The Application of Forensic Linguistics in Cyber Crime Investigations.

Forensic Linguistics

Forensic linguistics can be broadly defined as the study or analysis of language in legal settings (Kniffka, 2007; Rock, 2006). It is predominantly a sub-field of applied linguistics, in which linguistic knowledge, analysis and methodologies are applied to forensic and criminal situations. Svartvik (1968) was one of the earliest academics to call for forensic linguistics to be considered as a distinct field (Perkins & Grant, 2013). In 1965-1966 he applied existing linguistic knowledge to a series of statements of disputed authorship. Using qualitative and quantitative analysis he demonstrated that there were inconsistencies in the language used across the statements, and importantly, within the grammar of the incriminating sections. Through this he also demonstrated that applied linguistics (and particularly sociolinguistics) can contribute beyond the traditional realms of language teaching and machine translation, and be of use in forensic or criminal contexts too.

Forensic Linguistics began to develop an identity as a distinct field in the UK in the 1980s and 90s with the cases of Professor Malcolm Coulthard, the most famous of which was the Birmingham Six appeal. In 1993, the International Association of Forensic Linguists (IAFL) was established. Forensic Linguistics is now largely recognised as its own distinct field; it has spread around the world, broadening in scope and becoming recognised and utilised in a variety of jurisdictions and contexts.

Cybercrime relies very heavily on text based communication; in fact 'most forms of abuse online manifest textually' (Williams, 2001, p. 164). The growth and popularity of electronic and social media means that there are now many new opportunities for collecting evidence or data, benefiting both investigators and forensic linguists (Bhatia & Ritchie, 2013). Forensic linguists have been working with emerging technologies from cases involving phone SMS messages to more recent cases involving tweets and forum messages. It would be impossible to cover all the areas in which forensic linguistics can contribute to cybercrime investigation; this is in part because both fields are constantly evolving. This article will introduce some of the key areas where forensic linguistics has been documented to be of use, as well as discussing how future collaboration might be of benefit for all parties. It also presents findings from a research study on Native Language Influence Detection (NLID); showing that NLID is possible through a sociolinguistic explanation based approach, and indicating which features are of particular interest when considering native (L1) Persian speakers writing online in English. Moreover it also serves to demonstrate how linguists can contribute to developing systems that can have practical applications for cybercrime casework.

The majority of existing forensic linguistic work relates to three broad categories: written legal language (for example analysis of how PACE instructions are interpreted and understood), spoken legal language (such as analysing power in interviews), or investigative linguistics and the provision of evidence (Coulthard, Grant, and Kredens, 2011). It is this third category that is most closely allied to work done in relation to cybercrime investigations. Within the area of investigative linguistics and the provision of evidence, there are a variety of different tasks that forensic linguists perform; these include: comparative authorship analysis, sociolinguistic profiling, interactional meaning, determining meaning, trademark disputes and copyright infringement.

Comparative authorship analysis is usually a closed set analysis in which a text of anonymous or disputed authorship is credibly believed by investigators to be written by one of a limited number of authors. Forensic linguists can then compare the linguistic style and features of the questioned text to known texts by the suspect author or authors. Comparative authorship of long texts is increasingly dependent on heavily multivariate computational techniques, which can be shown to be reliable but offer little explanation as to the outcome. This validity deficit means that forensic analysts tend not to depend on such techniques and, in any case, such techniques often require more text than is available in forensic casework (Grant, 2007). Perhaps surprisingly, considerable progress in forensic comparative authorship analysis has been made with the very short texts found in SMS text messaging and other short form messages such as Twitter feeds. There have been a number of UK cases when a person is missing, presumed dead, but their mobile phone has continued to send text messages. In such cases, linguists have been consulted to see if the suspect messages are consistent with those of the missing person, the suspect, or neither (see Grant (2010) for a description of one such case and the analysis performed).

