# PhD Abstract

## Linguistic identifiers of L1 Persian speakers writing in English. NLID for authorship analysis.

**Ria Perkins**

Aston University

**Research and Teaching Associate**
**Centre for Forensic Linguistics**
**Aston University**
**Birmingham**
**UK**

**Awarding Institution**:
School of Languages and Social Sciences, Aston University, UK

**Date of award**: 2013

**Keywords**: Native language identification (NLID), authorship analysis, forensic linguistics, Persian, weblogistan, interlanguage.

This dissertation presents research on the linguistic identifiers of native (L1) Persian speakers writing weblogs in English, as a contribution towards Native Language Identification (NLID). NLID is a specific area of authorship profiling that focuses on identifying an anonymous author's native language. This research investigates what are the distinctive features of the language of a native L1 Persian speaker writing in English. It also focuses on 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.

Native Language Identification (NLID) is an understudied area of forensic linguistic authorship analysis, yet an area that holds considerable practical potential (Koppel *et al.*, 2005). The potential usefulness is even more significant when we consider that the majority of the world's population is bilingual (Thomason, 2001) and that English is one of the most widely spoken second languages with up a quarter of all people having some degree of competence in English (Bhatia and Ritchie, 2004: 519). 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 features in the ransom notes to hypothesis that the texts were most likely authored by a native German speaker. More recently cases documented by Kniffka (1996) and Hubbard (1996) involved degrees of NLID and demonstrate the potential of NLID as a tool for forensic authorship analysis. The data for this research comprises several corpora of internet blogs, this has many benefits for research from a forensic linguistic perspective, the most significant in this situation is 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 for 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 kind of forensic texts which may later benefit from the application of NLID analysis.

The objective of this research was to analyse and investigate the linguistic features of an L1 Persian speaker blogging in English, and to develop an implementable model that would form a useful tool for forensic authorship analysis. This can be broken down into the following six research aims:

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 identify specific linguistic choices as being indicative of influence 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)

These objectives were realised through three sub-studies; the first study devised a coding system to account for the interlingual features identified in a corpus of L1 Persian speakers blogging in English, and in a corpus of L1 English blogs. The second study looked at the features identified in Study One, with relation to other, related languages, namely; Azeri and Pashto, using collected blog data. The third study sought to determine if the features identified could distinguish between genuine L1 Persian authors and authors

attempting to disguise their language. It used elicited data from a questionnaire and a writing task. The final section considered the application of the results of the studies and developed an implementable model. Unlike previous research, this project focused predominantly on blogs, as opposed to student data, making the findings more appropriate to forensic casework data.

In summary this research showed that NLID is possible and can provide a valuable, reliable tool for forensic authorship analysis. The basic finding from Study One is that Native Language Identification (NLID) is possible, and that it can distinguish between texts produced 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. 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, and which combination of features formed the optimum set. Study Two compared the corpus of L1 Persian authors with a corpus of blogs by L1 Azeri and L1 Pashto speakers and demonstrated that it was possible to use the features to determine group membership of the authors, as well as determining which combination of features was the most discriminatory between the groups. Study Three distinguished between a genuine L1 Persian author and an author who was attempting to disguise their language to give the false impression that they are an L1 Persian speaker and determined that there was a clear difference between the groups. The implications for forensic authorship analysis are significant and this PhD forms part of a continuing study into this previously under-researched area.

## References

Bhatia, T. K. and Ritchie, W. C. (2004). Bilingualism in the global media and advertising. In T. K. Bhatia and W. C. Ritchie, Eds., *The Handbook of Bilingualism*. Oxford, UK: Blackwell Publishing.

Hubbard, E. H. H. (1996). Errors in court: A forensic application of error analysis. In H. Kniffka, S. Blackwell and M. Coulthard, Eds., *Recent Developments in Forensic Linguistics*, 123–140. Frankfurt am Main: Peter Lang.

Kniffka, H. (1996). On forensic linguistic "differential diagnosis". In H. Kniffka, S. Blackwell and M. Coulthard, Eds., *Recent Developments in Forensic Linguistics*, 75–122. Frankfurt am Main: Peter Lang.

Koppel, M., Schler, J. and 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, NY, USA: ACM.

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