## Article

Native Language Influence Detection for Forensic Authorship Analysis. Identifying L1 Persian Bloggers.

Dr Ria Perkins and Professor Tim Grant

Centre for Forensic Linguistics, Aston University

# Abstract

This article demonstrates and examines the potential use of interlingual identifiers for forensic authorship analysis and Native Language Influence Detection (NLID). The work focuses on the practical applications of native language (L1) identifiers by a human analyst in investigative situations. Using naturally occurring blog posts where the writer self-identifies as a native Persian speaker, a human analyst derived and coded sets of non-native features. Two logistic regression models were built, the first was used to select features to distinguish L1 Persian speakers from L1 English speakers in their English writings. The second developed a feature list to contrast L1 languages that are geographically and linguistically close to Persian. The results clearly demonstrate that interlingual identifiers have the potential to aid in determining the L1 of an anonymous author and can be used by a human analyst in a short forensically-realistic example text. This article demonstrates that Native Language Influence Detection is possible beyond the more common computational approaches and can form a useful tool in the forensic linguist’s toolbox. This study is not a statistical validation study, instead it demonstrates how a sociolinguistic approach can complement more traditional computational approaches.

**Keywords**

Native language identification, authorship analysis, linguistic profiling, Native Language Influence Detection, Persian

# Background

## 1.1 Introduction

Simply defined, Native Language Influence Detection (NLID) seeks to identify influence from an L1 or other language on an anonymous author’s writing in an L2. We prefer to use the terminology *Influence Detection* over that of *Identification* as it acknowledges the probability that any L2 might be influenced by several languages rather than a single L1. NLID work has a clear application in a forensic authorship analysis context (Dras & Malmasi, 2015; Li, 2013; Tetreault, Blanchard, & Cahill, 2013). Our approach takes a forensic linguistic approach, grounded in sociolinguistics and with a particular focus on the explanatory power of interlanguage. The term *interlanguage* was first coined by Selinker (1972) but the concept predates then, being evident in early research into contact linguistics and most notably the research of Weinreich (1953). 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: 35).

Until recent interest in NLID, the majority of studies into non-native varieties of English have been directed by pedagogic interest in interlanguage and cross-linguistic influence and are grounded within the field of second language acquisition. In 1957 Lado indicated that one could predict the errors a learner would make by studying the native language (L1 or NL) and target language (TL) or Second language (L2). This approach can be seen in the work of others, including Hopkins (1982), who wrote that contrastive analysis ‘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: 32). Although Hopkins was more explicitly referring to contrastive analysis, the prediction of errors is a common interest, as forensic analysis - where there is no pedagogical motive - needs to document features and understand their distribution. The contrastive approach to native language transfer has come up against much criticism, in particular Richards (1971), who demonstrated a systematic recurrence of errors that could not be attributed to influence from constructions in either the L1 or L2. The research here is observational, not predictive, but an understanding of this theory and potential influences enables a better understanding of the causes for the variations observed. That is not a rejection of all pedagogic research in the area, but rather a call to recognise the different aims.

Corder stated that ‘We must attempt to describe this language [interlanguage] in its own terms, at least in the first instance, and not in those of any other language’ (Corder, 1981: 56). This later view is much more closely allied to the current research, in which the focus is on identifying linguistic features of a particular sociolinguistic group (that of L1 Persian speakers writing in English) rather than understanding these features as errors. This article is therefore, focused on identifying an anonymous author’s native language through distinctive language use (not just language errors) and as such includes features such as constructions (at lexical or sentential level), word choices, word ordering, grammatical or spelling errors, and potentially even hypercorrection.

There are of course many difficulties with the concept of an L1, a native language, or a mother tongue. Many people do not have a monolingual linguistic background and multilingualism is increasingly prevalent (Thomason, 2001), as is the influence from other languages through, for example, globalised media. Our focus is on Persian and if we consider the history of Iran and the Middle East in general, we can see that available language maps do not match the political borders, and that there are a wide range of languages spoken in different areas. This has methodological implications as well as theoretical ramifications for selecting data – in our case 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: 555). As the data for this research are is collected from naturally occurring sources (rather than elicited from experiments) the linguistic background collected for each author is constrained by what the author has already written online. For this research an author’s L1 or native language, will be categorised by statements of self-identification. A further implication of the complexity surrounding being a native speaker of a language, is that of multilingualism where speakers can claim multiple L1s. In the case of bi- or tri-linguals, these speakers would likely have native-like influence from all the languages they consider to be their L1, for this research though, only authors who identified as having one L1 were selected. This research speaks of native language influence detection (NLID), rather than native language identification (NLI), in order to reflect these theoretical complexities which are likely to be mirrored in a casework situation.

NLID has been attracting a growing research interest but this has been mainly from a computational linguistic approach (Koppel, Schler, & Zigdon, 2005; Tomokiyo & Jones, 2001; Tsur & Rappoport, 2007; S.-M. J. Wong & Dras, 2009; S. J. Wong & Dras, 2011; S. J. Wong, Dras, & Johnson, 2011). 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. More recently there was the NLI (native language identification) shared task (Tetreault et al., 2013), which challenged teams of experts on a series of NLI related tasks; the majority of participants were from computational linguistics of Second Language Acquisition (SLA) research fields, and this is reflected in their approaches and findings. The current study is not intended as a statistical or computational validation study; rather it is an exploration of how linguistic features can be analysed in a forensically relevant data set, and hence how this can contribute as a complement to a more statistical approach.

Although computational and statistical research may have value in identifying some core interlingual features there are a number of problems which need to be addressed. A first issue is the nature of the training data, Koppel, Schler and Zigdon (2005) are followed by many others in using the ICLE corpus or similar data sets of learner English. This is perhaps exemplified by Brooke and Hirst (2012) who use five pre-existing corpora for their study, of which four are from formal learning or examination situations and one (Lang-8) they compiled from a language learning website where learners ‘write journal entries in their L2 to be corrected by native speakers’ (Brooke & Hirst, 2013: 191). For development of NLID work in the forensic context there is a real question of how the data sets are generically similar or different to possible forensic texts. Language learner essays are written for a specific purpose (to practise certain linguistic constructions) with a specific audience in mind (the language teacher) and to a fairly high level of formality. Although there may be no core genre of *forensic texts* the literature suggests that variety can be considerable and informal features typical of computer mediated communication may be key (see e.g. Grant 2008, and Grant 2010). Brooke and Hirst’s (2012) use of Lang-8, goes some way to addressing the issues with the change of mode (online), but the purpose and audience nonetheless remain firmly within the language learning domain.

A second issue is that *black box* classification typical of many computational approaches gives limited attention to the potential causes of features, leaving them vulnerable to misinterpretation. For example Koppel, Schler and Argamon (2009) identified the term *however* as being useful for identifying L1 Bulgarian speakers; it could be suggested that this feature does not have a structural interlingual explanation, but instead is evident in the student texts because the main text book used by the Bulgarian students advocates using this term above other potential synonymous options which might be favoured by a native speaker. The results from the 2013 NLI shared task (Tetreault et al., 2013) show that there were commonalities in the approaches taken, with most teams choosing to use Support Vector Machines for machine learning algorithms. The challenge also resulted in a ‘very similar set of standard features and machine learning methods’ (Malmasi, 2016: 46) from the entrants, the most common features being n-grams of character, word and part-of-speech as well as syntactic features. Unsurprisingly, this mirrors a lot of the existing research and literature in this field (Malmasi, 2016). For the investigative context, whether in providing intelligence or evidence for a courtroom, explanation is a key requirement. Forensic contexts require a weighing of different, often conflictual, sources of evidence and it is important to have the ability to provide explanation about why a feature indicates a particular L1 rather than being a typical error that a native writer might make, or indeed indicate a different language. Providing explanations alongside statistics of the predictive power of features adds to the validity of an analysis and enables the linguistic analysis to be evaluated alongside other sources of evidence. Malmasi & Dras, (2014) are the first authors to demonstrate that there are plausible explanations for the features they identify through computational methodology. They focus on the over- and under-occurrence of certain features such as function words, content words, syntactic dependencies, determiners and misspellings. However, their approach is still grounded in a computational approach with a descriptive linguistic understanding coming second. This current paper is proposing the opposite approach, grounded initially in sociolinguistic theory with descriptive linguistics being based on naturally occurring data.

There is, however, strong evidence that interlanguage analysis has the potential to be an invaluable investigative tool (Grant, 2008; Koppel, Schler and Zigdon, 2005). The increasing prevalence of multilingualism, with up to a quarter of the world’s population possessing some competency in English (Bhatia & Ritchie, 2004), means that English focussed NLID research can make a valuable contribution for forensic linguistic authorship analysis. The focus of this paper is the examination of writing in a variety of English, and determining the influences of other languages on their production.

## 1.2 Requirement and case history

There is a limited amount of academic literature detailing casework of forensic profiling of non-native writers. Perhaps the earliest cases relating to NLID to be fully reported from a forensic linguistic perspective are reported by Kniffka (1996), but evidence of the potential impact can also be seen in cases reported by the media (such as the Hauptmann trial), as well as in fiction. Sir Arthur Conan Doyle’s famous detective Sherlock Holmes, in the case of A Scandal in Bohemia says

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: 8)

Sherlock Holmes determines from the grammar used in an anonymous note that the author of the note must be a native German speaker. This is of course fictional, and the interlingual feature identified by Holmes was actually created by a Scottish fiction writer. It does still highlight that there is a common perception that one can determine someone’s L1 (native language) by the way they use an L2 (second or target language) language.

One of the earliest real cases to demonstrate the potential of NLID for forensic authorship analysis is the trial of Bruno Hauptmann for the kidnap and murder of the Lindbergh baby which received much media attention at the time. Hauptmann was a German immigrant living in America and in 1936 he was sentenced to death for the kidnap and murder of the one-year-old son of Charles Augustus Lindbergh, Jr. During the trial both the prosecution and the defence called a ‘handwriting expert’. Although the majority of the expert testimony focused on the shape and formation of individual letters, attention was also paid to phrase construction and grammar. The conclusion by the prosecution’s expert psychiatrist was that:

all the notes exhibited heavy Germanic influence, not just occasional German words but the pervasive “Germanic phraseology.” To an untrained observer, it might appear to be poor English grammar. If one “translated” the English back into German, however, the notes revealed an author who “had the ability to think correctly along German lines and thus we are forced to the conclusion that he is a German.” (*The International Herald Tribune*, 2010)

While this testimony has weaknesses and it is unlikely that it would be admitted in court today, it does highlight the concept that one can identify someone’s native language through the way they speak or write in a second language and that there is an investigative and legal application for such research.

More recently Kniffka (1996) detailed a German case on which he was asked to consult, which involved threatening letters being sent within a company. The content indicated that the anonymous author was most likely an employee of the company. Kniffka noted several interesting features which on their own were not significantly indicative as to who could have authored the documents, yet with the inclusion of the awkward lexical collocations and non-idiomatic use of German proverbs, he determined that the texts were most likely written by a non-native German speaker. The police using this information placed the only two L2 German speakers in the company under surveillance and eventually discovered the American employee in the process of writing another malicious letter.

This paper focuses on Persian L1 speakers writing in English. To date there are no documented forensic linguistic cases involving actual or suspected L1 Persian speakers writing in English, or native speakers of closely related languages. The research reported here, however, forms part of a wider string of research projects, focusing on a wider range of languages (and related NLID questions).

Persian is the most frequently spoken modern Iranian language (Comrie, 2001; Mahootian & Gebhardt, 2007), but it is only the native language (or L1) for approximately half of the population of Iran (Mahootian & Gebhardt, 2007). Persian is also frequently referred to as *Farsi* which is the Persian word for the Persian language. The use of the term *Farsi* however has political connotations for some and has provoked much debate and controversy (Suren-pahlav, 2007). Persian belongs to the Indo-Iranian group of the Indo-European language family and has approximately 57 million L1 speakers (Paul, Simmons and Fennig, 2015). 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, or grammatical variation, this research will adhere to the more prevalent view that they are varieties of Persian. As the data for this research is based on how the author self-identifies online, it would be very difficult to distinguish between the authors of the different dialects, as this information is rarely included in for example, blog posts.

Research into L1 Persian L2 English interlanguage features tends to focus on specific features, rather than seeking to provide an overview such as is required in NLID investigation. The most notable exception is a comparative linguistic work by Wilson and Wilson (2001). In an edited volume aimed at teachers, they identify features – predominantly errors – which may be encountered by L1 Persian speakers learning English. There is no explicit reference to evidence based empirical research (as opposed to purely hypothesised difficulties or anecdotal evidence) giving rise to these features. The paper does, however, provide an interesting comparison to the features that this research identifies using a data-driven approach.

This research has three aims:

* First, to determine how far it is possible to use interlingual identifiers to indicate the Persian L1 influence of an anonymous author.
* Second, to conduct an investigation which is forensically relevant. To further this aim the data must reflect a genre closer to those which occur in forensic investigations. Most existing research relies on elicited data from language students, which is not a key genre for forensic authorship analysis (though some computational approaches are now expanding to include blog posts, e.g. Brooke and Hirst, 2012). More than this the features need to be linguistic features open to linguistic explanation with regard to interlanguage.
* Third, to examine how far features developed for Persian can apply to culturally and linguistically related languages.

### 1.2.1 LADO

Language analysis for the determination of origin (LADO) holds many similar aims to NLID in that both seek to determine more about an author and their origin from the use of language. LADO was defined by Fraser (2012) as a form of applied linguistics ‘used by governments to assess asylum seekers applying for refugee status’ (Fraser, 2012: 9). The procedure varies considerably from country to country, but commonly, spoken interviews are analysed by a native speaker of the language (though not necessarily the dialect) sometimes in combination with a linguist.

There are however, some key differences between LADO and NLID. LADO uses elicited spoken interviews, rather than written collected data. LADO reports also play a key role in whether an asylum seeker is granted refuge, and as such is taken as evidence in decision proceedings. Forensic linguists and linguists working with LADO have recently made a significant effort to influence procedures relating to language analysis of refugee applicants. Eades, Fraser, Siegel, McNamara, and Baker (2003) produced a report documenting several key failings in the way LADO was being performed in Australia. Conversely, the current work is intended to provide investigative assistance and intelligence to investigations. It falls into the domain of forensic authorship analysis referred to as sociolinguistic profiling (Grant, 2008). For good linguistic and legal reasons no sociolinguistic profiling analysis has ever been accepted as evidence to a UK court or to the authors’ knowledge to other courts with Common Law jurisdictions, and there are limited examples of it being accepted in non-common law jurisdictions (de Vel, Anderson, Corney, & Mohay, 2001).

# Studies

This article presents findings from two complementary studies. Study One focuses on using interlingual features to determine which features are good for determining whether a text was written by an L1 Persian speaker writing in English, or an L1 English speaker. Study Two, a smaller sub-study, demonstrates that the features identified were indeed indicating L1 Persian influence rather than just non-native speech. For this purpose it looks at two other languages; Azeri and Pashto. This smaller study does not analyse these secondary languages in depth, but rather serves as a scoping study into these languages to see if the features identified as indicating L1 Persian influence, might also indicate influence of languages that were geographically or linguistically close, or whether the features can distinguish between them.

