Chapter 12
Native Language Identification (NLID) for Forensic Authorship Analysis of Weblogs

Ria Perkins
Aston University, UK

ABSTRACT

This chapter introduces Native Language Identification (NLID) and considers the casework applications with regard to authorship analysis of online material. It presents findings from research identifying which linguistic features were the best indicators of native (L1) Persian speakers blogging in English, and analyses how these features cope at distinguishing between native influences from languages that are linguistically and culturally related. The first chapter section outlines the area of Native Language Identification, and demonstrates its potential for application through a discussion of relevant case history. The next section discusses a development of methodology for identifying influence from L1 Persian in an anonymous blog author, and presents findings. The third part discusses the application of these features to casework situations as well as how the features identified can form an easily applicable model and demonstrates the application of this to casework. The research presented in this chapter can be considered a case study for the wider potential application of NLID.

1. BACKGROUND

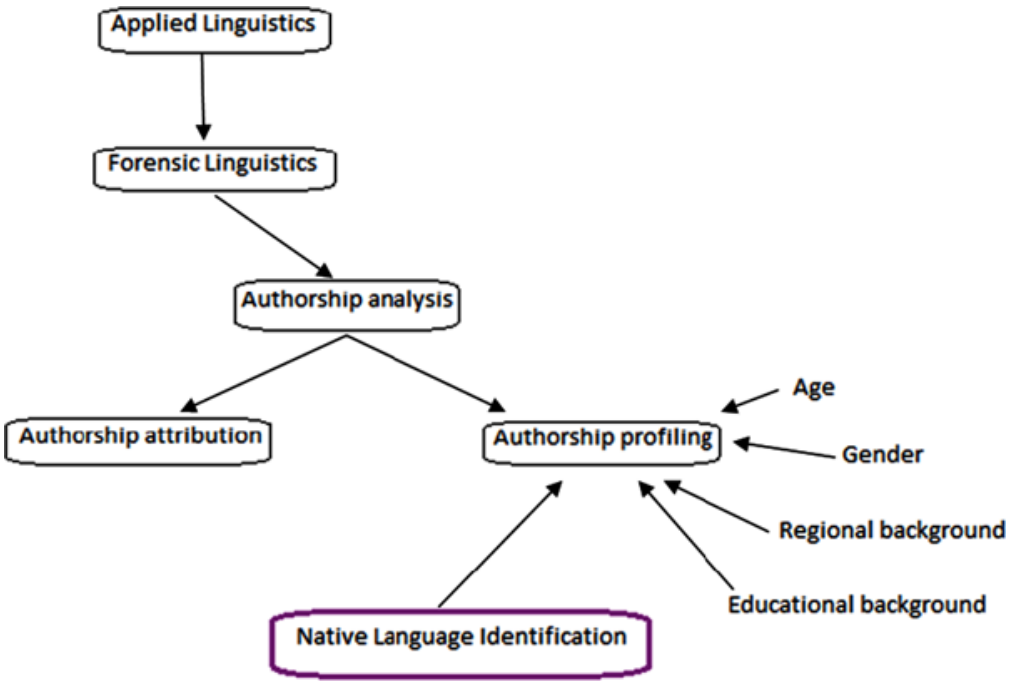
1.1 Introduction

Forensic Linguistics, Authorship Analysis and Native Language Identification

Native language identification is an aspect of forensic linguistic authorship analysis. Forensic linguistics is a branch of applied linguistics that relates to legal situations and processes. Loosely defined it relates to any overlap between language or linguistics, and legal or criminal situations. Coulthard, Grant and Kredens (2010) suggest that forensic linguistics, is application of linguistics in three main areas; "writ-

DOI: 10.4018/978-1-4666-8345-7.ch012

Figure 1. Authorship profiling



ten legal texts, spoken legal practices and the provision of evidence for criminal and civil investigations and courtroom disputes" (Coulthard et al., 2010, p. 529). For this research forensic linguistics is often considered a section of applied linguistics rather than a completely separate discipline. Forensic linguistic consultants are predominantly linguists, who apply their knowledge to forensic situations or contexts.

Native language identification (NLID) is a very specific question within the wider field of forensic linguistics, falling under the area of authorship analysis which relates specifically to the last of the three main areas of forensic linguistic interest that were identified by Coulthard, Grant and Kredens (2010); that being the provision of evidence. Authorship analysis has two main sub-areas; profiling and authorship attribution, of which native language identification belongs to the second (Figure 1 below shows the location of NLID to forensic linguistics and its sub-fields).

Authorship attribution seeks to answer which person out of a fixed group of suspects is the most likely author of an anonymous document. This usually requires a closed group of suspects and a comparison between the questioned or disputed texts and texts that are known to have been written by each suspect. A typical case might involve threatening emails being sent to a company, the content of which might indicate the author is most likely an employee of the company. The forensic linguist could then compare linguistic features within the threatening emails (questioned texts) and compare them to emails known to be authored by each of the employees (known texts), to determine which employee is the most likely author. Authorship profiling is the other area of authorship analysis, it is similar in many ways, but does not require a closed pool of suspects. A famous case of authorship profiling is the devil strip case (Leonard, 2005). Linguist Dr Roger Shuy was consulted on a case regarding a kidnapped girl whose parents had received a ransom note reading: