 Group (1=English, 2=Persian)	1	23	2	92.0
		2	2	23	92.0
	Overall Percentage				92.0

a. The cut value is .500

The Classification Table (Table 6-7 above) demonstrates how many authors are assigned to each group according to the features they exhibit (predicted), and how this contrasts with which group they actually belong to (observed). This table shows that out of the total of 50 authors, only 4 were incorrectly attributed. Each author (or case) is assigned a predicted group purely from the features contained within the data for that author; no reference is paid to the distribution of other authors. In order to better understand which cases were misattributed we can examine the Casewise List. Table 6-8 below is a modified version showing only the values for the misattributed cases.

Table 6-8 - Study 1 - Optimum Model Casewise List

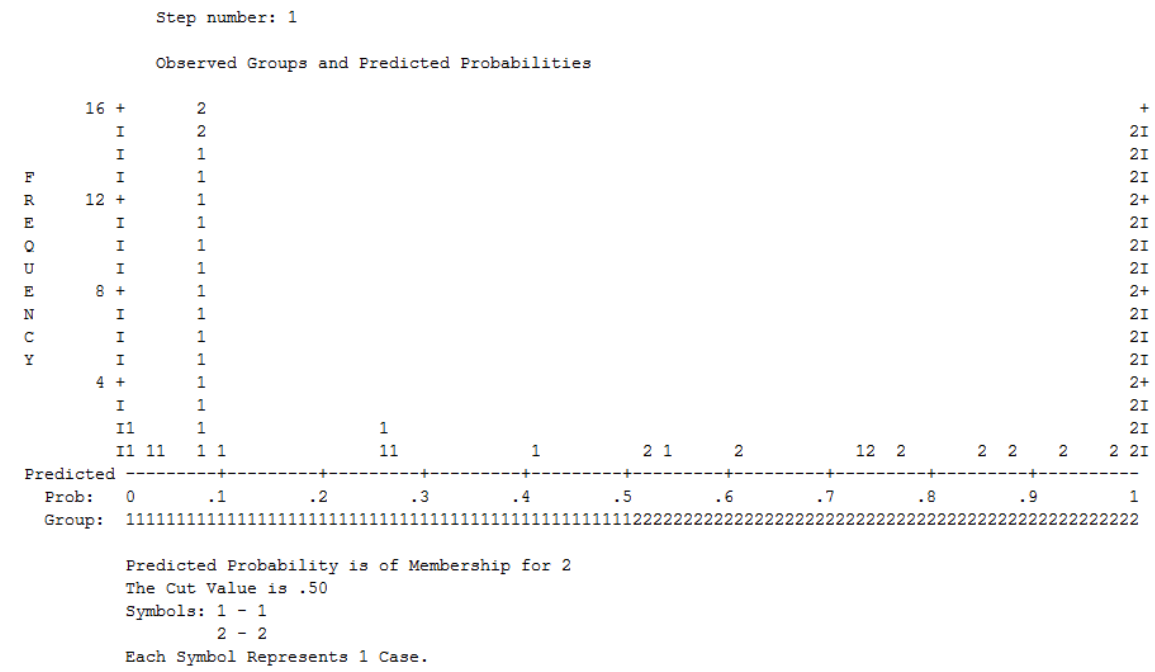
Casewise List

Case	Selected Status[a]	Observed	Predicted	Predicted Group	Temporary Variable	
		L1 Group (1=English, 2=Persian)			Resid	ZResid
11	S	1**	.532	2	-.532	-1.066
15	S	1**	.726	2	-.726	-1.628
33	S	2**	.073	1	.927	3.576
40	S	2**	.073	1	.927	3.576

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

A predicted probability over 0.5 (where the score is between 0 and 1) results in the author being assigned to Group 2, the group of L1 Persian speakers. A score of less than 0.5 results in the author being assigned to Group 1, the group of L1 English speakers. The output also represents this graphically (see Figure 6-2 below).

Figure 6-2 - Study 1 - OptimumModel Observed and Predicted Probabilities



Of the two L1 English speakers that were attributed as being L1 Persian, their probabilities are both close to the 0.5 threshold. Case 11 (which represents author *K: Jayne*) had a probability of 0.532, which is only 0.032 over the cut off point and Case 15 (which represents author *O: Kitty*) has a probability of 0.726 which is 0.226 over the 0.5 boundary. The two cases are within the bottom 4 cases that are predicted as belonging to Group 2 (L1 Persian speakers). Conversely the two L1 Persian speakers that were predicted as belonging to the L1 English group (Group 1) had probabilities that placed them firmly within that group, and were not close to the 0.5 boundary line. Both Case 33 (which represents author *AG:*

Ihateithere5150) and Case 40 (which represents author *AN: maryam owji*) have the same predicted probability of 0.073, which is -0.427 away from the cut off point of 0.5. The reasons for this are unclear at this stage, but are discussed further in Section 6.4 of this chapter.

6.4 Discussion

The findings above clearly indicate that the features and NLID methodology set out are able to discriminate between L1 Persian authors and L1 English blog authors. Using the optimum model, only four authors were attributed to the wrong groups. Table 6-9 below shows the probabilities for the wrongly attributed authors, along with the linguistic backgrounds that they stated in their blogs.

Table 6-9 - Study 1 - Optimum Model Case Probabilities and Linguistic Histories

Case	Author's name	Observed	Predicted	Predicted Group	Linguistic History given
		L1 Group (1=English, 2=Persian)			
11	*K: Jayne*	1**	.532	2	I am a native English speaker
15	*O: Kitty*	1**	.726	2	Thank you. Yes, my mother tongue is English
33	*AG: Ihateithere5150*	2**	.073	1	I am a native of Iran. I speak Farsi and English
40	*AN: maryam owji*	2**	.073	1	we have this Persian saying

Both the L1 Persian speakers demonstrate a near native level of competency of L2 English in their blogs, however, the L1 English speakers demonstrate no indications in the content of their blogs of having any Persian influence. This could explain why they are only just inside the boundary line of predicted group 2 membership, when the L1 Persian authors are considerably further away from the boundary probability. It is possible that the features they are exhibiting could be from influences from other languages that they have not mentioned. This possibility is no more than conjecture, due to a lack of background information. A better understanding of how different languages might influence the linguistic features may be of benefit.

Contained within the output there is more data that we have not used to its full potential. Most significantly the table *Variables in the Equation* (see Table 6-10 below) contains

information about each feature that we can use in a casework setting, where running a full analysis replicating this study might not be appropriate (this will be discussed further in Chapter 9 sections 2 and 3). In the common scenario in which an analyst is given a text and asked to provide profile information, if any of the 10 features are found in the text, then the analyst can use the values provided in Table 6-10 below to extrapolate more information about the profile. The *B* value indicates whether a feature is more indicative of membership to group 1 or group 2. A positive *B* value means the greater number of occurrences of that feature, the higher the probability of the author belonging to group 2, which in this case is the L1 Persian group. The *Exp(B)* value is the likelihood ratio which we can use in conjunction with any probabilities from prior information (such as a name which is three times more likely to belong to a Persian native speaker). Using the *Exp(B)* value, it is possible for an analyst to use these values in isolation without repeating the entire analysis. This is discussed further in Chapter 9 Section 3. The *B* value indicates whether a feature is more indicative of membership of group 1 or group 2. A positive *B* value means the greater number of occurrences of that feature, the higher the probability of the author belonging to group 2, which in this case is the L1 Persian group.

Table 6-10 Study 1. 10 Features. Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	@122LexicalMarkedPresence	26.623	10080.005	.000	1	.998	365120009131.117
	@66ConjunctionMarkedAbsence	-53.241	47098.245	.000	1	.999	.000
	@22AdverbMarkedPresence	-74.842	27316.667	.000	1	.998	.000
	@251VerbalMarkedChoice	1.727	.853	4.101	1	.043	5.626
	@102LexicalMarkedAbsence	1.355	1.467	.853	1	.356	3.877
	@265VerbalMarkedConstruction	-1.058	.829	1.629	1	.202	.347
	@200PronounMarkedChoice	80.921	15731.947	.000	1	.996	139119092697538060000000000000000000000.000
	@208PronounMarkedPresence	-16.168	2989.332	.000	1	.996	.000
	@33ArticleMarkedChoice	2.628	1.435	3.355	1	.067	13.844
	@45ArticleMarkedPresence	1.161	1.266	.840	1	.359	3.192
	Constant	-2.548	.848	9.034	1	.003	.078

a. Variable(s) entered on step 1: @122LexicalMarkedPresence, @66ConjunctionMarkedAbsence, @22AdverbMarkedPresence, @251VerbalMarkedChoice, @102LexicalMarkedAbsence, @265VerbalMarkedConstruction, @200PronounMarkedChoice, @208PronounMarkedPresence, @33ArticleMarkedChoice, @45ArticleMarkedPresence.

This study shows that the features identified are able to distinguish between L1 Persian authors writing in English, and L1 English authors. The initial statistical analyses indicate that the models containing all the higher level features, or all the lower levels, may be too

specific, and therefore not generalisable to other data. Through eliminating features with less predictive power, the optimum model was determined to contain the following ten features: *122LexicalMarkedPresence*, *66ConjunctionMarkedAbsence*, *22AdverbMarkedPresence, 251VerbalMarkedChoice, 102LexicalMarkedAbsence, 265VerbalMarkedConstruction*, *200PronounMarkedChoice, 208PronounMarkedPresence, 33ArticleMarkedChoice,* and *45ArticleMarkedPresence*. It is predicted that using the full range of features in the following studies will also lead to an over fitting of the models. It is likely that the most significant features in this study will vary compared to the most significant features in the following studies, however, there may be some overlap.

This study has shown that at the basic level it is possible to use features in the language of online blogs to distinguish between L1 Persian speakers blogging and L1 English speakers blogging. Many of the features can be attributed to likely L1 Persian influence. It is still possible however that what we are actually identifying through the analysis set out, is actually that a speaker is using English as their L2, rather than an indication that there may be an influence from Persian as an L1. In order to test exactly what the analysis is showing, we need to compare L1 English blog posts from authors with differing L1's. In an ideal situation, and in order to allow for all possibilities that may be encountered in a forensic situation, we should analyse L2 English blog posts by authors with as many differing L1's as possible. Clearly there would be severe logistical issues with this at a data management level, before we even consider the academic implications.

Languages are commonly arranged into linguistic trees. These indicate language families according to how the languages have evolved. However, due to the nature of language evolution, languages which are close in the linguistic tree are often very similar in their structure, grammar and lexis. This similarity could mean that rather than identifying an L1 Persian influence, we are actually indicating an L1 influence from a language in the Proto-Iranian language family. Study Two looks at related languages to test the extent to which these features can distinguish between languages.

This study does not allow for the possibility that an author might manipulate their language, especially in a forensic context, to give the impression that they are a native Persian speaker in order to cast suspicion onto another person. This will be explored in Study Three. In order to understand the limitations and implications of this analysis, we need to look at equivalent corpora where the L1 of the authors is different from, yet linguistically close to, Persian. This is undertaken in the next Study Two (Chapter 7).

Part Three – Applications

The chapters within this section of the thesis move beyond the initial research and consider the practical applications. This section is divided into three chapters. The first two are mini-studies (Study Two and Study Three respectively) which address some of the practical issues surrounding the application of NLID in a forensic context. The third chapter brings together all the areas of this study, discussing and evaluating the potential role of linguistic features and NLID as a tool for forensic authorship analysis. It should be noted that each sub-study here could be expanded considerably, to become its own research project, however, due to the time and size restrictions that are integral to any research project, these studies are relatively small in terms of data and analysis. They can be considered as exploratory studies which give indications as to potential future research, and to give a deeper contextual understanding of Study One (Chapters 5 and 6).

Chapter 7. Study Two – Other Languages

> ***The day we stop exploring is the day we commit ourselves to live in a stagnant world, devoid of curiosity, empty of dreams – (Tyson, 2012)***

- Chapter 7 - Study Two – Other languages
  - 7.1 Language selection
  - 7.2 Methodology and literature
  - 7.3 Findings
  - 7.4 Findings and Discussion

The aim of this chapter is to test whether we can use the NLID method and the linguistic features discussed so far to distinguish a native Persian speaker writing in English from speakers of other related languages. The underlying premise for this research is that there are distinct differences between the influences exerted by different L1 languages. The first section considers the context and motivation for analysing other languages, as well as considering language selection. The second section details the methodology for this study. The third section sets out the findings and the fourth section considers the implications of the findings.

7.1 Motivation, Plan, and Language Selection

Study One demonstrated that there is a clear distinction in the features that occur in the language of L1 Persian speakers, as opposed to the language of L1 English speakers. This took a bottom-up approach, with the features being identified solely through their appearance in the data, as opposed to examining features that are discussed in existing literature as they occur within the data. Little attention was paid, therefore, to the linguistic theory on the influences responsible for the features. In part this is due to the lack of consensus on exactly what affects the language produced by a non-native speaker (as discussed in Chapter 2). Study One has identified features that are unique to the language of L1 Persian speakers as opposed to L1 English speakers, but we do not know if these features can be attributed to universals in L2 English production, such as influence from learning strategies. Therefore, it is difficult to be sure if the features are indicative of authorship by L1 Persian speakers as opposed to indicating authorship by a non-native English speaker, regardless of the L1. In order to evaluate whether the features are L1 specific, we need to compare the distribution of the features in L1 Persian texts to the distribution in comparable corpora of English blogs by authors with different L1s.

The intended practical casework applications of this research mean that it would be more beneficial to investigate languages that are found within the same region and populations as Persian, rather than completely unrelated languages. Although Persian (Farsi) is the official language of Iran, there are also numerous minority languages that are widely spoken throughout the republic. These include; Baluchi, Kurdish, Azeri and Pashto. Dari and Tajik are not viable options as they are dialects of Persian, exhibiting similar lexis and grammar to the standard Tehran version. It is therefore unlikely that they would produce different features. Due to practical constraints it was decided that two languages would be a suitable number

to focus on for this exploratory study. The two languages chosen; Azeri and Pashto, had a significant proportion of their speakers in the same area as Persian, and would yield enough data for this study.

Pashto is the language of the Pashtun or Pakhtun people and an official language of Afghanistan, the other official language of Afghanistan being a variety of Persian called Dari. 52.3% of the Afghani population is Pashtun (Rahman, 1995) Pashto is also known as Afghani, or Pushto. It has approximately 50 million native speakers, is one of the largest Iranian languages (Rashidvash, 2012) and is spoken in Afghanistan and Pakistan as well as among the Pashtun diaspora. It is belongs to the Eastern Iranian language family, and can be linked back to Avestan. It is written using the Pashto Alphabet, which is a modified version of the Persian alphabet. There are two main dialects, which roughly correspond to the north and the south, but the differences are predominantly phonological.

Figure 7-1 - Iranian Languages Map

(Gippert, 1993-2010)

Azeri is a Turkic language, also known as Azerbaijani. It is an official language in both Azerbaijan and Russia. It is spoken predominantly in Azerbaijan and the North of Iran but extends to other regions too. The Azerbaijani have a mixed ethic background, with a heavy Turkic and Persian influence. Over history, there have been considerable population shifts, most recently due to conflict in Armenia. According to Bermel, Azeri "is spoken more widely outside Azerbaijan than inside it" (Bermel, 2006, p. 14) Exact figures are unknown, but unofficial estimates place the population of Azerbaijanis in Iran between 16-33% (Rashidvash, 2012): Bermel (2006) estimates that there are 23 million speakers in Iran of Southern Azeri. They are considered an established and well integrated linguistic minority (Rashidvash, 2012). Azerbaijanis are frequently bilingual, with Russian being the predominant second language in Azerbaijan, and Persian in Iran. Azeri is the most widely

spoken minority language in Iran, being understood or spoken by approximately 20% of the Iranian population (Rashidvash, 2012). Azerbaijani belongs to the southwestern branch of the Turkic languages and is mutually intelligible with Turkish[12]. Figure 7-2 shows the approximate distribution of the Azeri language.

Figure 7-2 - Turkic Languages Map

(Gippert, 1993-2010)

Both Azeri and Pashto have close links to Persian, yet are also distinct languages from Persian. They both have a relatively significant online blog presence (though considerably less so than Persian) making them ideal for this study. The close links between Azeri and Persian and Pashto and Persian, mean that focusing on these languages has more practical benefits than choosing languages without these links. In a casework scenario, it is feasible that either Azeri or Pashto may play a role in questioned authorship.

[12] It should be noted that Azeri and Azari are two different languages. Azari is an ancient language that existed in the area around Azerbaijan. Despite the similarity in geographical location, the two languages belong to different language family branches. Azeri is a Turkic language, and Azari is an Iranian language.

7.2 Analysis / Methodology and literature

This section outlines the methodology used in this study. The aim was to keep the methodology consistent with Study One across data collection and coding. The texts were collected using the same procedures as in Study One, these are detailed in Chapter 4. The authors were selected according to what they identified as being their native language, with care being taken at selection level that they did not contradict themselves. Due to practical limitations on this research, and the languages being less widely spoken, it was more difficult to collect an equally large volume of text as in Study One. Therefore, fewer authors were selected (5 for each language), also less text was collected for each author (and it was frequently the case that less was available).

The aim was to collect at least 200 words from each author, but in some cases it was possible to collect considerably more. Where possible, up to 2,000 words was collected as the feature counts were later normalised to account for the difference in word counts, both across and within the corpora. The word counts were normalised to show how many features there would be per 2,000 words. Therefore the number of items coded at each feature was divided by the total number of words collected for that particular author, then multiplied by 2,000.

The ethical considerations for this study were the same as they had been in Study One, as the data collection and methodology was consistent with the previous study. Greater discussion of the considerations can be seen in Chapter 4 Section 4. As all of the texts were publicly accessible at the time of collection, informed consent was deemed unnecessary. All data were handled in conjunction with Aston University's guidelines (Aston University Ethics Committee, 2007). The ethical considerations application as submitted to the Ethics Committee for the School of Languages and Social Sciences at Aston University can be seen in Appendix A. The strategy for analysis for this Study followed the structure of Study One. After an initial read through, the texts were coded using NVivo according to the coding structure devised previously (and discussed in Chapter 5 section 2). The coding matrix was then exported to see the distribution of features for each author. This can be seen in Appendix H which shows the amalgamated coding matrices for all the studies. Once the coding matrix was exported and normalised (Appendix I), then the statistical analyses could be performed using SPSS. Study One used binomial logistic regression to determine how accurate the features were at indicating membership to either the L1 Persian group or the L1 English group. Binomial logistic regression was the clear choice for Study One as the authors belonged to two distinct corpora. As the same questions are being asked of the data, it is

clear that logistic regression remains the most appropriate statistical technique. However, as Study Two looks at two other languages as well as Persian, it could be argued that there are three groups and that multinomial logistic regression would therefore be more appropriate than binomial. It is important to remember at this stage though, that the Azeri and the Pashto are not complete separate corpora, instead they are much smaller than the L1 Persian corpus. The features identified within the texts cannot be considered to be completely representative of L1 speakers of either language. Instead the two mini-corpora together form a sub-corpus of L1 speakers of 'other languages'. Binomial logistic regression is therefore again the most applicable analysis, as we have two discrete potential outcomes; membership of the L1 Persian speaker group, or membership of the 'other languages group' as either an L1 Azeri or Pashto speaker. At this level of analysis it is unrealistic to be able to separate the L1 Azeri speaker and the L1 Pashto, the size and amount of data would also severely limit any potential findings.

For clarity the group that contains the L1 Azeri and the L1 Pashto speakers is referred to throughout the analysis as *otherlanguages*. This reflects the concept of the study, which is comparing L1 Persian speaker to L1 speakers of other languages, specifically Azeri and Pashto. It is not intended to imply that all languages other than Persian or English would result in comparable results. During the data collection stages and in the coding matrix Group 3 specifically refers to the Azeri authors and Group 4 the Pashto authors. However, during the statistical analyses the distinction between these two corpora was discounted (as explained above); therefore in order to enable the analyses in SPSS, both sub-corpora were labelled as Group 3. The original information is retained and referred to for greater context and understanding. The progression of the statistical analysis mirrors Study One, and will be discussed in conjunction with the findings and results in the following section, Section 7.3.

7.3 Findings

This section sets out the main findings of Study Two. The methodology for analysis was kept consistent with the previous study, with the stages for finding the optimum model almost exactly mirroring Study One. First the analysis looks at the higher level features, then all the linguistic lower level features. The output from this second analysis is then used to rank the features according to their predictive ability, and then the model is refined by reducing the features to create the optimum model.

7.3.1 Higher level features

As in the previous study, performing the binary logistic regression with just the higher level features can give an interesting preliminary insight to the data. Unlike in the previous study the higher features alone, do not constitute an overly fitted model for this data. Using just the higher level features the model correctly assigns thirty one authors to the right group and misattributes four; a correct percentage of 88.6 percent.

Table 7-1 - Study 2 - Higher Level Features Classification Table

	Observed		Predicted		
			Study2otherlanguages		Percentage Correct
			2 – L1 Persian	3 – Other languages	
Step 1	Study2otherlanguages	2 – L1 Persian	24	1	96.0
		3 – Other languages	3	7	70.0
	Overall Percentage				88.6

a. The cut value is .500

The Hosmer-Lemeshow statistics demonstrate that the model using only the higher level features has a relatively high significance rating, with a Chi-square score of 4.103 and a significance of 0.768. The features with the highest Wald scores are *Capitalisation* and *Punctuation* with Wald scores of 4.063 and 3.965 respectively. These are followed by *Article* (3.105) and *Orderingandpositioning* (3.030). It is interesting to note that the two most significant higher features are the ones that could be considered more orthographic than linguistic; however, at this stage of the analysis, it is very difficult to infer any conclusions from this.

7.3.2 All lower level feats

The next stage of analysis was to consider the lower level linguistic features. These were selected exactly as in the previous study, resulting in 29 features which could be considered the lower branches of the coding tree. *Capitalisation* and *punctuation* were left at the higher level nodes, as they are more orthographic in nature than linguistic, and this stage of the research is concerned predominantly with the linguistics elements. As in Study One, it was predicted that using all 29 features in the binomial logistic regression, will create a model that is over fitted to the data. This was indeed the case, as all 25 L1 Persian speakers were correctly predicted as belonging to group 2, and all 10 L1 other language authors were correctly placed in group 3. The Hosmer-Lemeshow test (see Table 7-2 below) demonstrates the extent of overfitting, with a chi-square score of less than 0.000 and a significance approaching 1.000 it is apparent that the model contains too many features.

Table 7-2 - Study 2. All lower level features Hosmer-Lemeshow test

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	7	1.000

Before reducing the number of features, it is important to consider which features are the most significant when determining group membership. This is done by ranking the features according to the Wald Chi square score (as in Table 7-3 below).

Table 7-3 - Study 2. All linguistic lower level features Wald scores

Rank	Features	B	Wald	Exp(B)
1	@208PronounMarkedPresence	69.986	.0000104182598584	2.48045E+30
2	@251VerbalMarkedChoice	20.300	.0000074843370005	654937714.5
3	@106LexicalMarkedChoice	-11.931	.0000053197066646	6.58372E-06
4	@66ConjunctionMarkedAbsence	154.893	.0000027557604828	1.85766E+67
5	@80ConjunctionMarkedPresence	-116.357	.0000027229471839	2.92815E-51
6	@122LexicalMarkedPresence	50.206	.0000027190804504	6.37234E+21
7	@162PrepositionMarkedAbsence	-41.431	.0000025267314046	1.01577E-18
8	@102LexicalMarkedAbsence	60.291	.0000022237464460	1.52754E+26
9	@195PronounMarkedAbsence	14.089	.0000019857840480	1314317.981
10	@12AdverbMarkedAbsence	-382.966	.0000019161689331	4.7852E-167
11	@117LexicalMarkedConstruction	-2.998	.0000017776650539	0.049889267
12	@265VerbalMarkedConstruction	-25.022	.0000016443524822	1.35851E-11
13	@200PronounMarkedChoice	-60.326	.0000010583727959	6.31869E-27
14	@50Capitalisation	2.441	.0000008330193188	11.490118667
15	@181PrepositionMarkedPresence	38.953	.0000008150491648	8.26537E+16
16	@45ArticleMarkedPresence	14.989	.0000007983676795	3233139.931
17	@213Punctuation	-8.941	.0000005275522449	0.000130947
18	@131Other	-.318	.0000003630560365	0.727846537
19	@127Orderingorpositioning	-42.154	.0000003617185463	4.92949E-19
20	@22AdverbMarkedPresence	71.062	.0000002792669744	7.27612E+30
21	@247VerbalMarkedAbsence	360.896	.0000002484006036	5.4316E+156
22	@70ConjunctionMarkedChoice	53.529	.0000002410880995	1.76789E+23
23	@16AdverbMarkedChoice	54.905	.0000002129086945	6.99413E+23

Rank	Features	B	Wald	Exp(B)
24	@75ConjunctionMarkedOrderingorpositioning	1.483	.0000001821382249	4.405392223
25	@173PrepositionMarkedChoice	4.690	.0000001369070264	108.862084
26	@2AdjectiveMarkedChoice	20.939	.0000001017866442	1240793396
	Constant	-10.475	.0000000701118105	2.82315E-05
27	@27ArticleMarkedAbsence	2.479	.0000000679681788	11.9325587
28	@7AdjectiveMarkedPresence	-17.478	.0000000091499904	2.56753E-08
29	@33ArticleMarkedChoice	-3.908	.0000000064386251	0.020071499

The *B* score enables us to determine the direction of the feature, i.e. whether the presence of that particular feature increases the probability of membership of Group 2 (Persian L1 speakers) or of Group 3 (L1 other languages). A positive score means that a greater number of occurrences of that feature increases the probability the author is not an L1 Persian speaker, but belongs to Group 3.

Table 7.3 indicates that L1 Persian speakers' writings are more likely to contain *80conjunctionMarkedPresence* and *251VerbalMarkedChoice*, whereas *106LexicalMarkedChoice* is more indicative of membership of the L1 Azeri and Pashto (other languages) group. It would be interesting to compare this to existing Interlanguage research and comparative linguistics. Although this research is intended to mainly have implications for forensic authorship analysis, findings such as these might be of benefits to other areas of sociolinguistic research, bridging the gaps.

The Wald scores are larger than the Wald scores in the corresponding table for Study One (see Table 6-4), however they are still very small. This indicates that although the 29 features have a high predictive value when combined, each feature alone does not have a high predictive ability. As in Study One, the number of features can be reduced, removing the features with the lower predictive power first until the goodness-of-fit results for the model reflect that the model is no longer over-fitted, yet remains as accurate as possible.

When reducing the number of features to find the optimum model, it was found that using the 12 highest features with the highest Wald-scores produced the best fit. These features are: @208PronounMarkedPresence, @251VerbalMarkedChoice, 106LexicalMarkedChoice, 66ConjunctionMarkedAbsence, 80ConjunctionMarkedPresence, 122LexicalMarkedPresence, 162PrepositionMarkedAbsence, 102LexicalMarkedAbsence, 195PronounMarkedAbsence, 12AdverbMarkedAbsence, 117LexicalMarkedConstruction, and 265VerbalMarkedConstruction. Using these features in the model resulted in a significance

of 0.922 and a Chi-square score of 2.571 in the Hosmer-Lemeshow Test (see Table 7-5 - Study 2. 12 Features Model Summary. below). A significance score that is close to 1, without reaching it, means that the model is accurate without being over-fitted.

Table 7-4 - Study 2 - Lower Level Features Hosmer Lemeshow Test

Step	Chi-square	df	Sig.
1	2.571	7	.922

The Model Summary containing the Nagelkerke R square and the Cox and Snell R square scores (see Table 7-5 below) indicate that the model using the twelve features accounts for between 44.5 and 63.8 percent of the variability.

Table 7-5 - Study 2. 12 Features Model Summary.

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	21.258[a]	.445	.638

The classification Table (below) for this model gives a more visual representation of the predicted memberships according to the features contained in each text.

Table 7-6 Study 2. 12 Features Classification Table

Classification Table[a]

	Observed		Predicted		
			Study2otherlanguages		Percentage Correct
			2.00	3.00	
Step 1	Study2otherlanguages	2.00	23	2	92.0
		3.00	4	6	60.0
	Overall Percentage				82.9

a. The cut value is .500

Of the Cases (each case reflects an author) that were wrongly attributed using this model, we can see that two L1 Persian speakers were miss-assigned to the Other languages group, and 4 authors from the Otherlanguages group were falsely assigned to the L1 Persian group. Table 7-7 below demonstrates that the wrongly assigned texts all had predicted probabilities that were relatively close to the 0.5 boundary. Texts that had a predicted probability of over 0.5 were assigned to group 3, the group of other language speakers, and a predicted probability of less than 0.5 meant the author was indicated as belonging to the L1 Persian speaker group (Group 2).

Table 7-7 Study 2. 12 Features Misattributed Authors Casewise list.

Casewise List

Case	Author	Observed	Predicted	Predicted Group	Temporary Variable	
		Study2 otherlanguages			Resid	ZResid
34	AH : inside Iran	2**	.662	3	-.662	-1.400
36	AJ : Katayoun	2**	.510	3	-.510	-1.019
53	BA : Farida (Azeri)	3**	.485	2	.515	1.030
54	BB : Giridhar (Azeri)	3**	.347	2	.653	1.371
56	BD : Abudlhadi (Pashto)	3**	.110	2	.890	2.846
57	BE : FaerieBoy (Pashto)	3**	.276	2	.724	1.618
** = Misclassified cases.						

It is interesting to note that of the four authors that were falsely attributed to the L1 Persian authors group, two were L1 Azeri speakers and two were L1 Pashto speakers. This even split could indicate that despite Pashto being linguistically closer to Persian than Azeri, this does not result in a significantly greater chance of misattribution at this level. However it is important to note that this is only a preliminary study and further research using more data would be likely to uncover discrepancies. Despite the fact that the difference is not so great as to be observable within this study, the probability values for the wrongly assigned Pashto authors are lower than the probability values for the two Azeri authors. The lower probability values means that they are further away from the 0.5 boundary line, and hence have a great statistical probability of belonging to the L1 Persian group. One could hypothesise that this indicates an L1 Pashto speaker is more likely to be misattributed as an L1 Persian speaker than an Azeri speaker, however, this is only a preliminary hypothesis and conclusions cannot be drawn without further research involving a greater volume of data from L1 Azeri and L1 Pashto authors.

7.4 Findings and Discussion

This study indicates that it is possible to distinguish between authorship by an L1 Persian speaker, and speakers of co-existing languages; namely Azeri and Pashto. Pashto is an Iranian language like Persian, and Azeri is a Turkic language, however, the distribution of their speakers overlaps with the distribution of Persian speakers.

The optimum model comprised twelve features; @208PronounMarkedPresence, @251VerbalMarkedChoice, 106LexicalMarkedChoice, 66ConjunctionMarkedAbsence, 80ConjunctionMarkedPresence, 122LexicalMarkedPresence,

162PrepositionMarkedAbsence, 102LexicalMarkedAbsence, 195PronounMarkedAbsence, 12AdverbMarkedAbsence, 117LexicalMarkedConstruction, and 265VerbalMarkedConstruction. Of these, six were also incorporated in the optimum model for Study One, these are marked below in Table 7-8 in green.

Table 7-8 - Study 1 and 2 - Optimum Model Features

Rank	Study One	Wald	B	Study Two	Wald	B
1	@251VerbalMarkedChoice	4.101	1.727	@66ConjunctionMarkedAbsence	3.535	4.364
2	@33ArticleMarkedChoice	3.355	2.628	@208PronounMarkedPresence	2.463	6.313
3	@265VerbalMarkedConstruction	1.629	-1.058	@162PrepositionMarkedAbsence	2.134	-1.632
4	@102LexicalMarkedAbsence	0.853	1.355	@106LexicalMarkedChoice	2.109	0.242
5	@45ArticleMarkedPresence	0.84	1.161	@102LexicalMarkedAbsence	0.852	-2.52
6	@122LexicalMarkedPresence	0	26.623	@117LexicalMarkedConstruction	0.83	0.103
7	@66ConjunctionMarkedAbsence	0	-53.241	@122LexicalMarkedPresence	0.657	-1.691
8	@22AdverbMarkedPresence	0	-74.842	@265VerbalMarkedConstruction	0.385	-0.383
9	@200PronounMarkedChoice	0	80.921	@195PronounMarkedAbsence	0.352	0.385
10	@208PronounMarkedPresence	0	-16.168	@251VerbalMarkedChoice	0.053	0.073
11				@80ConjunctionMarkedPresence	0	-40.801
12				@12AdverbMarkedAbsence	0	-26.574

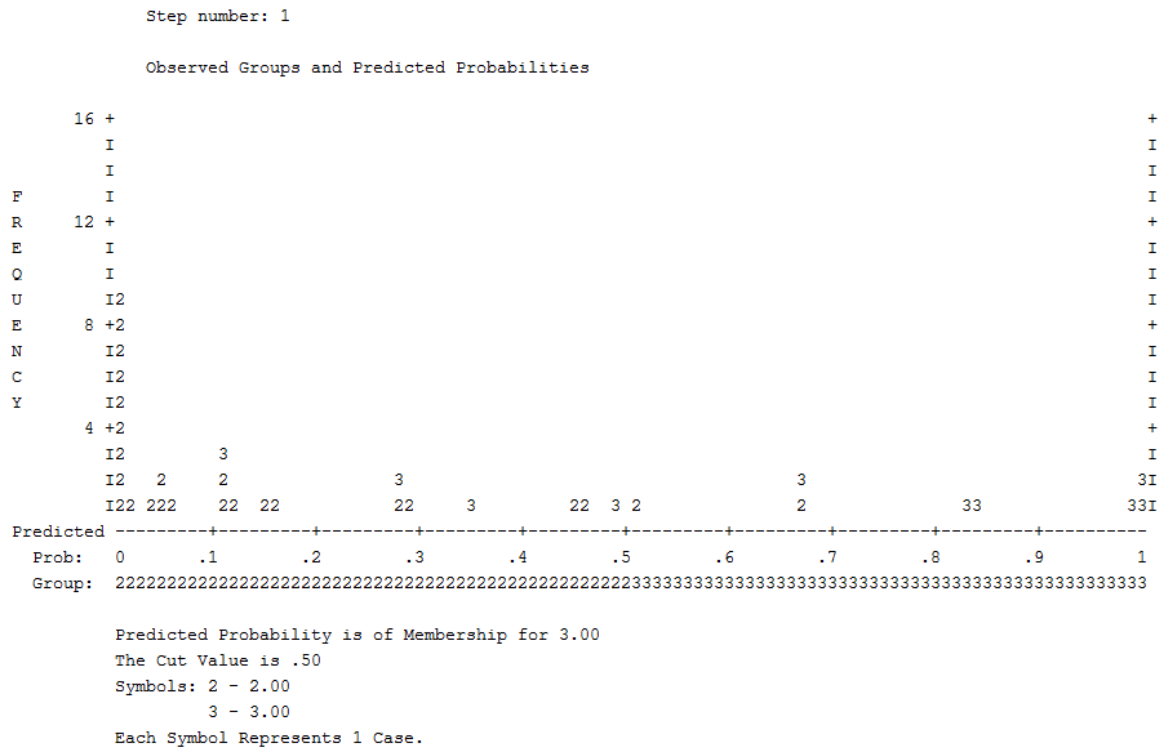
Study One demonstrated that it was possible to distinguish between L1 English and L1 Persian authors writing in English. However, one of the limitations was that it was impossible to tell if the features and NLID system designed were just indicating the difference between native and non-native speakers, or whether NLID was able to indicate a particular language. This study demonstrates that the system can distinguish between languages, and even between languages from the same branch of a linguistic family tree; i.e. between two Iranian languages.

The polarity of the B value indicates whether the presence of a feature increases the probability that an author belongs to either the first or second group. The analysis was such in SPSS, that in Study One a positive B value increased the probability that an author

belonged to the L1 Persian group (rather than L1 English), but in Study Two a negative B value increased the probability the author was an L1 Persian speaker (rather than belonging to the otherlanguages group). In each Study the following 3 features indicated an increased probability that the author was an L1 Persian speaker: @66ConjunctionMarkedAbsence, @208PronounMarkedPresence, and @122LexicalMarkedPresence.

Of the authors that were assigned to the wrong groups using the optimised model, four texts were falsely attributed as being L1 Persian, two of which were L1 Azeri and two L1 Pashto. It is interesting to note though that the Pashto authors had group membership probabilities that put them closer to the L1 Persian boundary line than the L1 Azeri authors (see Figure 7-3 below). Two L1 Persian authors were misattributed to the *otherlanguages* group. It is possible that these two authors had influence from either Azeri or Pashto, especially due to the geographical locations of the languages; however, this level of information is not available within the collected data. The prevalence of Persian in the areas that Azeri and Pashto are commonly spoken would imply that many L1 Azeri and L1 Pashto speakers will have an influence from L1 Persian. It is therefore not surprising that more authors from the otherlanguages Group were attributed as L1 Persian speakers, than the other way around.

Figure 7-3 - Study 2 - Optimum Model Predicted Probabilities



The main limitation of this study is that it is only an exploratory study into NLID, with reference to Azeri and Pashto. Part One of this thesis set out the potential need for further study into NLID, although the focus was primarily on L1 Persian speakers, the principle remains the same for speakers of all languages. Due to the limited amount of data analysed, it is very difficult to draw conclusions about potential features for these specific languages. However, findings of this study do indicate that Azeri and Pashto exhibit distinct interlingual features compared to Persian, and it would be beneficial to expand this sub-study into two full research projects; one investigating the interlingual features of L1 Azeri speakers writing in English, and one focusing on L1 Pashto speakers.

This study also suffers the same limitation as Study One in that is does not consider the possibility of authors disguising their language (or attempting to). The next study, Study Three, will consider whether the NLID features and system devised so far can distinguish between genuine L1 Persian authors and authors disguising their language.

Chapter 8. Study Three – Disguise Data

> ***Le vrai moyen d'être trompé, c'est de se croire plus fin que les autres.***
>
> *The truest way to be deceived is to think yourself more knowing than others.*
>
> (De la Rouchefoucauld, 1665-1678)

- Chapter 8 – Study Three – Disguise data
  - 8.1 Forensic context and casework motivation
  - 8.2 Methodology and Analysis
  - 8.3 Findings
  - 8.4 Discussion

This chapter details the third and final sub-study of this research, Study Three. This study looks at disguise data, and questions whether the features identified so far can distinguish an L1 Persian speaker writing in English from a native speaker of another language who is disguising their language to give the appearance of being an L1 Persian speaker writing in English. The first section considers the forensic context and motivation for analysing disguise data. The second section details the methodology and the analysis that has been carried out. The findings are set out in the third section, and the fourth section considers the implications of these findings, particularly with reference to the wider research project.

8.1 Forensic context and casework motivation

Shuy (2001) discussed a series of threat notes which were written in different handwriting styles, in an apparent attempt to disguise the author's identity. Shuy noted that the features exhibited indicated that the author was likely to have a Hindi-Urdu influence acting on the English he produced, yet when disguising the language he had not attempted to alter the features that are of most interest to forensic authorship analysis. Shuy concluded that when "anonymous writers attempt to disguise their prose style, such effort usually involves the more conscious aspect of language use [in this case handwriting] rather than the major features analyzed in the linguistic profile." (Shuy, 2001, Loc. 10105). This has very positive implications for forensic authorship analysis, however, it should not be considered a static situation. Forensic linguistics is gaining more and more media attention, both in the factual and fictional realms. As television shows (such as Criminal Minds) and novels (such as

Danuta Reah's *Night Angels*, Kathy Reich's *Bones to Ashes*, Kylie Brant's *Deadly Intent*, or Jeffery Deaver's *Devil's Teardrop*) introduce forensic authorship analysis to a wider audience, we must consider the possibility of this impacting on casework data. Put plainly, authors of potential forensic texts may become aware of the importance of linguistic features, and therefore consciously seek to alter these to some extent.