## 2.1 Data

The data for the main study comprises two corpora; one of L1 Persian speakers blogging in English, and a control corpus of L1 English speaker blogs. The blogs covered a wide range of topics. Data was collected from 25 authors for each corpus, all who self-identified as being L1 (or native or mother-tongue) speakers of the relevant language without contradicting themselves. While this does not guarantee the validity of their linguistic background it diminishes the probability that a significant number of the chosen authors were selected for the data based on false information.

The data for the second smaller study, which investigates related languages, comprises blogs from 5 L1 Azeri authors and 5 L1 Pashto authors blogging in English online. The criteria for data collection mirrored that of the main study in that authors were selected based on their self-identification of their L1. Table 1 below shows the distribution of collected data across all the corpora and studies.

Table 1 - Data Breakdown

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Corpus** | **L1** | **Number of authors** | **Total number of Blog posts** | **Total Word Count** | **Studies** |
| L1 Persian Corpus | Persian | 25 | 180 | 49530 | 1 |
| L1 English Corpus | English | 25 | 150 | 55357 | 1 & 2 |
| Other languages | Azeri | 5 | 33 | 6723 | 2 |
| Corpora | Pashto | 5 | 20 | 5981 | 2 |
| Total: | Azeri & Pashto | 10 | 53 | 12704 | 2 |
| Grand Total |  | 60 | 383 | 117591 |  |

The data sets are small, as data with reliable linguistic background information was prioritised over quantity of data. This also replicates the data that one is more likely to find in a casework situation.

## 2.2 Feature identification

A sub-set of texts were analysed to determine which features were present, where, and in what quantity. This stage of coding focused on documenting all potential features to a very fine-grained level; a feature is any marked language use. By marked language, we mean language that is unlikely to be used by an L1 English speaker in the given context. This involved traditional linguistic analysis and expertise, examining the texts for apparent language errors and non-native features. Through this detailed qualitative analysis certain trends within the features were identified; this enabled the expansion of the observed features through the creation of a feature tree (which accounts for all the features present, as well as a range of potential new features and establishes all features into linguistically plausible groups so that overarching trends can be analysed. Figure 1 below shows the underlying structure of the feature tree.

Figure 1 - Feature Tree Structure

[xxx Figure 1 near here]

The features were loosely categorised into groups (see left hand column) and within this they were then categorised into why the language was marked; whether it was a marked presence, absence choice, etc. Under virtually all the broad grammatical categories there was the possibility for: *Marked Presence*, *Marked Absence, Marked Choice, or Marked Construction.* In some situations *Marked Position* or *Ordering* was also a possibility*.* These constitute the 29 mid-level features.Within this framework the marked feature would then be further coded to represent how marked it was. Furthermore specific codes were also added to identify trends at a lower level; for example, in the case of a missing article, these would state which article was missing (when it could be easily determined without relying on a significant assumption). Additional information was also collected in the form of features, to help distinguish potential influences such as typing errors or homophone confusion. This system minimises assumptions from the analyst, but still enables the codes to reflect the difference between a genuine spelling error and a likely typing error. It should be noted that while this information was collected, the secondary ‘evaluative’ codes are not feeding in to the models being discussed in this paper. Instead they are feeding in to the wider project and later stages of analysis. An example of the coding can be seen below:

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 the 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 for a native speaker 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 it is coded as Non-standard rather than Awkward.

Naturally there is an element of subjectivity to this coding and judgements between codes, for example of 1-Awkward or 2-Non-standard. A number of approaches were used to ensure consistency of coding; a single analyst (Perkins) performed all the coding and developed a written coding framework which was developed in discussion with a secondary analyst (Grant). Reference to the coding framework helped ensure consistency of coding within the analyst. For this study we were less concerned with inter-rater reliability as the focus is more on validity and explanation of features. Variability within the coding and the risk of analyst error is discussed later and is part of the on-going project.