Some crimes are inherently linguistic in that they are committed through language, for example: threatening, extorting, and bribing. Shuy (1996) termed these ‘language crimes’ (also discussed by Solan & Tiersma, 2005). In his work, Shuy (1996, 2005) demonstrates that covertly recorded conversations involving an undercover agent can make for poor forensic evidence of what was said and what was meant. He demonstrates how the imbalance in knowledge between the participants in the conversation can warp interpretation of the communications, leading to prosecutions on the basis of linguistically questionable evidence. The role of forensic linguists and linguists in determining meaning is perhaps more apparent when considering multilingual texts; but even within monolingual situations, a forensic linguist can have much to offer, particularly when slang is involved. Grant (2017) identifies four main roles a linguist can have when seeking to determine slang meaning, with each role or situation requiring a different combination of methodologies. An example of one variety is Grant’s work in a conspiracy to murder case (Coulthard, Grant, & Kredens, 2011; Grant, 2017), which took place over internet relay chat (IRC).The suspects were Grime musicians that spoke Multicultural London English, a variety of East London slang which draws heavily on Jamaican English. One key phrase from the IRC chat transcript was ‘I’ll get da fiend to duppy her den’. In this instance Grant was able to explain to the Court the origin and the meaning of the verb ‘to duppy’ (which can be traced back to Jamaican English and its approximate meaning of ‘ghost’) and that it did indeed indicate a threat against the victim.

Sociolinguistic profiling is directly descended from the field of sociolinguistics and is based on the concept that an individual's linguistic output is influenced by a number of social factors including age, gender, geographical background, other languages spoken, and educational status. In sociolinguistic profiling casework, the forensic linguist will aim to determine information about an anonymous author or the origins of the text. A linguist may not make psychological observations about the author or their intentions but, dependent on the features within the text, they might be able to describe the author's social origins or background. Sociolinguistic profiling has been used extensively with computer mediated communications, and there have been numerous documented cases of it being beneficial to the outcome of a case and the provision of justice (Kniffka, 1996; Leonard, 2005; Schilling & Marsters, 2015). Conclusions about the likely social background of an anonymous author are unlikely to ever be certain enough to provide evidence for courtroom use, but as evidenced through previous casework, they can be used investigatively to good effect.

Native Language Influence Detection

One area of sociolinguistic profiling that is of increasing interest and that holds much potential for impacting law enforcement work is native language influence detection (NLID) (Dras & Malmasi, 2015; Grant, 2008; Koppel, Schler, & Zigdon, 2005; Li, 2013; Malmasi, 2016; Tetreault, Blanchard, & Cahill, 2013). A simplified definition of NLID is that it seeks to indicate an author's native language, also termed L1, from the way they write in a second language (or L2). As multilingualism is becoming increasingly prevalent and there are now more multilingual than monolingual speakers in the world (Thomason, 2001), application of NLID holds much potential benefit. While it is difficult to define exactly what level of expertise is required for someone to be considered a speaker of a second language, it is estimated that the number of second language (L2) English speakers could outnumber the number of native English (L1) speakers (Bhatia & Ritchie, 2004). Unsurprisingly, this trend continues online, with approximately 80% of the 40 million internet users communicating in English (Bhatia & Ritchie, 2013). It is therefore logical to conclude that a considerable number of English language forensic texts are likely to be produced (or at least potentially produced) by non-native English speakers. Bhatia and Ritchie (2013) highlighted the growing link between computer mediated communication, multilingualism and forensic linguistics, stating 'In a world connected by social media and globalization, the role of the study of multilingualism in forensic linguistics is increasing rapidly.'(Bhatia & Ritchie, 2013, p. 672).

There is an established social belief that one can identify a person's L1 from the way they use a second language, and the link to potential forensic application is not new. A similar concept can be seen in the Bible with the Gileadites using the term 'Shibboleth' to distinguish whether a person was a Gileadite or an Ephraimite based on their pronunciation of the first phoneme. It can also be witnessed through fictional literature, in a Scandal in Bohemia (Doyle, 1892), Sherlock Holmes uses interlanguage principles and the positioning of a verb to identify that the author of an anonymous note is a native German speaker. Whereas Parker Kincaid, Jeffery Deaver's (1999) fictional forensic document expert, uses linguistic typologies to determine that an anonymous author is merely pretending to be a non-native English speaker, as the features do not indicate a specific language.