In the *Devil's Teardrop* a forensic document analyst performs a sort of forensic authorship analysis on a threat note. The FBI agent highlights some marked phrases, suggesting that they indicate the author is a non-native English speaker because; "that's how foreigners talk," (Deaver, 1999, p. 93). He is then corrected by the expert, who notes that some features are indicative of Slavic or Germanic languages, whereas another feature is more likely to be from influence from an Asian language; which he concludes indicates that the author "'just threw in random foreign-sounding phrases. Trying to fool us into thinking he's foreign. To lead us off." (Deaver, 1999, p. 93). Clearly this is a fictional example created by a novelist, however, it should be noted that most forensic texts are not written by linguistic experts. There are numerous documented cases of authors trying to disguise their language. The Devil-strip case discussed in Section 3.2 contained deliberate misspellings and slightly marked grammatical constructions, which Shuy interpreted as those of an educated man who was attempting to disguise his language to appear uneducated (Leonard, 2005). The Lindbergh kidnapping as discussed in Section 1.2 also contained elements that related to disguise. There were differences between the ransom notes and texts that were later elicited from Hauptmann (the native German speaker who was found guilty and executed for the kidnapping, but the conviction is considerably disputed). These differences were attributed to attempts by Hauptmann to disguise his writing style and hence his identity (Solan & Tiersma, 2005). This indicates that an understanding of disguise language is needed; not just to uncover people attempting to disguise their L1, but also so that features are not attributed to being attempts at disguise, when there might be a different influence or cause present.

Kniffka (2000) documented a case he worked on that involved the possibility of someone disguising their language to alter perceptions of their L1 influence in an extortion letter that contained blackmail and death threats. Law enforcement suspected a "naturalized businessman of Hungarian descent" (Kniffka, 2000, p. 180) and approached Kniffka with the data asking whether the speaker was an L1 German speaker or not and if there is "any evidence that the anonymous author disguises his linguistic identity, and if so with what

probability?" (Kniffka, 2000, p. 181). His analysis uncovered features that could be indicative of an L1 German speaker disguising his command of German, however, he determined that some of his analysis indicated authorship by an L1 Hungarian speaker writing in German who was not making a particular effort to disguise their language. This is pertinent to this research, as it raises the issue of being able to distinguish disguised language from genuine Interlanguage, and the degree of certainty to which this can be done.

It is predicted that the features produced by non-L1 Persian speakers pretending to be L1 Persian speakers will be significantly different from the features produced by genuine L1 Persian speakers. This study examines whether this hypothesis is correct. Testing this hypothesis is important for the practical applications of Native Language Identification, as it is important to understand the potential impact of an anonymous author attempting to disguise their language. The next section outlines the methodology for this study.

8.2 Methodology and Analysis

This study was designed to mirror the previous two studies as much as possible. The aim was to build a corpus of authors attempting to disguise their language to give the impression that they are L1 Persian speakers. Due to the complexity of the linguistic history required, such data can only be elicited. This is significant difference to the previous studies, which both relied on collected rather than elicited data. However, in this situation eliciting the data will not devalue it, as the aim will be the same as if it was collected – the aim will be to deceive the reader (or analyst). Collecting that data also enabled greater control over the variables, and the ability to collect more in-depth information about the authors and their linguistic histories. It is only possible to consider the attempt at deception in one direction; that of someone with a different L1 attempting to give the impression that Persian is actually their L1. There is no clear theoretical linguistic difference between an L1 Persian speaker attempting to disguise themself as an L1 English speaker, and an L1 Persian speaker who is writing in English without disguising their linguistic history. The only difference in this situation would be content based.

When collecting data, the predominant aim was to target L1 English speakers who could pretend to be L1 Persian, however, due to the difficulty in collecting data, and the multilingual histories of the people being surveyed, it was decided that limiting participants purely to monolingual L1 English speakers would overly restrict the data it was possible to collect. Each participant was asked to provide a linguistic history, so that the impact of the differing language influences could be considered if necessary during the analysis. It was

decided that in order to reach a greater number of participants, an online survey would be the best method. The free online software SurveyMonkey (www.surveymonkey.com) was used to create and distribute the survey. The link to the survey is http://www.surveymonkey.com/s/7Z8G568 a printed version can be seen in Appendix O. The quiz had several aims; firstly participants were informed about the purpose and aims of the study and asked to give their informed consent in order to participate (this and the further ethical considerations are discussed later in this section). Next the aim was to collect a comprehensive linguistic history of the participant, and their knowledge of the Persian language. Participants were also asked to list the features they would consider to be features of L1 Persian speakers writing in English. Finally, participants were asked to provide a text in the form of a blog, which although written in English, gives the impression that they are L1 Persian speakers. The full questionnaire can be seen in Appendix O.

A very high dropout rate was expected. Quite a lot was asked of the participants in the form of completing the questionnaire and the corresponding task, also in understanding the concept and motivations behind the disguise study. The quiz and task were designed to be as clear as possible and to not ask too much from the participants in terms of work or time required. However, in order to collect data that would be appropriate and comparable to the previous data sets, it was necessary to ask participants to provide as much writing as possible, as well as answering a significant number of questions. The questionnaire was designed to elicit as much information as possible, without unnecessarily burdening or demotivating the participants, as they were not recompensed in any way for their time or effort. It was expected that a large number of participants would start the quiz, but not complete the task section, this prediction held true. Participants that started the questionnaire, but did not complete the final writing exercise still provided valuable data (as will be discussed further in the next section).

Informed consent from participants was requested as part of the questionnaire. If a participant did not give their informed consent then the questionnaire did not progress to the following questions, meaning that no-one could participate in this section of research without first stating that they had understood the conditions of the study and were willing to participate. It was made clear that there was no form of compensation for participation and that overtly identifying information would be removed from the data. As with the previous studies, approval was granted by Aston University's School of Languages and Social Sciences Ethics committee (ethical applications can be seen in Appendix A). The study was

designed to adhere to Aston Universities Ethical Guidelines, the British Association of Applied Linguistics Recommendations on Good Practice and my own moral code. The underlying ethical principles are discussed more thoroughly in Chapter 4 Section 4.

Despite the difference in data collection, the structure of the analysis for this study predominantly follows the same structure as in previous chapters. The initial observations are based on a close reading of the disguise texts, and also the answers to the questionnaires. The next stage of analysis involved coding all the texts produced (some texts were excluded from this due to being incomplete and extremely brief, or because the author listed Persian as their L1). The texts were coded using NVivo and the same coding system as developed in Study One (see Chapter 5). The L1 Persian corpus from Study One formed the control corpus (as it did in Study Two), and the texts collected from the participants formed the studied corpus. Following the procedure employed in the previous two studies, the results were imported to SPSS, in order to perform statistical analyses to determine if there is a significant difference between the groups. Binary logistic regression was the most appropriate test, as this study is focusing on texts from two groups; the L1 Persian speaker blogs, and the texts from non-L1 Persian speakers trying to disguise their language to appear as L1 Persian speakers.

8.3 Findings

8.3.1 Initial findings from texts and questionnaires:

The initial questions in the survey yielded a few interesting observations. Participants explicitly referred to their experiences of Persian speakers when coming up with possible features. Some also referred to the processes of language acquisition, but this is likely to be linked to the number of language related professionals who participated. One participant credited their observations to how L1 Persian people speak in English, hypothesising that idiosyncrasies would be carried over into written text. I would hypothesise that although this is the only participant who explicitly attributed the features they expected to spoken language, this influence is probably more prevalent. Three participants predicted that Persian speakers might have a more artistic approach to writing, stating that they expected: “longer more ornate sentences”, “their writing to be more “flowery””, and “Elegant and descriptive prose” (responses to Question 9, Appendix Q). This is mirrored by the observation in Study One, Section 5.3 that L1 Persian speakers had a slightly higher average word length than the L1 English speakers.

The number of people that did not answer this question was considerably higher than the previous questions, which focused on linguistic background. This was expected, as it is not a simple question and requires a degree of relative experience with L1 Persian speakers; a few participants explicitly stated that they did not know what features to predict. Interestingly the predictions that were made by participants did not always match the features contained in their texts. This reflects the understanding in forensic linguistics that it is harder than generally expected to manipulate your language. Kniffka wrote that "Not even a professor of linguistics specializing in the analysis of anonymous authorship of criminal letters, would be able to manipulate his language (including orthographic) behaviour in so many different areas of **so many variables** at a time **over a long period of time**." (Kniffka, 1996, p. 90). The next step is to analyse the features that were actually contained within the produced texts and see whether they are significantly different from a statistic perspective with the L1 Persian texts.

8.3.2 Statistical analyses

The statistical analyses for this study mirror the analyses for Study One and Study Two as much as possible. As in both the previous two studies, binomial logistic regression is the most appropriate statistical analysis. The first stage of the analysis was to perform binomial logistic analysis in SPSS, using only the 12 higher level features from the tree of features. The output (see Appendix R) demonstrates that the model using just these features is over-fitted to these data. This was also experienced in Study One, but not in Study two. Table 8-1 below ranks the higher level features according to their Wald Chi Square score, which indicates how good each individual feature is at predicting group memberships; the higher the Wald Chi score, the greater the predictive ability of that particular feature.

Table 8-1 Study 3. Higher level features by Wald

Feature	B	Wald	Exp(B)
@101Lexical	3.017	.0000140869452	20.426
@194Pronoun	-5.598	.0000066034767	.004
@131Other	-2.018	.0000043093197	.133
@161Preposition	-1.202	.0000028390433	.301
@127Orderingorpositioning	-7.891	.0000026426252	.000
@65Conjunction	-1.889	.0000019832671	.151
@246Verbal	1.005	.0000019602684	2.733
@213Punctuation	-2.311	.0000007522950	.099
@11Adverb	2.457	.0000003357316	11.673

Feature	B	Wald	Exp(B)
@50Capitalisation	.306	.0000001416263	1.358
@1Adjective	-2.828	.0000000209900	.059
@26Article	.043	.0000000014789	1.044

It is interesting that Lexical is the most effective feature at discriminating between the disguise group and L1 Persian. Shuy's Devil Strip case saw the author deliberately misspell words such as *kop*, *trash kan* and *dautter*, all of which would be classified here as lexical features. Comprehension of the extortion note was required, however, the author wanted to disguise their language (and identity), so included these deliberate misspellings. Lexical features are also among the most frequent features in the Disguise corpus, indicating that a key strategy when disguising language is for the author to alter the lexis through marked construction, marked absence, marked presence, or marked choice.

Wilson and Wilson (2001) (as discussed in section 2.2.3) predicted several areas in which pronouns and prepositions may be of difficulty to L1 Persian speakers writing in English. It is interesting therefore that they are the second and fourth most predictive higher level features respectively. The *B* value indicates which way the feature works; a positive *B* value means the higher number of occurrences of that feature, the higher probability that the text belongs to the disguise group. Therefore, a negative B value means the more of that feature, the greater the probability that the author is an L1 Persian speaker. The *B* value of the higher level features *pronoun* and *preposition* are negative, meaning that while marked pronoun features may occur within disguise texts, they are more indicative of L1 Persian speaking authors. There is only a little information that can be uncovered from the higher level features, the lower level features provide a more accurate picture of the patterns in the data.

As predicted, using all 29 lower level linguistic features resulted in the model being overly fitted to the data, this was also witnessed in both previous studies. This resulted in a significance of 1, and a chi-square of less than 0.0005, which as discussed in Section 6.2 indicates the model is overly fitted and therefore not generalisable. The next stage is to reduce the number of features in order to determine the optimum model. In order to do this the features must be ranked according to their Wald Chi Square score, which indicates how good the individual features are on their own at determining group membership. The higher the Wald Chi Square score, the better the feature's predictive power. The Wald Chi Square score is determined from Block 1 of the logistic regression (see output in Appendix R). This is

less accurate than the Wald Chi Square scores given in Block 2, however SPSS was unable to proceed fully with Block 2 due to the overfitting of the model.

Most significant features:

Table 8-2 - Study 3 - Features Ranked by Wald Score

Rank according to Score	Feature	Score
		32.709
1	@106LexicalMarkedChoice	15.341
2	@265VerbalMarkedConstruction	10.284
3	@251VerbalMarkedChoice	9.679
4	@127Orderingorpositioning	8.715
5	@195PronounMarkedAbsence	8.563
6	@247VerbalMarkedAbsence	8.055
7	@2AdjectiveMarkedChoice	7.706
8	@173PrepositionMarkedChoice	7.399
9	@131Other	7.307
10	@117LexicalMarkedConstruction	6.064
11	@16AdverbMarkedChoice	5.438
12	@162PrepositionMarkedAbsence	5.425
13	@27ArticleMarkedAbsence	5.225
14	@12AdverbMarkedAbsence	4.521
15	@102LexicalMarkedAbsence	4.001
16	@33ArticleMarkedChoice	3.523
17	@200PronounMarkedChoice	3.014
18	@45ArticleMarkedPresence	2.450
19	@66ConjunctionMarkedAbsence	2.287
20	@80ConjunctionMarkedPresence	1.114
21	@122LexicalMarkedPresence	.988
22	@208PronounMarkedPresence	.906
23	@22AdverbMarkedPresence	.763
24	@213Punctuation	.451

Rank according to Score	Feature	Score
25	@7AdjectiveMarkedPresence	.371
26	@70ConjunctionMarkedChoice	.371
27	@181PrepositionMarkedPresence	.370
28	@75ConjunctionMarkedOrderingorpositioning	.151
29	@50Capitalisation	.006

It was discovered that the optimum model comprised the top 6 features; @106LexicalMarkedChoice, @265VerbalMarkedConstruction, @251VerbalMarkedChoice, @127Orderingorpositioning, @195PronounMarkedAbsence, and @247VerbalMarkedAbsence. This resulted in a Chi-square value of 3.813 with a significance value of 0.874 (see Table 8-3 below). This is considerably higher than 0.05, indicating that the model is reliable. It is also less than 1.000, indicating that the model is not over-fitted to the data. This is supported by the Model Summary (Table 8-4 below).

Table 8-3 - Study 3 - Optimum Model Hosmer-Lemeshow Test

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.813	8	.874

Table 8-4 - Study 3 - Optimum Model, Model Summary

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	17.507[a]	.473	.691

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

The Cox and Snell and Nagelkerke pseudo R square statistics signifies that the model accounts for between approximately 47.3 and 69.1 percent of the variability. The classification table (Table 8-5) below demonstrates that using the model of 6 features, only 4 cases are misattributed to the wrong group.

Table 8-5 - Study 3 - Optimum Model Classification Table

Classification Table[a]

	Observed		Predicted		
			Disguisegroup		Percentage Correct
			2.00	5.00	
Step 1	Disguisegroup	2.00	24	1	96.0
		5.00	3	6	66.7
	Overall Percentage				88.2

a. The cut value is .500

Group 2 represents the L1 Persian speakers, and Group 5 represents the disguise authors. Three disguise authors are misattributed as being native Persian speakers, and one Persian speaker is falsely assigned to the disguise group. This shows that while some authors may be very good at pretending to be L1 Persian speakers, only one Persian speaker out of 25 is mistaken as belonging in the disguise group.

The following table (Table 8-6) demonstrates the Wald and B scores for the six features within the model. The best predictor is *106Lexicalmarkedchoice* (as it has the highest Wald Chi Square score). It is interesting to note that of the six features that comprise the optimum model, three relate to Verbal features. The B values (see Table 8-6 below) also signifies that a higher number of occurrences of the three verbal values all indicate a greater probability that the author belongs to the disguise group rather than the L1 Persian group.

Table 8-6 – Study 3 - Optimum Model Variables in Equation

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	@106LexicalMarkedChoice	.220	.124	3.148	1	.076	1.246
	@265VerbalMarkedConstruction	.001	.136	.000	1	.996	1.001
	@251VerbalMarkedChoice	.089	.062	2.048	1	.152	1.093
	@127Orderingorpositioning	-1.762	1.368	1.660	1	.198	.172
	@195PronounMarkedAbsence	-.100	.470	.045	1	.832	.905
	@247VerbalMarkedAbsence	1.542	1.429	1.165	1	.280	4.676
	Constant	-4.165	1.498	7.727	1	.005	.016

a. Variable(s) entered on step 1: @106LexicalMarkedChoice, @265VerbalMarkedConstruction, @251VerbalMarkedChoice, @127Orderingorpositioning, @195PronounMarkedAbsence, @247VerbalMarkedAbsence.

8.4 Discussion

Casewise List

Case	Author	Observed	Predicted	Predicted Group
		Disguisegroup		
46	AT : Sammy	2**	.619	5
63	BK : Respondent 40	5**	.357	2
67	BO : Respondent 5	5**	.155	2
69	BQ : Respondent 7	5**	.098	2

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

Table 8-7 - Study 3 - Optimum Model Casewise List

This study shows that there are clear differences between L1 Persian authors and authors who are pretending to be L1 Persian speakers. The optimum model comprises six features: @106LexicalMarkedChoice, @265VerbalMarkedConstruction, @251VerbalMarkedChoice, @127Orderingorpositioning, @195PronounMarkedAbsence, @247VerbalMarkedAbsence. This model has a high accuracy, with a significance of 0.874 and a Chi-square score of 3.813. Using this model results in 4 authors out of 34 authors being misattributed to a group according to the features they exhibit. One L1 Persian author out of 25 is attributed as belonging to the disguise group, and 3 disguise authors are attributed as belonging to the L1 Persian group. This has important ramifications from a casework perspective, as it is considerably more likely that a disguise author will be attributed as an L1 Persian author, than the other way round.

It should be borne in mind, that as discussed in Study One, the nature of collecting data from the internet means that we cannot verify the author's stated L1. It is therefore conceivable that a low number of them may not actually be L1 Persian speakers. It could be the case that the author who was classified as belonging to the disguise group is actually not an L1 Persian speaker, but falsely claims to be so on their blog. This is however, purely speculation and cannot be verified within this research. Table 8-8 below shows that the L1 Persian author who was misattributed to the disguise group has a probability of 0.619; this is quite close to the 0.5 boundary line.

It is possible to trace the three wrongly attributed disguise authors to check their linguistic history and see what knowledge of the Persian language they have. Table 8-8 below contains their answers for the relevant questions.

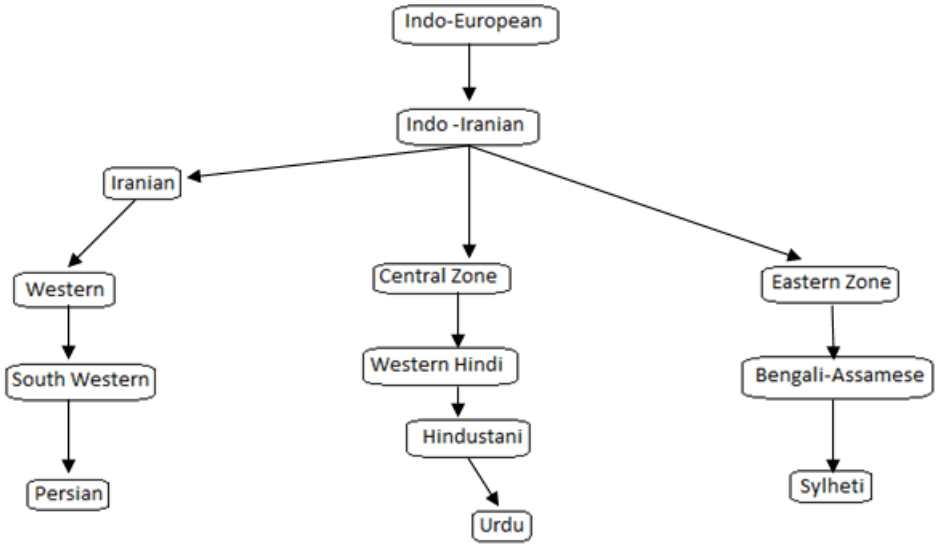
Table 8-8 - Study 3 - Authors' Linguistic Histories

Question	Respondent 5	Respondent 7	Respondent 40
Your Nationality	BRITISH	British	British
Your Native language	ENGLISH	English	English
Father's mother tongue(s):	SYLHETI	English	English
Mother's mother tongue(s):	SYLHETI	English	English
Partner's mother tongue(s):	N/A	English	Flemish
Please list any language(s) that are spoken in your home (if more than one, please give the average % use of each):	-	-	German (10%), Flemish (40%)
Please describe what contact you have had with the Persian (Farsi) language (years learning it, and courses or classes taken in Persian) and what contact you have had with Persian speakers:	As a speaker of Urdu, I often come across many words of Persian origin. I also have some friends who are from Iran.	Very little. I have Iranian friends and friends of Persian descent who have taught me odd words and I have heard them converse, but only very casually.	None
Please describe any stays in Persian-speaking countries, including when, where and for how long:	None	None	None
What features would you expect to find in the language of a native Persian speaker writing in English and why?	I'd expect grammatical errors, especially in syntax. This is because I've found Farsi (and Arabic) speakers to struggle with English grammar.	I wouldn't know to be honest.	None

Two out of the three authors do not list the features they would expect. This indicates that they are disguising their language using an intuitive approach, rather than focusing on the features they would expect, then inserting them into the text. Respondent 30, does not mention any specific contact with Persian speakers, or knowledge of the Persian language, however out of the three authors Respondent 30's probability score is closest to the 0.5

group membership boundary line, meaning they have a lower probability of belonging to the L1 Persian group than the other two authors. Respondents 7 and 5 both state that they have L1 Persian friends. Respondent 5 speaks Urdu and both parents are Sylheti speakers, this is interesting as both languages belong to the Indo-Iranian branch of the linguistic family tree (see Figure 8-1 - Linguistic Family Tree below).

Figure 8-1 - Linguistic Family Tree



(Created from information from Ethnologue (Lewis, 2009))

Study Two demonstrated that NLID can distinguish between influence from different languages that belong to the same branch of the linguistic tree, however there were indications that L1 speakers of the more closely related language, Pashto, were more likely to be misattributed as L1 Persian speakers than the less closely related language, Azeri. Respondent 5 could be exhibiting a greater probability of belonging to the L1 Persian group due to influence from Urdu and Sylheti, rather than from the conscious attempt to disguise their language.

Like Study Two, this study should be considered as exploratory, rather than complete. Due to the limited amount of data, there is a limited amount of analysis that can be performed, and we are restricted in the number of firm conclusions that can be drawn. The data were also elicited in an artificial manner, through an online survey. It is not known how a forensic

context would alter the language produced by an author attempting to disguise their language. However, this exploratory study still has many useful indications. The most significant conclusion we can draw from this study is that there is a considerable difference between the language and features produced by a genuine L1 Persian speaker and those of someone pretending to be an L1 Persian speaker. It is interesting from a casework perspective that although a disguise author may be misattributed as an L1 Persian speaker, the opposite is much less likely to happen. As in the previous studies, there is considerable information contained in the SPSS output (see Appendix R) that can be of use from a casework perspective, this is discussed more fully in Section 9.3.

Chapter 9. Discussion

> ***"If I claim full justice for my art, it is because it is an impersonal thing – a thing beyond myself. Crime is common. Logic is rare. Therefore it is upon the logic rather than upon the crime that you should dwell." – (Conan Doyle, 1892)***

This final chapter brings together all the findings and conclusions from the previous chapters and considers the wider implications. The first section, 9.1, examines the findings of the three studies with reference to the aims as set out in Chapter 4 section 1. Section 9.2 considers the practical aspects of this research, in particular the limitations. Section 9.3 then takes a practical look at casework applications , examining in greater depth how NLID, and this research in particular, can be of benefit to forensic authorship analysis. The fourth section (9.4) looks to the future, and drawing on findings from this research proposes ways in which NLID could progress.

- Chapter 9 - Discussions and Conclusions
  - 9.1 Summary of findings – answers to aims
  - 9.2 Practical limitations
  - 9.3 Casework applications and potential
  - 9.4 The future for this research and NLID

9.1 Summary of findings – answers to aim

The main aims of this research as set out in Chapter 4 Section 1 are as follows:

1. To determine if interlingual features in L2 writing can be used to indicate an author's native tongue (Study One, and throughout other studies)
2. To develop a methodology of NLID (Native Language Identification) (Study One)
3. To determine what features indicate authorship by a native Persian speaker (Study One)
4. To determine if we can attribute influence to being from a specific language rather than a language family, and to determine whether we can distinguish between two languages from the similar geographical area (Study Two).
5. To determine if it is possible to distinguish between a genuine native Persian speaker writing in English and someone who is trying to disguise their language to give the false impression that they have an L1 influence from Persian (Study Three)

6. To understand with what degree of accuracy we can draw conclusions based on the analysis involved (throughout all studies)

The first study sought to answer the first three aims. The basic finding from Study One is that Native Language Identification (NLID) is possible, and that it can distinguish authorship by an L1 English speaker and an L1 Persian speaker writing in English. A template of features was created. This template can theoretically be applied to any collection of texts. It is important to note that all features were data driven. These features were then tested using logistic regression to see if they were able to distinguish between authorship by L1 English and L1 Persian speakers. When used in combination the features indicated that they were over-fitted to the data, so the number of features was reduced in order to find the optimum model. The optimum model for distinguishing between authorship by an L1 English and an L1 Persian speaker incorporates ten features; @251VerbalMarkedChoice, @33ArticleMarkedChoice, @265VerbalMarkedConstruction, @102LexicalMarkedAbsence, @45ArticleMarkedPresence, @122LexicalMarkedPresence, @66ConjunctionMarkedAbsence, @22AdverbMarkedPresence, @200PronounMarkedChoice, and @208PronounMarkedPresence.

The fourth aim was to distinguish between influence from specific languages and influences that were merely indicative of non-nativeness. This was accomplished by a second study, which was smaller than the first one. Study Two compared the corpus of L1 Persian authors with a corpus of blogs by L1 Azeri and L1 Pashto speakers to determine if it was possible to distinguish which group an author belonged to, and to discover how the features altered. Study Two demonstrated that it was possible to use the features to determine group membership of the authors. As in Study One, the number of features had to be reduced to find the optimum model which accounted for as much variation in the data as possible without being over-fitted. In determining authorship between close languages it was determined that the optimum model contained the following twelve features: @66ConjunctionMarkedAbsence, @208PronounMarkedPresence, @162PrepositionMarkedAbsence, @106LexicalMarkedChoice, @102LexicalMarkedAbsence, @117LexicalMarkedConstruction, @122LexicalMarkedPresence, @265VerbalMarkedConstruction, @195PronounMarkedAbsence, @251VerbalMarkedChoice, @80ConjunctionMarkedPresence, and @12AdverbMarkedAbsence.

The fifth aim related to distinguishing between genuine L1 Persian authorship and an author who was attempting to disguise their language to give the false impression that they are an L1 speaker. Study Three sought to analyse this and determined that there was a clear difference between the groups. The optimum model for this mini study comprised six features: @106LexicalMarkedChoice, @251VerbalMarkedChoice, @127Orderingorpositioning, @247VerbalMarkedAbsence, @195PronounMarkedAbsence, and @265VerbalMarkedConstruction.

The sixth aim related to error rates. It was the case in all three studies, that the features tended towards over-fitting rather than a lack of reliability. The optimum model for Study One has a Chi-square value of 3.128 with a significance value of 0.793. This is considerably higher than 0.05 yet less than 1.000 indicating that using the ten selected features generates a model that is very reliable, but not over fitted to the data. The Cox and Snell R Square and Nagelkerke R Square values signify approximately how much of the variability is explained by the chosen features, which in this case is between 59.5 percent and 79.3 percent. The higher the percentage (or the closer the score is to one) the more accurate the model is. This is supported by the fact that in using the 10 features that comprised the optimum model, it was possible to correctly predict which group 92% of the authors belonged to. For Study Two the optimum model resulted in a significance of 0.922 and a Chi-square score of 2.571 in the Hosmer-Lemeshow Test, with the Nagelkerke R square and the Cox and Snell R square scores indicating that the model using the twelve features accounts for between 44.5 and 63.8 percent of the variability, as well as correctly assigning 82.9% of authors to the correct group. In Study Three the optimum model resulted in a Chi-square value of 3.813 with a significance value of 0.874, accounts for between approximately 47.3 and 69.1 percent of the variability, and correctly assigns 88.2% of the texts to the right group. It is clear that the models produced from each study are accurate at determining group membership.

The final underpinning aim of this research was to spark interest and promote research in the area of NLID. Findings from this research have already been presented at conferences, including at the International Association of Forensic Linguists biennial conference. This has helped to start the academic conversation about NLID that will encourage competing theories and views. There are also plans to publish findings in peer review forums, such as journals. This research is, moreover, not posited as being complete, but as part of an on-going research into NLID within forensic linguistics.

9.2 Practical limitations

This section discusses the limitations of this research and the implications of these limitations. The error rates for each study have already been thoroughly discussed within the relevant studies (Chapters 6, 7, and 8); this section does not focus on the reliability of the models created, but on the limits of the research. It is however worth noting here that none of the modes claim to be 100% reliable. If they did, this would give cause for concern (and exhibit over fitting as discussed within each study). Language and language users have a

considerable amount of variation, therefore it is unreasonable and potentially dangerous to expect a system of language analysis to have 100% reliability.

The difficulty of any research that has casework applications is that it is necessary to simplify situations in order to research them. This research has focused on blogs by authors who identify themselves as being L1 Persian speakers. As stated in Section 1.1 any text can become a forensic text, regardless of the genre. The fact that this research focuses on blogs does not mean that the findings are purely applicable to NLID casework that centres on blog data, or even computer mediated communication (CMC) data. It is necessary for the analyst to understand how different genres might affect the features within the data, and account for this in their analyses and reports. This is one of many reasons why it is so important that NLID is performed by linguists who have accrued experience in the relevant areas of language analysis.

Perhaps the main issue is that of intra-coder reliability. Due to the constraints of this research, it has been impossible to investigate the impact that having different coders performing the analysis would have. As already stated in Section 5.2 and Section 6.4, there is a degree of subjectivity in the coding system (as there is in any method of coding language). In order to minimise the effect of subjectivity, extensive notes were made to ensure consistency with coding decisions. The use of only one analyst added to the consistency of interpretation of the codes and the decisions when coding data. It is unknown how great the difference in interpretation would be across different analysts. The practical application for this research is that it will form a part of the forensic authorship tool-kit, meaning it is to be used by different analysts. A further study should consider to what extent findings and results of NLID analysis are replicable across analysts.

This research has not made explicit allowance for the possibility of multiple authors of a document. A casework situation which contained a document written by multiple authors would still benefit from NLID. It is possible that the interplay of different authors may affect the features that are contained within the text. This limitation is particularly significant, as NLID forms a part of wider authorship analysis methodology and an analyst should make allowance for the possibilities of multiple authorship as part of their wider considerations.

The coding of the texts holds considerably more information than has been considered here. It is possible that examining the lowest levels of codes further may slightly alter the understanding of the findings in this thesis. However, this research is not considered static,

this is discussed further in section 9.4 (see particularly Figure 9-1 Evolution of NLID). Studies Two and Three are only preliminary studies. They comprise a limited amount of data, from a small number of authors. Due to this limitation the conclusions that can be drawn from the findings are restricted. It is because of the limited data that the L1 Azeri and Pashto authors together form one corpus of 'other languages' (see Chapter 7 for further information). In order to draw conclusions about the features contained within the English of L1 speakers of these languages individually, it would be necessary to extend these mini-studies into full studies, which might show trends that are not visible here due to the reduced volume of data.

This research has also assumed a relatively simple linguistic history of each author. In part this is based on the information given by the authors, who have stated their L1 as being Persian, English, Azeri or Pashto (in Studies 1 and 2). In a forensic situation it is likely that an author (or suspect) might have a complex linguistic history (this is discussed further in the following two sections). Section 9.4 considers the future for NLID, as well as suggesting potential areas of future research. The majority of section 9.4 builds on the limitations discussed in this section. Section 9.3 considers the analysis so far from a casework perspective, which lays the groundwork for the future of NLID research.

It is important that any future related research or casework recognises the constraints and limitations of NLID. While the analysis from detailed in this project has much promising investigative potential, it is not intended to be used as evidence (as discussed in Section 1.2). It is also restricted by the same limitations as language analysis for the determination of origin of asylum seekers (see Section 3.3), in that it requires a linguistic expert and cannot indicate a person's origin. NLID features contained in an author's language can indicate a native-like influence from a certain language (in this case Persian), however, this is not the same as indicating a person's origin, and any analyst must be aware of the surrounding linguistic complexities.

9.3 Casework applications and potential

This coursework has a very clear casework application, and is therefore of interest to law enforcement agencies. One thing that cannot be overstressed is that Native Language Identification, as it is discussed in this research, should only be performed by qualified linguists with the necessary experience with linguistic analysis. This links to debates within language analysis for the determination of origin (LADO), although LADO is considerably different from NLID (see discussion in Chapter 3 Section 3. It is also intended that any future

applications of this research would respect the limitations (as discussed in Section 9.4) and the ethical considerations (as discussed in Section 4.4).

The aim of this research is to develop a methodology and discover features that can have a direct casework application. It is interesting therefore to consider how this research could benefit an actual case. One such case that occurred recently was from a Canadian Law Enforcement Organisation regarding a series of threatening emails that were sent to various judges, legal officials, police officers and private citizens. The main suspect was a man with a Persian-Italian linguistic background. He denied having authored the emails, despite the evidence against him. The question was thus whether evidence could be found in the texts to either support or contradict the claim that he authored them. There was a considerable amount of questioned data, predominantly in the form of emails. There was however, very limited data that was known to have been written by the suspect and was undisputed (known texts). The known data were also of a different genre, predominantly comprising of text from a completed online form.

This case raises several questions and interesting considerations. NLID aims to identify the native language of an author, but often there is already a suspect, or suspect pool. It is interesting to consider what role NLID can play in this situation. Focusing on the questioned texts, it is possible to perform a sociolinguistic analysis, of which NLID is a part. This could give information on whether the suspect's background was consistent with the profile, or help narrow down the suspect pool. As was the case in this situation though, if a case involves an L1 speaker of another language, then it is highly probably that there is a community or people surrounding the case that have a similar linguistic background. Understanding the features that may be indicative of authorship by an L1 speaker of a certain language (or languages) enables the analyst to look for features that may be distinct within the language; in addition to the features that are present from the L1 influence. This can also be extended to the influence of multiple languages. In this case the suspect had a Persian and Italian linguistic history, and the emails were written in English. The texts should therefore demonstrate features that are indicative of an influence from both languages. Due to the growth of multilingualism (Blommaert, 2010) it would perhaps be interesting and valuable to research how the interplay of multiple language influences would affect sociolinguistic profiling and, more specifically, feature identification in NLIID. This will be discussed further in Section 9.4.

There are also many practical issues surrounding casework, time being a key one. Performing a full analysis as detailed in these studies is very time consuming. While there may be a smaller volume of data than involved in this project, it is still not a quick process to import the data to NVivo, code it according to the full coding structure, then export the coding matrix to SPSS (via excel) and run the logistic regression. This would also require a very specific level of expertise with this exact analysis. Therefore the logical progression is to consider if we can look at the most distinguishing features in isolation. Table 2.2-1 below lists the features that comprise the optimum model for each study.

Table 9-1 - All Studies - Optimum Model Features

Rank	Study One	Wald	B	Study Two	Wald	B	Study Three	Wald	B
1	@251VerbalMarkedChoice	4.101	1.727	@66ConjunctionMarkedAbsence	3.535	4.364	@106LexicalMarkedChoice	3.148	.220
2	@33ArticleMarkedChoice	3.355	2.628	@208PronounMarkedPresence	2.463	6.313	@251VerbalMarkedChoice	2.048	.089
3	@265VerbalMarkedConstruction	1.629	-1.058	@162PrepositionMarkedAbsence	2.134	-1.632	@127Orderingorpositioning	1.66	-1.762
4	@102LexicalMarkedAbsence	0.853	1.355	@106LexicalMarkedChoice	2.109	0.242	@247VerbalMarkedAbsence	1.165	1.542
5	@45ArticleMarkedPresence	0.84	1.161	@102LexicalMarkedAbsence	0.852	-2.52	@195PronounMarkedAbsence	0.045	-.100
6	@122LexicalMarkedPresence	0	26.623	@117LexicalMarkedConstruction	0.83	0.103	@265VerbalMarkedConstruction	0	.001
7	@66ConjunctionMarkedAbsence	0	-53.241	@122LexicalMarkedPresence	0.657	-1.691			
8	@22AdverbMarkedPresence	0	-74.842	@265VerbalMarkedConstruction	0.385	-0.383			
9	@200PronounMarkedChoice	0	80.921	@195PronounMarkedAbsence	0.352	0.385			
10	@208PronounMarkedPresence	0	-16.168	@251VerbalMarkedChoice	0.053	0.073			
11				@80ConjunctionMarkedPresence	0	-40.801			
12				@12AdverbMarkedAbsence	0	-26.574			

Contained within this information is the *B* value. The B value indicates the extent to which the presence of that particular feature alters the probability of membership to the second group. Due to the order of analysis the second group for Study One was the L1 Persian corpus (the L1 English group was the first group), but for Studies 2 and 3, the L1 Persian corpus was the first group. The order of the groups affects the polarity of the B value. A higher positive B value means the greater number of occurrences of that feature, the higher the probability that the author belongs to the second group.

The practical implication of this is that it is possible to focus on these features in isolation, rather than analysing the texts for all potential features. This is best demonstrated through a practical example. The following is an extract from a blog that did not form part of the data for this research:

> ***My name is Jaleh but my friends call me Jamigen, I am a student at the University of Tehran [Pardis] where I *265* studying in the Faculty of Law and Political Sciences. My professor is Davoud Agahie. I *251* start this blog site as a school project. I *251* provide information on world affairs but mostly I know my own country Iran the best. My native language is Persian but I know some English and Arabic. This blog site will *251* write in English since I *265* trying to speak and write *127* English better. I will update this blog *122* site often. (Jaleh, 2011)***

The features are marked in the text with the corresponding number (see Table 9-1). The number is placed between two *symbols* as the author had already used square brackets. The text contains the following features:

- 251VerbalMarkedChoice x3 (Study 1, 2, & 3)
- 265VerbalMarkedConstruction x2 (Study 1, 2 & 3)
- 127Orderingorpositioning x1 (Study 3)
- 122LexicalMarkedPresence x1 (Study 1 & 2)

Each study much be considered in turn. Firstly, the features from Study 1 can be translated into the following equation to determine the probability that the text is written by an L1 Persian speaker

> ***Likelihood of membership to second group = (B value of feature for specific study x number of occurrences) + (B value of next feature x number of occurrences) [...]***

Study 1 Likelihood of L1 Persian author = (1.727 x 3) + (-1.058 x 2) + (26.623 x 1) = 5.181 + -2.116 + 26.623 = 29.688 times more likely to be L1 Persian

This is reassuring, as although there is a greatly reduced volume of text, the author identifies themselves as being an L1 Persian speaker and the features agree. The likelihood ratio constitutes moderate evidence by the standards of Champod & Evett, (1999)'s scale.

Study 2 likelihood of L1 otherlanguages author = (0.073 x 3) + (-0.383 x 2) + (-1.691 x 1) = -2.238 times more likely to be an L1 other languages speaker = 2.238 times more likely to be an L1 Persian speaker

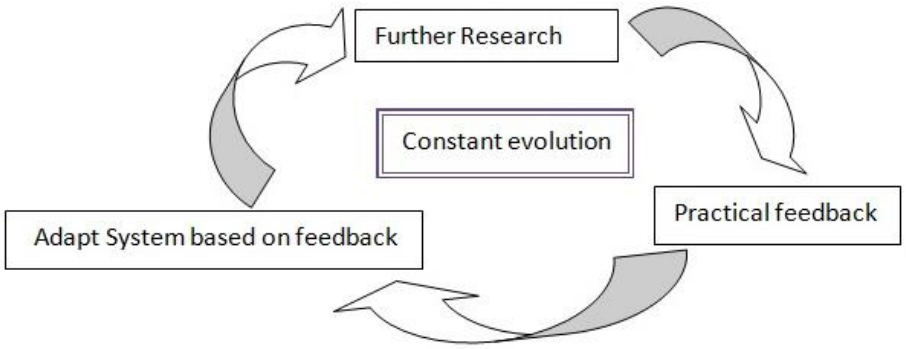
Study 3 likelihood of L1 disguise author = (0.089 x 3) + (0.001 x 2) + (-1.762 x 1) = -1.493 times more likely to be an L1 disguise author = 1.493 times more likely to be an L1 Persian speaker

This shows that even with very limited data, it is possible to focus purely on the specific features above in order to perform NLID. The values for Studies Two and Three only show limited evidence to support the group membership (according to Champod and Evett's 1999 scale). However, the text that is focused on is only 103 words long. A deliberately small section of text was chosen in order to demonstrate the procedure. The fact that even with a reduced volume of text, the results are as expected, speaks to support the reliability of the features. It is likely that with a greater volume of text, there would be more features, and the weight of evidence might be greater. Understanding *B* values avoids a black box approach, but is also open to greater bias (conscious or unconscious) than the full analysis. Analysts should be aware of this, and as strenuously stated throughout this research should be used in accordance with the relevant ethical guidelines (see discussion in Section 4.4).

