

# Native Language Identification for Forensic Authorship Analysis



**Dr Ria Perkins** completed her PhD in 2013 at Aston University. Her thesis focused on **interlingual identifiers** of L1 Persian speakers blogging in English. Since finishing her PhD Ria is moving on to research linguistic structures in English and Swedish language texts of online extremism and radicalisation. She is now working at Aston University as a Teaching Associate.

## Context

Grounded in Forensic Authorship analysis, this research focused on online blogs from weblogistan to identify interlingual features of L1 Persian speakers writing in English and to develop an implementable model for forensic authorship cases. The potential importance of Native Language Identification (NLID) can be witnessed through the prevalence of multilingualism as well as documented cases such as those of Kniffka (1996 & 2000) and Hubbard (1996). Despite potential practical applications research into NLID from a forensic perspective is limited ; a void which this research seeks to fill.

Name	In Folder	Created On	Created By	Modified On	Modified By
Arrested of 'er'	Free Nodes	28/05/2011 11:27	R	28/05/2011 11:27	R
Abbreviation	Free Nodes	25/05/2011 11:34	R	25/05/2011 11:34	R
Adjective string construction	Free Nodes	06/06/2011 11:54	R	06/06/2011 11:54	R
ambidextrous	Free Nodes	06/06/2011 12:07	R	06/06/2011 12:08	R
'cause in place of because'	Free Nodes	24/05/2011 15:08	R	30/05/2011 16:27	R
Conjunction between 't' and 'y'	Free Nodes	26/05/2011 11:06	R	26/05/2011 11:06	R
Eller string article in non string instead of plur	Free Nodes	24/05/2011 15:03	R	24/05/2011 15:03	R
Less influence by homophone	Free Nodes	16/05/2011 11:26	R	16/05/2011 11:26	R
emba word	Free Nodes	24/05/2011 15:10	R	25/05/2011 12:11	R
Emba word 'u'	Free Nodes	12/04/2011 14:42	R	13/05/2011 14:47	R
Emba word 'u'	Free Nodes	13/05/2011 15:04	R	13/05/2011 15:04	R
Grammatical	Tree Nodes	14/04/2011 14:28	R	14/04/2011 14:28	R
Grammatical/Additional Article	Tree Nodes	14/04/2011 14:27	R	16/05/2011 13:29	R
Grammatical/Additional Article/Additional article 'u'	Tree Nodes	14/04/2011 14:27	R	31/05/2011 17:13	R
Grammatical/Additional Article/Additional article 'u'	Tree Nodes	14/04/2011 14:27	R	25/05/2011 11:37	R
Grammatical/Adjective creation based on overlap	Tree Nodes	24/05/2011 16:22	R	24/05/2011 16:22	R
Grammatical/'er' instead of 'r'	Tree Nodes	31/05/2011 14:06	R	06/06/2011 12:14	R
Grammatical/'er' instead of 'er' or 'r'	Tree Nodes	31/05/2011 14:06	R	31/05/2011 14:06	R
Grammatical/'er' instead of 'er'	Tree Nodes	31/05/2011 14:06	R	31/05/2011 14:06	R
Grammatical/Article construction	Tree Nodes	24/05/2011 16:25	R	24/05/2011 16:25	R
Grammatical/Article use	Tree Nodes	16/05/2011 10:14	R	16/05/2011 10:14	R
Grammatical/Emba making a plural	Tree Nodes	27/04/2011 14:47	R	27/04/2011 14:47	R
Grammatical/Emba word 'er'	Tree Nodes	24/05/2011 16:25	R	24/05/2011 16:25	R
Grammatical/Missing article	Tree Nodes	24/05/2011 16:25	R	24/05/2011 16:25	R
Grammatical/Missing article 'u' (or 'er')	Tree Nodes	27/04/2011 14:46	R	30/05/2011 13:42	R
Grammatical/Missing article 'er'	Tree Nodes	16/05/2011 10:29	R	06/06/2011 12:42	R
Grammatical/Missing negative	Tree Nodes	16/05/2011 10:13	R	16/05/2011 10:13	R
Grammatical/Missing negative 'er'	Tree Nodes	27/04/2011 14:47	R	27/04/2011 14:47	R
Grammatical/Missing positive	Tree Nodes	16/05/2011 10:34	R	06/06/2011 12:00	R
Grammatical/'er' instead of 'er'	Tree Nodes	27/04/2011 14:46	R	27/04/2011 14:46	R
Grammatical/That possessive when as a single pr	Tree Nodes	16/05/2011 10:19	R	16/05/2011 10:19	R
Grammatical/Single instead of plural	Tree Nodes	27/04/2011 14:47	R	28/04/2011 15:21	R
Grammar	Tree Nodes	24/05/2011 16:26	R	24/05/2011 16:26	R
Grammar/'er' instead of 'r'	Tree Nodes	24/05/2011 16:26	R	24/05/2011 16:26	R
Grammar/'er' instead of 'er'	Tree Nodes	24/05/2011 16:26	R	24/05/2011 16:26	R
Grammar/'er' instead of 'er'	Tree Nodes	24/05/2011 16:26	R	24/05/2011 16:26	R
Grammar/'er' instead of 'er'	Tree Nodes	24/05/2011 16:26	R	24/05/2011 16:26	R
Grammar/Emba word 'er'	Tree Nodes	16/05/2011 10:20	R	16/05/2011 10:20	R
'r' in place of 'r'	Free Nodes	30/05/2011 13:57	R	30/05/2011 13:57	R
Logical	Tree Nodes	14/04/2011 14:34	R	16/05/2011 11:57	R
Logical/anyone!	Tree Nodes	27/04/2011 14:38	R	27/04/2011 14:38	R

## Data and Analysis

3 sub-studies involving different corpora:

- L1 Persian and L1 English Weblogs
- Other languages (Azeri and Pashto)
- Disguise language

A coding system was developed to account for all the features. The data were then coded using NVivo. Logistic regression was used to determine which features had the higher discriminatory power. The features were then refined to discover the optimum models.

## Findings

This research demonstrated that interlingual features in L2 writing can be used to indicate an author's L1. It developed a new methodology for NLID and distinguished which features best indicate authorship by an L1 Persian speaker . The features identified were determined to have a statistically high level of reliability at determining group membership. An implementable model was constructed to help determine if an anonymous text was written by an L1 Persian speaker, as well as distinguishing from close languages and author's attempting to disguise their L1. NLID and the model created have a strong potential to be a valuable tool for forensic authorship analysis and may prove useful for criminal and intelligence investigations. It is intended as part wider research and there is clear potential for further research.