Having established the structure of a feature tree using the subset of data, more specific features were identified during the full coding of the texts, which fit within this existing framework. A total of over 300 features were identified, many of which related to very precise descriptive markers. The total frequencies for the mid-level features can be seen in Figure 2 below; it is interesting to note that they more or less follow Zipfian distributions which indicates that they are following expected trends within language and linguistics (Zipf, 1932).

Figure 2 - Feature Frequencies

[xxx Figure 2 near here]

## 2.3 Statistical design

After the blogs were coded for all of the potential features, statistical classification was used to determine the discriminatory power of different features. Logistic regression can be used to predict 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. In the first study the potential outcomes are that any given text could belong to the group of native Persian speakers, or the group of native English speakers; in the second study the outcomes are that the text is authored by an L1 Persian speaker, or, alternatively, that the author is a speaker of one of the closely related languages. Logistic regression has received recognition as a useful tool for criminal justice and forensic analysis, most explicitly from Weisburd and Britt (2007).

The first stage of the analysis used all of the 29 mid-level features, as outlined above in Section 2.2. The analysis also produceda range of statistical output, including Wald X2 scores for each feature. This information could be used to help reduce the number of features in the model to determine which features comprised the best, most statistically valid model for further analysis and casework.

## 2.5 Findings

Statistical analyses for Study One created a model using all 29 mid-level features in combination that could correctly determine group membership for each of the 50 authors. Given the number of initial variables, the Hosmer-Lemeshow test unsurprisingly suggested the model was over-fitted to the data and hence would not be generalizable or adaptable to other data. Such over-fitting is particularly undesirable in the forensic context and for potential casework applications of NLID. One strategy for reducing over-fitting is to reduce the number of predictive features. Features with the strongest predictive power can be identified by ranking them using the highest Wald X2 scores. Using this, the number of features was reduced until a more optimal balance was achieved between good prediction and over-fitting.

The features that were the most discriminating and comprise the optimum model for Study One were (in descending order of discriminatory power); Verbal Marked Choice, Article Marked Choice, Verbal Marked Construction, Lexical Marked Absence, Article Marked Presence, Lexical Marked Presence, Conjunction Marked Absence, Adverb Marked Presence, Pronoun Marked Choice, and Pronoun Marked Presence. These can be seen in Table 2 below with the along with the Wald score.

For Study Two the features were slightly different, but there was significant overlap in the features that comprised the optimum model. For Study Two the optimum model included 12 features, of which 7 were also present in the model for Study One, and five were new. The full table representing both the models can be seen below in Table 2.

Table 2 Optimum Model Features for Studies One and Two

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Rank** | **Study One** | **Wald** | **B** | **Study Two** | **Wald** | **B** |
| 1 | Verbal Marked Choice | 4.101 | 1.727 | Conjunction Marked Absence | 3.535 | 4.364 |
| 2 | Article Marked Choice | 3.355 | 2.628 | Pronoun Marked Presence | 2.463 | 6.313 |
| 3 | Verbal Marked Construction | 1.629 | -1.058 | Preposition Marked Absence | 2.134 | -1.632 |
| 4 | Lexical Marked Absence | 0.853 | 1.355 | Lexical Marked Choice | 2.109 | 0.242 |
| 5 | Article Marked Presence | 0.84 | 1.161 | Lexical Marked Absence | 0.852 | -2.52 |
| 6 | Lexical Marked Presence | 0 | 26.623 | Lexical Marked Construction | 0.83 | 0.103 |
| 7 | Conjunction Marked Absence | 0 | -53.241 | Lexical Marked Presence | 0.657 | -1.691 |
| 8 | Adverb Marked Presence | 0 | -74.842 | Verbal Marked Construction | 0.385 | -0.383 |
| 9 | Pronoun Marked Choice | 0 | 80.921 | Pronoun Marked Absence | 0.352 | 0.385 |
| 10 | Pronoun Marked Presence | 0 | -16.168 | Verbal Marked Choice | 0.053 | 0.073 |
| 11 |  |  |  | Conjunction Marked Presence | 0 | -40.801 |
| 12 |  |  |  | Adverb Marked Absence | 0 | -26.574 |

# Application of statistical models

The method proposed can be demonstrated through a worked example using new data (not used in building the model). Below is a very short extract from a blog which has been coded by the researcher for the features in the optimum models. These features are marked using the feature number between \* symbols (the features and their corresponding numbers can be seen below).

My name is [name] but my friends call me [name], I am a student at the University of [city] where I \*265\* studying in the Faculty of Law and Political Sciences. My professor is [first name] [surname]. 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 English better. I will update this blog \*122\* site often. (2011)

The text contains the following features:

* 251 Verbal Marked Choice x3 (Study One & Two)
* 265 Verbal Marked Construction x2 (Study One & Two)
* 122 Lexical Marked Presence x1 (Study One & Two)

This straightforward mark-up can be combined with the model from the logistic regression where

Likelihood of group Membership = (B value of feature for specific study x number of occurrences) + (B value of next feature x number of occurrences) [...]

For Study One the likelihood of the author having Persian influence can be calculated.

**Study One** 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 than L1 English*

Even though the text is very brief, the author’s non-native features can be used to identify them as a potential L1 Persian speaker. Champod and Evett (1999) propose a verbal equivalent scale for interpreting such likelihood measures in a forensic science context and against this scale this conclusion constitutes ‘moderate’ evidence, for the writer being an L1 Persian speaker.

**Study Two** likelihood of L1 author belonging to the group labelled ‘other the languages’ = (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 than a speaker of the related languages* which were modelled. Based on Champod and Evett’s scale this amounts to ‘limited’ evidence to support this assertion.

Through this worked example we can demonstrate a transparent human analyst driven version of NLID and at each stage of the analysis the reasoning can be interrogated for linguistic meaning.

# Discussion

What has been demonstrated in this study is how an approach to NLID which focuses on linguistically interpretable interlingual features is a viable alternative, or useful addition, to black-box computational systems. Using a logistic regression model built from Persian blog posts we are able to provide an informed analysis of a further potentially queried text. The features coded rely on human analytic skills which might be automated, but the method used allows a human analyst to code a brief non-standard text to arrive at a conclusion which is explicable in terms of the individually identified features. Where results are inconclusive or disputed, the features themselves can be examined and discussed and their contribution to the conclusion drawn is also explicit.

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 can be seen that a greater amount of text will allow for more features to contribute to the prediction and provide a greater weight of evidence towards different conclusions. Avoiding a black-box approach is particularly important in a forensic context where an analyst needs to consider great variability within the data available across different cases.

The approach here can be developed in a number of ways:

In addition to the two studies discussed in this article a third study addressed the question of whether an anonymous author pretending to be a native Persian speaker would intuitively use the right features in their disguised language (and hence discredit the findings in Studies One and Two), so that this system of NLID would falsely conclude that they are an L1 Persian speaker. The results indicated that this was not the case. When looking at the sample text above, the model indicated it was not likely to be an author disguising their language. This supports Shuy’s (2001) assertion that it is difficult to disguise one’s language, and an anonymous writer attempting to disguise their linguistic characteristics usually focuses more on the ‘conscious aspect of language use rather than the major features analysed in the linguistic profile’ (Shuy, 2001, Loc. 10105). This adds strength to the conclusion that the features identified in Studies One and Two (as presented here) are good indicators of an L1 influence from Persian.