There are few real cases involving NLID that have been publicised, likely due to the sensitive situations surrounding them. Two real life cases that involve forensic linguistics have been documented by Kniffka (1996) and Hubbard (1996). Kniffka discussed a case in which he was consulted about threatening letters being sent within a German company. The content indicated that the anonymous author was one of the company's employees. Kniffka's analysis uncovered occurrences of marked linguistic constructions of the German language including; unusual spelling errors with umlauts, awkward lexical collocations and non-idiomatic use of German proverbs. He concluded that the author was likely a non-native German speaker with a high level of German fluency. This information fed into the investigation with police changing their focus from an L1 German suspect, to the two L2 German employees, one of whom was later found writing another threatening letter.

The field of NLID is strongly influenced by the concepts of interlanguage and cross-linguistic influence which developed from second language acquisition studies from a pedagogic perspective. In this field, researchers, for example Lado (1957) and Hopkins (1982), indicated that an understanding of a learner's first language (L1) and their target or second language (TL or L2) can be used to predict the errors they might make. Similarly after successfully using linguistic analysis to aid in a prosecution on a South African case involving the questioned authorship of a series of extortion letters and an L1 Polish speaking suspect, Hubbard (1996) concluded that 'error analysis can have forensic value' (Hubbard, 1996, p. 137). Although these areas have different motivations to NLID, and NLID is interested more in general linguistic patterns than errors, they still set up a theoretical precedence.

Native Language Identification (NLI) is a very closely related field to Native Language Influence Detection (NLID), approaching the same question of indicating an author's native language, but from a computational perspective. The field of NLI was pioneered by computational researchers such as Tomokiyo & Jones (2001), Jarvis, Castaneda-Jiménez, & Nielsen (2004), and Koppel, Schler, & Zigdon (2005). Koppel et al. (2005) in particular have been taken as the standard for future research.

Koppel et al. drew their data from the ICLE corpus (International Corpus of Learner English), which comprises classroom essays on common topics across the different language sub-corpora. The use of language student data has been replicated by many other studies. Malmasi (2016) noticed a trend emerging in 2012 for research using data other than from the ICLE corpus; the motivation seemed mainly to prevent topic bias, rather than to better mimic forensic data as the majority of studies still focused on data from second language learners. In keeping with this, the majority of new data sets were still based on language learner texts. In a 2013 shared task on NLI (Tetreault et al., 2013), the majority of the participating teams based their work on the TOEFL11 corpus test data (Blanchard, Tetreault, Higgins, Cahill, & Chodorow, 2013). Those that found other data used other corpora of English learners, arguably the most interesting being the use of the Lang-8 (www.lang-8.com) corpus by (Brooke & Hirst, 2013). Lang8 is an online learning resource where users post diary journal entries which are then corrected by native speakers of the language. This is potentially more valid data for the development of forensic and intelligence applications, as much forensic data is also produced online. However the purpose and audience are still firmly grounded in the

language-learning domain. While there is little consistency in what constitutes a forensic text, they rarely resemble student texts which are written in unique conditions and for the purpose of evaluating the author's language, rather than for communicating content.

In contrast to most existing NLI studies, the present NLID study uses real life data collected from weblogs. This contrasts with elicited data or student data which is written for a particular purpose. NLID takes a data-driven bottom-up approach using sociolinguistic explanations to indicate which languages might have an influence over the language that is being analysed. This allows a more nuanced perspective of why certain features are important and indicative, and means that the analyst can better explain what is happening and the analysis better adapt to genre changes as well as authors with more complicated linguistic histories. Some NLI studies do incorporate interlingual explanations of the features, most notably Brooke & Hirst (2012) and Bykh, Vajjala, Krivanek, & Meuers (2013). The latter used linguistically informed features, rather than just surface focused n-grams, to develop a more accurate system in the 2013 NLI shared task. This demonstrated that linguistic explanations can make automatic analytical systems more accurate, as well as being of use to the human analyst.

NLI also tends to take a closed-set approach to the problem; identifying an author's language from a limited set of options. As there are approximately 7,000 languages in the world (Simons & Fennig, 2017), and the majority of speakers have contact with more than one language, a closed set approach is of limited use to forensic profiling. NLID (Native Language Influence Detection) and OLID (Other Language Influence Detection) are more focused on influence, using explanations to indicate potential influences on an anonymous author's language, and allowing for more complex linguistic backgrounds, which many authors and potential authors have.

Study - Methodology

This article presents findings from two studies within a wider series of studies: the first looks at features of native Persian speakers writing online, as opposed to native English speakers; The second was a sub-study to indicate whether the features identified were indicative of Persian, as opposed to non-native speech or the wider Persian language family. For this purpose, it analysed weblogs from L1 authors of two related languages: Azeri and Pashto. The study does not analyse these languages fully; instead they are used as a scoping study to determine if the features identified as indicating L1 Persian influence can distinguish between languages that are geographically or linguistically close to Persian.

The data comprises publically accessible weblogs from authors writing in English. The data for the main study contains two corpora; one of weblogs written by native Persian speakers and a control corpus of weblogs by L1 English speaking authors. The blogs were collected from a range of sources and cover a variety of topics. 25 authors who self-identified as being an L1 (or mother-tongue or native) speaker of the relevant language without apparent self-contradiction were selected for each corpus. The data for the second smaller study into related languages, comprises 5 L1 Azeri authors and 5 L1 Pashto authors blogging in English. The L1 Persian corpus from the main study served as the control corpus.

Feature identification was based on a data-driven approach. Initial analysis was undertaken through a close analysis of a sub-set of the data, any occurrence of marked language was noted. Marked language in this case is the use of language in a way that an L1 English speaker would unlikely do; this includes errors and grammatically correct (but unusual) preferences.

It was noticed that the marked language clustered around certain features. Many of these features had clear interlingual explanations. The features were loosely grouped into hierarchical categories, of which this article is focusing on the mid-level features. There are both higher level features which reflect the broad grammatical class of the feature (e.g. preposition, ordering and positioning, or lexical), and lower level features that contain more specific information about the marked language, how marked it is, and potential influences. Under the higher level categories the features clustered around several areas within each category: *Marked Presence*, *Marked Absence, Marked Choice, or Marked Construction.* In some situations *Marked Position or Ordering* was also a possibility. This resulted in the 29 mid-level features including: *Verbal Marked Choice, Article Marked Choice, Pronoun Marked Choice, and Pronoun Marked Presence*.

After the full feature set was established, the entire data set was coded (with lower level specific descriptive features being fitted into the framework as they appeared). A total of over 300 features across the levels were identified, many of which were very precise lower level descriptors.

Findings and Application

After the blogs were coded for all of the features, logistic regression analysis was used to determine which of the mid-level features had the highest discriminatory power. Logistic regression is a statistical analysis that has been demonstrated to be a useful tool in forensic analysis and criminal justice (Weisburd & Britt, 2007). It predicts the outcome of a situation, based on a set of variables, which in this case are the 29 mid-level features. The potential outcomes for the first study are that any given author could belong to the L1 Persian speaker set, or the L1 English speaker set. In the second sub-study the potential outcomes are that the author could belong to the L1 Persian speaker set, or the group of authors from the closely related languages (L1 Azeri and L1 Pashto).

Initially, all the 29 mid-level features were used, but given the high number of variables, it is not surprising that the Hosmer-Lemeshow test indicated that the model was over-fitted to the data and hence would not be generalizable or adapt well to other data. This is a particular issue in the forensic context where there is no standard genre of forensic texts; instead there is great variability in what a forensic text can look like. The number of variables included in the model can be reduced by eliminating features with lower predictive power as identified through low Wald X^2 scores, to find an optimal model that is a balance between good prediction and over-fitting.