9.4 The future for this research and NLID

Native language identification is a very under researched area. This research answers many questions, but it is intended as a preliminary step in this field. In order for Native Language Identification to be a useful tool for forensic authorship analysis, the field needs to evolve much further, and the best way for such evolution is more research. In discussing the progression of the wider field of forensic linguistics Coulthard (1994) stated that "the future [for forensic linguists] must lie in the creation of a better standardized and more widely used methodology." (Coulthard, 1994, p. 40). This can be applied to the narrower area of NLID. The methodologies discussed within this thesis have many merits, and have much potential on their own to aid forensic authorship analysis. However, it is only through testing competing theories and methodologies that NLID can evolve to its full potential. The practical aspect of this is also key: the need for NLID analysis was demonstrated through the cases discussed in Chapter 1 Section 2, and future NLID research should also be guided by casework. Figure 9-1 represents the proposed evolution of NLID, a significant feature of which is the cyclical and thus constantly evolving nature.

Figure 9-1 Evolution of NLID



The coding system devised contained considerably more information than could be thoroughly examined in this thesis. Future research is planned to investigate the lower levels of coding in more depth. This research indicates several areas for potential future research, which will be discussed in this section, but these should not be considered all encompassing. Different practitioners and researchers will have different perspectives, which could spark exciting new directions for NLID research. One main area of potential research that has arisen from this thesis is the need to compare different practitioners of NLID to better understand the inter-analyst reliability. Section 9.2 discusses the limitations of focusing on one genre, that of blogs, and the limitation of having only one coder. These limitations can both be diminished through future research, and are therefore a good starting point for future research.

This study has focused on a very narrow range of potential languages, predominantly L1 Persian speakers, but with the very limited inclusion of L1 Azeri and L1 Pashto speakers. The natural progression for NLID is to examine other languages, in order to build a better picture and understanding of features across languages and language groups. Examining other languages has several aspects. One suggestion would be the expansion of the corpora of Azeri and Pashto authors writing in English, as well as other languages that are both closely related to Persian and completely unrelated. The research discussed here focuses only of authors writing in English, but it would be interesting to compare L1 Persian authors writing in other languages. There is potentially a very large number of language combinations to consider. Chapter 2 discussed Dari and Tajik and their status as dialects of Persian, therefore when considering different languages to research it is interesting to re-visit the definitions of what constitutes an L1 speaker of a language. The difference between a dialect and a language is predominantly political (Weinreich, 1968; Winford, 2003), but it could be

possible that certain dialects result in differing features. While the opportunities for academic research into NLID seem endless, it is perhaps useful to remember the casework applications and to focus predominantly on language combinations that will aid in improving the theory and understanding that underpins the casework aspect.

This then raises the question that was discussed in Section 9.2 of the interplay between authors with complex linguistic histories; and how this affects the features of the individuals writing. Section 9.2 also discusses the issue of genre; this research focuses on blogs. It would be interesting to see how different genres affect the features and the process of NLID. It is also completely unknown whether the features exhibited in the research data examined here would transfer to spoken language. There is existing pedagogical research that examines interlanguage in the realm of spoken language, it may be of forensic value to extend this NLID research to spoken corpora.

The casework orientation of NLID brings the focus to the role of the expert in the future of NLID. Section 4.4 focused on ethics, and discussed the role of the expert practitioner from an ethical view point. While it is impossible to control how this research is used, it is designed to be used according to the ethical considerations discussed in Section 4.4, and in view of the limitations discussed throughout (and especially in Section 9.2). Where possible linguistic consultants should feed back to the academic and professional community on cases they have been involved with (within the confines of ethical responsibility to those involved) in order to help inform research and other casework (Linguistic Society of America Executive Committee, 2011).

It is clear that there are many potential directions in which this research can be expanded, and suggestions for the future that are contained in this thesis should not be considered as limiting, as it is impossible to discuss all areas of potential benefit. Perhaps the key element of this research is that it intends to spark a debate, and the potential future for NLID as discussed here, is just my perspective. In order to truly grow and develop to its full potential, input from a range of academics and practitioners is needed.

Bibliography

Addley, E. (2011). Syrian lesbian blogger is revealed conclusively to be a married man | World news | The Guardian. *The Guardian*. Retrieved November 15, 2012, from http://www.guardian.co.uk/world/2011/jun/13/syrian-lesbian-blogger-tom-macmaster

Androutsopoulos, J. (2010). Localizing the Global on the Participatory Web. In N. Coupland (Ed.), *The Handbook of Language and Globalisation* (Kindle., pp. 203–231). Maldon & Oxford: Wiley-Blackwell.

Angelis, G. De. (2005). Interlanguage Transfer of Function Words. *Language Learning*, *55*(3), 379–414.

AoIR ethic working committee, & Ess, C. (2002). *Ethical decision-making and Internet research* (pp. 1–33). Retrieved from w.aoir.org/reports/ethics/pdf

Appel, R., & Muysken, P. (1987). *Language Contact and Bilingualism.* London: Edward Arnold.

Aston University Ethics Committee. (2007). *School of Languages and Social Sciences Policy on Research Ethics*. Birmingham.

Barnbrook, G. (1996). *Language and Computers*. Edinburgh: Edinburgh University Press.

Baron, N. (2000). *Always On; Language in an Online and Mobile World* (Kindle.). Oxford & New York: Oxford University Press.

Bassett, E. H., & Riordan, K. O. (2002). Ethics of Internet research : Contesting the human subjects research model. *Ethics and Information Technology*, *4*(1), 233–247.

BBC. (2010). BBC News - Iran blogger Hossein Derakhshan temporarily released. *BBC News - Middle East*. Retrieved November 9, 2012, from http://www.bbc.co.uk/news/world-middle-east-11962032

BBC News. (2011). BBC News - Syria Gay Girl in Damascus blog a hoax by US man. *BBC News - Middle East*. Retrieved from http://www.bbc.co.uk/news/world-middle-east-13744980

Berghel, H. (2003). The discipline of Internet forensics. *Communications of the ACM*, *46*(8), 15.

Bermel, N. (2006). *Linguistic authority, language ideology, and metaphor : the Czech orthography wars*. Berlin: Mouton de Gruyter.

Bhatia, T. K., & Ritchie, W. C. (2004). Bilingualism in the Global Media and Advertising. In T. K. Bhatia & W. C. Ritchie (Eds.), *The Handbook of Bilingualism* (pp. 513–546). Oxford: Blackwell Publishing Ltd.

Bhatt, R. M. (2012). World Englishes. *Annual Review of Anthropology*, *30*(2001), 527–550.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Blommaert, J. (2010). *The Sociolinguistics of Globalization*. Cambridge: Cambridge University Press.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-sneddon, G., & Anderson, A. H. (1997). The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, *23*(1), 13–31.

Champod, C., & Evett, I. W. (1999). A. P. A. Broeders (1999) “Some observations on the use of probability scales in forensic identification”,. *International Journal of Speech Language and the Law*, *6*(June), 228–241.

Chaski, C. E. C. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, *8*(1), 1–65. Retrieved from http://www.extenza-eps.com/EPL/doi/abs/10.1558/sll.2001.8.1.1

Civil Procedure Rules. (2010). PRACTICE DIRECTION 35. Retrieved from http://www.justice.gov.uk/courts/procedure-rules/civil/rules/part35

Clyne, M. (2003). *Dynamics of Language Contact*. Cambridge: Cambridge University Press.

Comrie, B. (2001). Languages of the World. In M. Aronoff & J. Rees-Miller (Eds.), *The Handbook of Linguistics* (Kindle Edi., p. Loc. 384 – 718). Oxford: Blackwell Publishing Ltd.

Conley, J., & Peterson, D. (1996). When Ethical Systems Collide: The Social Scientisst and the Adversary Process. *Recent Developments in Forensic Linguistics1* (pp. 345 – 358). Frankfurt Am Main: Peter Lang GmbH.

Corder, S. P. (1974a). The Significance of Learners’ Errors. In J. Richards (Ed.), *Error Analysis. Perspective on Second Language Acquisition* (pp. 19–30). London: Longman.

Corder, S. P. (1974b). Idiosyncratic Dialects and Error Analysis. In J. Richards (Ed.), *Error Analysis. Perspective on Second Language Acquisition1* (pp. 158–171). London: Longman.

Coulthard, M. (1994). On the use of corpora in the analysis of forensic texts. *Forensic Linguistics*, *1*(1), 27–43.

Coulthard, M., Grant, T., & Kredens, K. (2010). Forensic Linguistics. In B. Johnstone, R. Wodak, & P. Kerswill (Eds.), *The SAGE Handbook of Sociolinguistics* (pp. 529–544).

Crystal, D. (2001). *Language and the Internet*. *New York*. Cambridge: Cambridge University Press.

Crystal, D. (2005). The scope of Internet linguistics. *Proceedings of American Association for the Advancement of Science Conference* (pp. 17–21). Washington, DC: American Association for the Advancement of Science Conference.

De la Rouchefoucauld, M. (1898). *Reflexions ou Sentences et Maximes*. Paris: Ollendorf.

Deaver, J. (1999). *The Devil’s Teardrop* (Kindle Edi.). London: Hodder and Stoughton.

Doyle, A. C. (1892). *The Adventures of Sherlock Holmes*. New York: Harper and Brothers.

Eades, D., Fraser, H., Siegel, J., McNamara, T., & Baker, B. (2003). Linguistic identification in the determination of nationality: a preliminary report. *Language Policy*, *2*(April), 179–199.

Federal Committee Rule 702. Testimony by Expert Witnesses | Federal Rules of Evidence | LII / Legal Information Institute (2011). Retrieved from http://www.law.cornell.edu/rules/fre/rule_702

Foster, D. (2001). *On the Trail of Anonymous*. New York: Henry Holt & Company.

Fraser, H. (2012). Language Analysis for the Determination of Origin (LADO). In C. A. Chappelle (Ed.), *Encyclopedia of Applied Linguistics* (pp. 9–11). Wiley-Blackwel.

Gao, X., Kong, H., & Sar, H. K. (2010). Ethical Challenges in Internet-based Research on Language Learners ’ Autonomous Learning : Personal Reflections, *46*(August), 13–17.

Gippert, J. (2010a). Iranian Languages. *TITUS: Thesaurus Indogermanischer Text- und Sprachmaterialien*. Retrieved December 1, 2012, from http://titus.uni-frankfurt.de/didact/karten/iran/iranm.htm

Gippert, J. (2010b). Turkic Languages. *TITUS: Thesaurus Indogermanischer Text- und Sprachmaterialien*. Retrieved December 1, 2012, from http://titus.uni-frankfurt.de/didact/karten/turk/turklm.htm

Grant, T. (2007). Quantifying evidence in forensic authorship analysis. *International Journal of Speech Language and the Law*, *14*(1), 1–25. Retrieved from http://www.equinoxjournals.com/ojs/index.php/IJSLL/article/view/3955

Grant, T. (2008). Approaching questions in forensic authorship analysis. In J. Gibbons & M. T. Turell (Eds.), *Dimensions of Forensic Linguistics* (pp. 215–229). Philadelphia, PA: John Benjamins Publishing Company.

Grant, T., Kredens, K., & Perkins, R. (2010). *Identifying an Author ' s Native Language Phase 2 + Finding and training the bilingual language expert* . Birmingham.

Haugen, E. (1966). *Language conflict and language planning: The case of modern Norwegian*. Cambridge MA: Harvard University Press.

Heine, B., & Kuteva, T. (2005). *Language contact and grammatical change*. Cambridge: Cambridge University Press.

Hendelman-baavur, L. (2007). Promises and Perils of Weblogistan: Online Personal Journals and the Islamic Republic of Iran. *Middle East Review*, *11*(2), 77–93.

Hopkins, E. (1982). Contrastive Analysis, Interlanguage, and the Learner. In W. Lohnes & E. Hopkins (Eds.), *The contrastive Grammar of English and German* (pp. 32–48). Michigan: Karoma Publishers Inc.

Hubbard, E. H. H. (1996). Errors in Court: A Forensic Application of Error Analysis. In H. Kniffka, S. Blackwell, & M. Coulthard (Eds.), *Recent Developments in Forensic Linguistics* (pp. 123–140). Frankfurt Am Main: Peter Lang GmbH.

Hundt, M., Nesselhauf, N., & Biewer, C. (2007). Corpus Lingusitics and the Web. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus Linguistics and the Web* (pp. 1–6). Amsterdam & New York: Rodopi B.V.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Internet World Stats. Usage and Population statistics. (2012). Top Ten Internet Languages - World Internet Statistics. *Internet World Users By Language*. Retrieved December 4, 2012, from http://www.internetworldstats.com/stats7.htm

Jaleh. (2011). Jamigen's Iranian Affairs Blog Site. Retrieved July 19, 2011, from http://jamigen.com/index.htm

Johnstone, B. (1996). *The Linguistic Individual. Self-Expression in Language and Linguistics*. Oxford: Oxford University Press.

Kachru, B. B., & Nelson, C. L. (1996). World Englishes. In S. L. McKay & N. H. Hornberger (Eds.), *Sociolinguistics and Language Teaching* (pp. 71–102). Cambridge: Cambridge University Press.

Kelly, B. J., & Etling, B. (2008). Mapping Iran's Online Public : Politics and Culture in the Persian Blogosphere. *Most*, 1–36.

Kent, R. (1953). *Old Persian: Grammar, Texts, Lexikon* (2nd rev. E.). New Haven: American Oriental Society.

Kent, R. G. (1936). The Present Status of Old Persian Studies. *American Oriental Society*, *56*(2), 208–225.

Khanlari, P. (1979). *A History of the Persian Language*. New York: Sterling Publishing Company.

Kiss, J. (2010). Iranian "blogfather" Hossein Derakhshan could face death penalty. *Guardian*. Retrieved November 9, 2012, from http://www.guardian.co.uk/media/pda/2010/sep/21/hossein-derakhshan-hoder-iran-blogger

Kniffka, H. (1996). On Forensic Linguistic "Differential Diagnosis". In H. Kniffka, S. Blackwell, & M. Coulthard (Eds.), *Recent Developments in Forensic Linguistics* (pp. 75–122). Frankfurt Am Main: Peter Lang GmbH.

Kniffka, H. (2000). Forensische Linguistik: anonymous authorship analysis without comparison data? A case study with methodological implications. *Linguistische Berichte*, *182*(182), 179–198. Retrieved from http://cat.inist.fr/?aModele=afficheN&cpsidt=1555504

Koppel, M., Schler, J., & Zigdon, K. (2005). Determining an author's native language by mining a text for errors. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining - KDD '05* (pp. 624–628). New York,: ACM Press. doi:10.1145/1081870.1081947

Kredens, K. (2000). *Forensic linguistics and the Status of Linguistic Evidence in the Legal Setting*. University of Lodz.

Language and National Origin Group. (2004). Guidelines for the use of language analysis in relation to questions of national origin in refugee cases. *International Journal of Speech, Language and the Law - Forensic Linguistics*, *11*(2), 261–266.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus Linguistics and the Web2* (pp. 133–150). Amsterdam & New York: Rodopi B.V.

Leonard, R. A. (2005). Forensic Linguistics. *The International Journal of the Humanities*, *3*, 65–70.

Leung, C., Harris, R., & Rampton, B. (1997). The Idealised Native Speaker, Reified Ethnicities, and Classroom Realities. *TESOL Quarterly*, *31*(3), 543–560. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21344620

Lewis, P. (2009). Ethnologue: Languages of the World. (P. Lewis, Ed.). Dallas: SIL International. Retrieved October 1, 2012, from http://www.ethnologue.com/.

Linguistic Society of America. (2009). *Linguistic Society of America Ethics Statement May 2009*. Retrieved from www.linguisticsociety.org/files/Ethics_Statement.pdf

Linguistic Society of America Executive Committee. (2011). *Code of Ethics for Linguists in Forensic Linguistics Consulting*. Retrieved from www.linguisticsociety.org/files/code-of-forensic-consulting.pdf

Mackey, A., & Gass, S. (2005). *Second Language Research* (Kindle.). Mahwah, New Jersey and London: Lawrence Erlbaum Associates.

Mahootian, S., & Gebhardt, L. (2007). *Persian* (Kindle.). London and New York: Routledge.

Maryns, K. (2004). Identifying the asylum speaker : reflections on the pitfalls of language analysis in the determination of national origin, *11*(2).

McMenamin, G. R. (2001). Style markers in authorship studies. *Forensic Linguistics*, *8*(2), 93–97. doi:10.1558/sll.2001.8.2.93

McNamara, T., & Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics*, *18*(1).

Mina, N. (2007). Blogs, Cyber-Literature and Virtual Culture in Iran. *Occasional Paper Series*, (15), 1–36.

Muehleisen, S. (2002). *Creole Discourse*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Nemser, W. (1974). Approximative Systems of Foreign Language Learners. In J. Richards (Ed.), *Error Analysis. Perspective on Second Language Acquisition1* (pp. 55–63). London: Longman.

Olsson, J. (2004). *Forensic Linguistics. An Introduction to Language, Crime and the Law*. London and New York: Continuum.

Pallant, J. (2010). *SPSS Survival Manual* (4th ed.). Maidenhead: McGraw-Hill.

Patrick, Pete. (2010). Linguistic Rights in the Asylum Context. *CamLing VI Presentation*. Cambridge. Retrieved from http://privatewww.essex.ac.uk/~patrickp/papers/CamLingLADO_Dec2010.pdf

Patrick, Peter. (2010). Linguistic Rights in the Asylum Context.

Perkins, R. (2009). *Interlingual Identifiers of an L1 German speaker writing in English .* Aston University.

Perkins, R., & Grant, T. (2013). Forensic linguistics. *Encyclopedia of Forensic Sciences*.

Rahimi, B. (2008). The Politics of the Internet in Iran. In M. Semati (Ed.), *Media, Culture and Society in Iran: Living with Globalization and the Islamic State* (Kindle., pp. 37–57). London and New York: Routledge.

Rahman, T. (1995). The Pashto language and identity-formation in Pakistan. *Contemporary South Asia*, *4*(2), 151–170. Retrieved from http://www.informaworld.com/openurl?genre=article&doi=10.1080/09584939508719759&magic=crossref||D404A21C5BB053405B1A640AFFD44AE3

Rampton, M. B. H. (1990). Displacing the "native speaker": expertise, affiliation, and inheritance. *ELT Journal*, *44*(2), 97–101. Retrieved from http://eltj.oupjournals.org/cgi/doi/10.1093/elt/44.2.97

Rashidvash, V. (2012). The Iranian and Azari languages. *Research on Humanities and Social Sciences*, *2*(5). Retrieved from www.iiste.org

Reporters Sans Frontiers. (2010). *Countries under Surveillance: Iran* (Vol. 435, pp. 247–248). doi:10.1038/435247b

Richards, J. (1971). A Non-Contrastive Approach to Error Analysis. *English Language Teaching*, *25*(3), 204–219.

De la Rochefoucauld, M. (1665-1678). *Reflexions ou Sentences et Maximes*. Paris

Sapir, E. (1949). *Selected Writing of Edward Sapir*. (D. G. Mandelbaum, Ed.). Berkley and Los Angeles: University of California Press.

Selinker, L. (1972). Interlanguage. *IRAL*, *10*, 209–31.

Selinker, L. (1974). Interlanguage. In Jack Richards (Ed.), *Error Analysis. Perspective on Second Language Acquisition* (pp. 31–54). London: Longman.

Shakespere, W. (1825). *King Richard II*. London.

Shuy, R. (2001). Forensic Lingusitics. In M. Aronoff & J. Rees-Miller (Eds.), *The Handbook of Linguistics* (Kindle., pp. 683–691). Oxford & Malden: Blackwell Publishing Ltd.

Siemund, P. (2008). Language Contact: Constraints and common paths of contact induced language change. In P. Siemund & N. Kintana (Eds.), *Language Contact and Contact Languages* (pp. 3–11). Amsterdam & Philadelphia: John Benjamins Publishing Company.

Silva, R. S., & Laboreiro, G. (2011). Automatic Authorship Analysis of Micro-Blogging Messages. *ReCALL*, 161–168.

Skjaervo, P. O. (2009). Iranian Languages. In K. Brown & S. Ogilvie (Eds.), *Concise Encyclopedia of the Languages of the World* (pp. 537–542). Oxford: Elsevier Ltd.

Solan, L. M., & Tiersma, P, M. (2005). *Speaking of Crime: The Language of Criminal Justice*. Chicago and London: The University of Chicago Press.

Sorell, T. (2011). Preventive Policing, Surveillance, and European Counter-Terrorism. *Criminal Justice Ethics*, *30*(1), 1–22. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/0731129X.2011.559057

Stubbs, M. (1996). *Text and Corpus Analysis*. Oxford: Blackwell Publishing Ltd.

Suren-pahlav, S. (2007). Persian NOT Farsi: Iranian Identity Under Fire : An Word " Farsi " for the Persian Language. *The Circle of Ancient Iranian Studies*, (July), 1–14.

The International Herald Tribune. (2010). From the International Herald Tribune - 100, 75, 50 Years Ago - NYTimes.com. *International Herald Tribune*. Retrieved December 4, 2012, from http://www.nytimes.com/2010/01/25/opinion/25iht-oldjan25.html?_r=1

The Law Commission. (2011). *Expert Evidence in Criminal Proceedings in England and Wales*. London. Retrieved from www.lawcom.gov.uk/expert_evidence.htm

Thomason, S. (2001). *Language Contact: An Introduction*. Baltimore: Georgetown University Press.

Tomokiyo, L. M., & Jones, R. (2001). You're Not From 'Round Here, Are You? Naive Bayes Detection of Non- native Utterance Text. In Association for Computational Linguistics (Ed.), *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies.* (pp. 1–8). Association for Computational Linguistics.

Tsur, O., & Rappoport, A. (2007). Using Classifier Features for Studying the Effect of Native Language on the Choice of Written Second Language Words. In P. Buttery, A. Villavicencio, & A. Korhonen (Eds.), *Cognitive Aspects of Computational Language Acquisition* (pp. 9–17). Madison: Omnipress.

Tyson, N. de G. (2012). Twitter @neiltyson. Retrieved December 1, 2012, from https://twitter.com/neiltyson/status/231893951483879425

University of Missouri-Kansas City. (2000). The Case Against Bruno Hauptmann: Key Prosecution Evidence. Retrieved from http://law2.umkc.edu/faculty/projects/ftrials/Hauptmann/incriminevidence.html

Warschauer, M., Black, R., & Chou, Y.-L. (2010). Warschauer, M., Black, R. W., & Chou, Y.-L. (2010). Online Englishes. In A. Kirkpatrick (Ed.), The Routledge Handbook of World Englishes (pp. 490-505). New York: Routledge. In A. Kirkpatrick (Ed.), *The Routledge Handbook of World Englishes* (pp. 490–505). New York: Routledge.

Weinreich, U. (1953). *Languages in contact*. The Hague: Mouton & Co.

Weinreich, U. (1968). *Languages in Contact: Findings and Problems* (pp. 89–99). Paris: Mouton.

Weisburd, D., & Britt, C. (2007). *Statistics in Criminal Justice* (3rd Editio.). New York: Springer.

Williams, A. (2009). Pahlavi. In E. Brown & S. Ogilvie (Eds.), *Concise Encyclopedia of Languages of the World* (pp. 827–828). Oxford: Elsevier B.V.

Wilson, L., & Wilson, M. (2001). Farsi Speakers. In M. Swan & B. Smith (Eds.), *Learner English: A Teacher's Guide to Interference and Other Problems* (Second Edi.). Cambridge: Cambridge University Press.

Windfuhr, G. (2009). Persian. In B. Comrie (Ed.), *The World's Major Languages* (2nd ed., pp. 445–459). Abingdon & New York: Routledge.

Winford, D. (2003). *An Introduction to Contact Linguistics*. Oxford: Wiley-Blackwell.

Wittgenstein, L. (1922). *Logico-Tractatus Philosophicus*. London: Routledge.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell. Translation by G.E Anscombe

Wong, S. J., & Dras, M. (2011). Exploiting Parse Structures for Native Language Identification. In Association for Computational Linguistics (Ed.), *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 1600–1610). Edinburgh.

Wong, S. J., Dras, M., & Johnson, M. (2011a). Topic Modeling for Native Language Identification, 115–124.

Wong, S. J., Dras, M., & Johnson, M. (2011b). Exploring Adaptor Grammars for Native Language Identification, (July), 699–709.

Wong, S.-M. J., & Dras, M. (2009). Contrastive Analysis and Native Language Identification. In L. A. Pizzato & R. Schwitter (Eds.), *Australasian Language Technology Association Workshop (ALTA)* (pp. 53–62). Sydney. Retrieved from http://www.alta.asn.au/events/alta2009/index.html

Zipf, G. K. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge, MA: Harvard University Press.

Appendix List

A. Ethical applications
B. Data overview spreadsheet (all studies)
C. L1 English Corpus (for full corpus see USB)
D. L1 Persian Corpus (for full corpus see USB)
E. Full feature tree
F. Main PhD Analysis – NVivo Project
G. Study 1 Word Frequency List
H. Coding Matrix Results (all studies)
I. Normalised Coding Matrix (2,000w)
J. Study 1. Graph of Frequency of Lower Level Features
K. Study 1 – SPSS Output
L. Study Two - L1 Azeri Corpus (for full corpus see USB)
M. Study Two – L1 Pashto Corpus (for full corpus see USB)
N. Study Two – SPSS Output
O. Study Three - Blank SurveyMonkey questionnaire
P. Study Three – Anonymised Questionnaire Answers
Q. Study Three - L1 Disguise Language Corpus (and predicted features)
R. Study Three – SPSS Output
S. SPSS Data Set (all studies, on USB)

Please contact Ria Perkins (perkinrc@aston.ac.uk) or Dr. Tim Grant (t.d.grant@aston.ac.uk) to access appendices.