In addition, using this method moves away from a traditional identification task. It will be possible to analyse a single text and conclude the likelihood that it contains influences from more than one language. Work in this direction is already indicated by Study Two but fuller exploration of this problem is on-going. In particular it is important to understand features which might co-predict more than one language and features which provide evidence against a particular language influence. As stated previously, this is part of a wider series of projects in which we are addressing a range of questions, one of which is potential for automation. This is intended to bridge the gap between traditional computational approaches to NLI (Native Language Identification) and sociolinguistically informed NLID (Native Language Influence Detection).

# Bibliography

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* 513--546. Oxford: Blackwell Publishing Ltd.

Brooke, J., & Hirst, G. (2012). Robust, lexicalized native language identification. In Proceedings of *COLING 2012* (391--408). Mumbai: The COLING 2012 Organizing Committee. Retrieved in October 2016 from <http://www.aclweb.org/anthology/C12-1025>

Brooke, J., & Hirst, G. (2013). Using Other Learner Corpora in the 2013 NLI Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* 188--196. Retrieved in October 2016 from <http://www.aclweb.org/anthology/W13-1725>

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*, 6:228--241.

Comrie, B. (2001). Languages of the World. In M. Aronoff & J. Rees-Miller (Eds.), *The Handbook of Linguistics* 19--43. Oxford: Blackwell Publishing Ltd.

Corder, S. P. (1981). *Error Analysis and Interlanguage*. Oxford: Oxford University Press.

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

Davies, A. (2003). *The native speaker: myth and reality*. Clevedon: Multilingual Matters Ltd.

de Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining e-mail content for author identification forensics. *ACM SIGMOD Record*, 30(4): 55--64.

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

Dras, M., & Malmasi, S. (2015). Multilingual native language identification. *Natural Language Engineering*, 1(1):1–53.

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

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

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

Grant, T. (2010). Text Messaging Forensics: Txt 4n6: Idiolect free authorship analysis? In M. Coulthard & A. Johnson (Eds.), *The Routledge Handbook of World Englishes* 508--522. London and New York: Routledge.

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

Jaleh. (2011). Jamigen’s Iranian Affairs Blog Site. Retrieved December 2012 from <http://jamigen.com/index.htm>

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

Koppel, M., Schler, J., & Argamon, S. (2009). Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology*, 60(1), 9--26.

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

Lado, R. (1957). *Linguistics Across Cultures.* Ann Arbor: University of Michigan Press.

Leung, C., Harris, R., & Rampton, B. (1997). The Idealised Native Speaker, Reified Ethnicities, and Classroom Realities. *TESOL Quarterly*, 31(3), 543--560.

Li, B. (2013). Recognizing English Learners’ Native Language from Their Writings. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* 119--123. Retrieved in October 2016 from <http://www.aclweb.org/anthology/W13-1715>

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

Malmasi, S. (2016). Native Language Identification : Explorations and Applications (PhD Thesis) Macquarie University. Retrieved from <https://www.researchonline.mq.edu.au/vital/access/services/Download/mq:50040/SOURCE1?view=true>

Malmasi, S., & Dras, M. (2014). Language transfer hypotheses with linear SVM weights. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP ’14), (2013)*, 1385--1390. Retrieved October 2016 from http://www.aclweb.org/anthology/D14-1144

Rampton, M. B. H. (1990). Displacing the “native speaker”: expertise, affiliation, and inheritance. I 44(2), 97--101.

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

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

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

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

Tetreault, J., Blanchard, D., & Cahill, A. (2013). A report on the first native language identification shared task. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications,* 48--57. Retrieved September 2016 from http://aclweb.org/anthology/W/W13/W13-1706.pdf

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

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.* 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* 9--17. Madison: Omnipress.

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

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

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* 1600--1610. Association for Computational Linguistics: Edinburgh.

Wong, S. J., Dras, M., & Johnson, M. (2011). Topic Modeling for Native Language Identification, In *Proceedings of Australasian Language Technology Association Workshop* 115--124.

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)* 53--62. Sydney. Retrieved October 2017 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.