The features that comprise the optimum models for each study can be seen in Table 1 below. For Study 1, the optimum model contained 10 features as follows (in order of descending 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. The optimum model for Study 2 contained 12 features, of which five were different to Study 1 and seven repeated. The statistical output also gives us the *B* value for each feature (see Table below), which relates to how much the presence of that particular feature alters the probability of membership to each group. The polarity indicates which group the feature relates to. A positive B value in Study One increased the probability that the author belonged to the second group, that of L1 Persian speakers. In Study Two, a negative B value increases the probability of L1 Persian authorship. In both studies, the following 3 features indicated an increased probability that the author belonged to the L1 Persian speakers group: Conjunction Marked Absence, Pronoun Marked Presence, and Lexical Marked Presence.

[Table 1 preferred location]

It is also possible to use these optimum models and the information contained in the table above to perform analysis in case-work situations. This information also enables much greater understanding of the features, avoiding a 'black-box' approach to analysis.

The optimum models have the added benefit that they are much more easily implementable, as the analyst can focus on these features in isolation, and hence only has to code for the seventeen distinct mid-level features, rather than over 300 features. This means that the analysis is practical for casework, which tends to be very time sensitive.

The application can be demonstrated as follows. Below is a short extract from an online blog that did not constitute part of the data for this research. In an ideal situation, the forensic linguistic analysis would use as much data as it is possible to gather. However, often there is not much data that the linguist is able to access. Coulthard (1994) estimated that forensic texts tend to be between 400 and 700 words in length. The increase of forensic data linked to computer mediated communication, such as text messages or tweets, means that even briefer texts are becoming more relevant to forensic contexts (Silva & Laboreiro, 2011). It is unlikely that data as short as the text would constitute good forensic data, but it serves as a useful example here.

> *My name is [Username] 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 [Full Name]. 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.* (Jaleh, 2011)

The features are marked in the text with *numbers* that correspond to the relevant features (see list below). The text contains the following features:

- *251* = Verbal Marked Choice x3 occurrence
- *265* = Verbal Marked Construction x2 occurrence
- *122* = Lexical Marked Presence x1 occurrence

Each study much be considered in turn. Firstly, the features relating to Study 1 can be input to the following equation to determine the probability that the text was authored 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 that the author belongs to the L1 Persian group = (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 demonstrates, that despite the reduced volume of text, the features indicate that the author is an L1 Persian speaker, which matches how the author self-identified within the blog. The likelihood ratio constitutes moderate evidence by the standards of Champod & Evett, (1999)'s scale.

Study 2 Likelihood that the author belongs to the L1 'other 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

This again is in keeping with the blogger's self-identification as an L1 Persian speaker. The reduced likelihood ratio only constitutes limited evidence on Champod and Evett's 1999 scale for evaluating likelihood scales as evidence. However, the fact that even with a very small section of text, the results are as expected, supports the reliability of the features and their use in forensic situations. It is likely that a greater volume of text would yield more features, and hence the weight of evidence might increase.

The study above indicates that native language influence detection is possible, it can be used by forensic linguists to indicate influences from other languages on an anonymous author's second language. It can distinguish between languages that are closely related (especially linguistically, geographically, or culturally). It can be useful with very short texts (such as one might find in forensic contexts). Previous research (Perkins, 2013) has also demonstrated that it is not easily susceptible to confusion by authors disguising their language.

The research presented above is a small element of a wider series of research projects that has been progressing at Centre for Forensic Linguistics at Aston University. There are many more questions that cannot be presented here that are being considered, and analysed. This includes questions around inter- and intra-rater reliability, considering more languages, distinguishing between related languages, as well as developing features that are computationally tractable and developing a semi-automated system that is grounded within interlingual explanations. Through this approach, the projects seek to develop a system that incorporates elements of computational NLI approaches, combined with the theoretical

grounding of NLID, to develop a system that will be of benefit to investigators and law enforcement agents.

What is clear is that as technology and globalisation enables more regular contact with different L1 speakers around the world, understanding the impact of cross linguistic influence and developing multilingual forensic linguistics systems of analysis will become more important. It is also likely that advances in technology will significantly influence this field, not just through the evolution of software available to law enforcement and researchers, but also with those seeking to evade detection, potentially with methods such as spoofing software. This in turn will necessitate further work to understand the implications and impact of this, as well as how it can be countered.

The author of this article has undertaken several cases which required elements of native language influence detection, sometimes to great effect. Casework experience has indicated that NLID is seldom used on its own, but as part of a wider profile, and hence it forms a useful tool in the sociolinguistic profiler's toolbox. It is expected that with continuing globalisation through technology, NLID will be of even more use in the future.

Discussion and Conclusions

This article has introduced some of the key areas in which forensic linguists are currently supporting cyber investigations, using NLID as an example to demonstrate how research motivated by casework situations can be of use. The cases on which forensic linguists consult on are needs-driven and reflect the socio-political climate. It is interesting to note that, increasingly, cases revolve around language data produced as part of computer mediated communication, as well as cases involving speakers of multiple languages. However, it would be erroneous to think that the work presented in this article represents the full extent of what is possible within the field of forensic linguistics. Cyber criminals are constantly evolving, leveraging technological developments and adopting new strategies and organisational structures across countries. This results in law enforcement agencies and officers having to adapt and evolve to keep up (Choo & Smith, 2008). Similarly, language evolves and changes in many ways (Heine & Kuteva, 2005) and it can do so very quickly (Keller, 1994). It is therefore unsurprising that the work of a forensic linguist is constantly evolving and growing. Technological advances have led to an increase in perceived and actual anonymity online, leading to a greater emphasis on fields such as forensic linguistics to help investigators (Hughes et al., 2008).

The key message of this article, and the main take home for police or investigative practitioners, is that an awareness of forensic linguistics, what it is and does, can be of great use in cybercrime investigations. An understanding of potential linguistic features can help investigators know when and how forensic linguistic analysis might be of use. The majority of forensic linguists who work with cybercrime are also academics, and as such they are involved with three main areas of work: casework, research, and teaching. It is the view of this article, that collaboration across all three areas will enhance the utility of forensic linguistics, meaning greater success in cybercrime investigations, and the delivery of justice.

One of the growing areas in which forensic linguists have been of use, is in the delivery and facilitation of linguistic training for investigators. Once such example is the linguistic training delivered as part of the Pilgrim course for police staff who work with online material (HM Inspectorate of Constabulary and Fire & Rescue Services (HMICFRS), 2014). Officers are trained to recognise different levels of linguistic feature, which in turn enhances their ability to assume different (and sometimes very specific) identities online (Grant & Macleod, 2016). This is an area that is having demonstrable investigative impact, thanks to research collaboration. Similarly the author of this article was seconded to a British Policing Unit as part of a previous forensic linguistic research project. As far as we are aware, this is the first situation of a forensic linguist being embedded into an investigative unit. Working in close proximity for an extended period enabled not just better results for that particular research project, but also an ongoing understanding in of each other's work that would not have been possible otherwise. It is the contention of this article that collaboration in any form is positive. Developing a working relationship has led to casework, further research, and training, enhancing the practices on both sides and ensuring long lasting impact.

It is impossible to accurately predict the future direction of forensic linguistics. What is clear is that through collaboration with casework, research, and training forensic linguists can continue to support those working to prevent and solve cybercrime. Similarly, through engaging with forensic linguists, investigators and law enforcement officers can enable us to better support investigations and the provision of justice, at case level and beyond.

Tables

Table 1 - Study 1 and 2 - Optimum Model Features

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

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* (pp. 513–546). Oxford: Blackwell Publishing Ltd.

Bhatia, T. K., & Ritchie, W. C. (2013). Multilingualism and Forensic Linguistics. In *The Handbook of Bilingualism and Multilingualism* (pp. 671–701). Malden & Oxford: Blackwell Publishing Ltd.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2013). *TOEFL11 : A Corpus of Non-Native English*.

Brooke, J., & Hirst, G. (2012). Robust , lexicalized native language identification. In *Proceedings of COLING 2012* (pp. 391–408). Mumbai: The COLING 2012 Organizing Committee. Retrieved 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* (pp. 188–196). Retrieved from http://www.aclweb.org/anthology/W13-1725

Bykh, S., Vajjala, S., Krivanek, J., & Meuers, D. (2013). Combinging Shallow and Linguistically Motivated Features in Native Language Identification. In *NAACL / HLT 2013 Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 197–206). Atlanta, Georgia.

Champod, C., & Evett, I. W. (2000). A. P. A. Broeders (1999) “Some observations on the use of probability scales in forensic identification”,. *International Journal of Speech Language and the Law*, *7*(2), 228–241.

Choo, K. K. R., & Smith, R. G. (2008). Criminal exploitation of online systems by organised crime groups. *Asian Journal of Criminology*, *3*(1), 37–59. https://doi.org/10.1007/s11417-007-9035-y

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. (2011). Forensic Linguistics. In B. Johnstone, R. Wodak, & P. Kerswill (Eds.), *The SAGE Handbook of Sociolinguistics* (pp. 529–544).

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.

Dras, M., & Malmasi, S. (2015). Multilingual native language identification. *Natural Language Engineering*, *1*(1), 1–53. https://doi.org/10.1017/S1351324915000406

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

Grant, T. (2017). Duppying yoots in a dog eat dog world, kmt: Determining the senses of slang terms for the Courts. *Semiotica*. https://doi.org/10.1515/sem-2015-0082

Grant, T., & Macleod, N. (2016). Assuming Identities Online: Experimental Linguistics Applied to the Policing of Online Paedophile Activity. *Applied Linguistics*. https://doi.org/10.1093/applin/amv079

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

HM Inspectorate of Constabulary and Fire & Rescue Services (HMICFRS). (2014). *An inspection of undercover policing in England and Wales*. Retrieved from https://www.justiceinspectorates.gov.uk/hmicfrs/wp-content/uploads/an-inspection-of-undercover-policing-in-england-and-wales.pdf

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.

Hughes, D., Rayson, P., Walkerdine, J., Lee, K., Greenwood, P., Rashid, A., … Brennan, M. (2008). Supporting Law Enforcement in Digital Communities through Natural Language Analysis , *5158/2008*, 122–134.

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

Jarvis, S., Castaneda-Jiménez, G., & Nielsen, R. (2004). Investigating L1 lexical transfer through learners' wordprints. In *Second Language Research Forum (SLRF)*. State College, PA.

Keller, R. (1994). *On Language Change: The Invisible Hand in Language*. London & New York: Routledge.

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. (2007). *Working in Language and Law*. Basingstoke: Palgrave Macmillan. https://doi.org/10.1057/9780230590045

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* (pp. 624–628). New York,: ACM Press. https://doi.org/10.1145/1081870.1081947

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

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

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}* (pp. 119–123). Retrieved from http://www.aclweb.org/anthology/W13-1715

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

Perkins, R.C., 2013. *Linguistic Identifiers of L1 Persian speakers writing in English . NLID for Authorship Analysis* . Aston University.

Perkins, R., & Grant, T. (2013). Forensic linguistics. In *Encyclopedia of Forensic Sciences* (Second Edi, pp. 174–177). Elsevier.

Rock, F. (2006). Looking the other way : Linguistic ethnography and forensic linguistics. *UK Linguistic Ethnography Forum Papers*, (May), 1–23.

Schilling, N., & Marsters, A. (2015). Unmasking Identity: Speaker Profiling for Forensic Linguistic Purposes. *Annual Review of Applied Linguistics*. https://doi.org/10.1017/S0267190514000282

Shuy, R. (1996). *Language crimes: the use and abuse of language evidence in the courtroom*. Oxford: Blackwell.

Shuy, R. (2005). *Creating language crimes: How law enforcement uses (and misuses) language*. Oxford: Oxford University Press.

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

Simons, G., & Fennig, C. (2017). How many languages are there in the world? *Ethnologue: Languages of the World, Twentieth edition*. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

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

Svartvik, J. (1968). *THE EVANS STATEMENTS: A Case for Forensic Linguistics*. *Gothenburg Studies in English*. Goetburg: University of Goteburg.

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 from http://aclweb.org/anthology/W/W13/W13-1706.pdf

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.

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

Williams, M. (2001). The Language of Cybercrime. In *Crime and the Internet* (pp. 152–166). London: Routledge.