Developing a framework for the explanation of interlingual features for native and other language influence detection

Krzysztof Kredens, Ria Perkins & Tim Grant

Aston University, UK

10.21747/21833745/lanlaw/6_2a2

Abstract. *This article demonstrates the benefit of taking an explanation-based approach in the development of features for computationally supported systems used for linguistic analysis in forensic contexts. As a focal point it considers Other Language Influence Detection (OLID) as well as its related field of Native Language Identification (NLI). An explanation-based approach allows the forensic linguist to understand the implications of the presence or absence of features as they vary across the contexts and situations s/he might encounter. The authors present a qualitative framework for types of explanation and show how different types of explanations are needed to develop a full and rich language-influence feature set. The authors are not advocating a strict or inflexible typology of feature explanation but are seeking a richness of explanation at a variety of levels of analysis instead. This in turn can be developed into computational approaches, which the authors contend will therefore be stronger and more applicable to forensic casework contexts.*

Keywords: *Native language identification, interlanguage, language typology, other language influence detection.*

Resumo. *Este artigo mostra as vantagens de uma abordagem explicativa no desenvolvimento de características usadas na análise linguística em contextos forenses por sistemas assistidos por computador, com enfoque na Deteção da Influência de Outras Línguas (OLID – Other Language Influence Detection), bem como aa área de Identificação da Língua Nativa (NLI – Native Language Identification). Uma abordagem de natureza explicativa permite ao/à linguista forense compreender as implicações da presença ou ausência de determinadas características, que variam conforme os contextos e as situações com que se depara. Os autores apresentam um enquadramento dos tipos de explicação de natureza qualitativa e mostram como são necessários diferentes tipos de explicações para desenvolver um conjunto de características de influência linguística abrangente e aprofundado. Os*

autores não defendem uma tipologia de explicações das características rígida ou inflexível; antes, procuram uma gama diversificada de explicações numa série de níveis de análise, o que, por sua vez, permite o desenvolvimento em abordagens computacionais que os autores defendem ser, portanto, mais robustas e aplicáveis em contextos de trabalho forense.

Palavras-chave: *Identificação da língua nativa, interlíngua, tipologia linguística, deteção da influência de outras línguas.*

Introduction

Much progress has been made over the last decade in computational identification of the native language of individuals writing in English. This task was originally introduced as an aspect of authorship analysis through which 'stylistic idiosyncrasies [are] used to identify the native language of the author of a given English language text' (Koppel *et al.*, 2005: 624). We choose to characterise the task more generally as the identification of linguistic features in language A, which derive from an individual's language contact with language B. The problem is most typically referred to in the literature as Native Language Identification (abbreviated as NLID or NLI; 'NLI' henceforth) and is characterized as the identification of an individual's first language (L1) from the features present when they author a text in their other languages (L2/L*n*). It is typically approached as a classification task in which a closed set of L1 languages are considered. Our framing of the task at the more general level allows us to avoid the complication of implicitly relying on a strong definition of an L1 (and L2/L*n*) and also allows for the possibility that a person's linguistic behaviour in one language may be influenced by strong language contact with multiple other languages. Following this thinking, we would like to propose the term 'Other Language Influence Detection' (OLID) (but see also Perkins and Grant, 2018).

In this article we approach the OLID problem with a determined focus on explanation of how features might help an analyst to draw any conclusions in forensic contexts. We have collected corpora of texts which are comparable to typical forensic linguistic casework applications and we derive features using bilingual language informants, who are tasked with identifying potential features in written English which index their own L1. In addition to identifying potentially useful features, the informants are also asked to provide explanations as to why an L1 speaker will use a particular feature in English, where a native English speaker will not. This article does not focus on describing the feature sets or evaluating the discriminating power of each feature, and neither does it describe the algorithms that can be used to measure the degree of influence of the L1 on English; that is done elsewhere, e.g. Koppel *et al.* (2005); Malmasi *et al.* (2017), or Tetreault *et al.* (2013). Our focus here is on how OLID can be explained as being a function of language difference and language contact. In behavioural science terms, we are thus interested in the *validity* of features, which we believe has been under-researched, as opposed to their reliability. In attempting to redress the balance, we wish to sketch a rough taxonomy of types of explanation and it is our hope that explanatory categories might be used to help understand not only features derived in our own work but also be applied to those used in more traditional NLI projects.

Language transfer is a specific form of cross-linguistic influence (CLI). Perkins and Grant (2018) trace the origins of CLI research back to Weinreich (1953) and a major breakthrough was the publication of Selinker (1972), where the term *interlanguage* was

first used. Interlanguage is 'a separate linguistic system based on the observable output which results from a learner's attempted production of a TL [Target Language] norm' (Selinker, 1972: 35) and it provides a useful conceptual departure point for work in OLID contexts.

CLI research has traditionally focused on language learners, and how a dominant language (usually an L1) might affect a non-dominant language (or L2) (see e.g. Grosjean, 1982, 1999 and, for a distinction between of L1, L2 and L3, Hammarberg, 2001). Transfer can be positive or negative, in that it can aid the production of standard or natural-sounding constructions in a second language, or ones that are not native-like. In their edited volume, Jarvis and Crossley (2012) highlight how an understanding of language transfer can be of use when taking a machine-learning approach to understanding patterns in L2 writing. Drawing conclusions from the studies discussed in the volume, Jarvis (2012) surmised that there is 'some indication that the variables used to predict the L1 of the L2 writers are transfer related.' (Jarvis, 2012: 181).

In 2013 the Association for Computational Linguistics (ACL) ran a shared task in which different teams of researchers used a common data set to develop NLI classification systems, which were evaluated in a series of blind tests (Tetreault *et al.*, 2013). In 2017 the task was broadened to three areas: 'NLI on the essay only, NLI on the spoken response only (based on a transcription of the response and i-vector acoustic features), and NLI using both responses' (Malmasi *et al.*, 2017: 62). The 2013 challenge resulted in a 'very similar set of standard features and machine learning methods' (Malmasi *et al.*, 2017: 46). In this article we use the results of this 2013 NLI shared task alongside our own developing feature sets as a starting point to discuss linguistic explanations for NLI features more generally. Competitions such as the shared task are quite naturally evaluated on the basis of success in classification; our focus, however (and crucially), is to develop NLI for use in forensic settings, and for these applications whilst accurate prediction is a crucial criterion for success, so too is explanation. Cheng (2013: 547) suggests an aspiration for forensic linguistic evidence is that it should be 'sufficiently transparent to permit reasoned decision making'. Equally, in the forensic domain, the analyst evaluating the origin of a communication typically works with a variety of technical and human intelligence on top of the linguistic data, and in these contexts they are required to measure the worth of conflicting information and make predictions to produce the most useful overall conclusions. This process of integration of available information requires an explanatorily rich approach. This need for explanation in evidential and investigative linguistics is a key issue we wish to address in this article.

This article then posits that NLI features should be explicable and we report on a developing framework for *types* of explanation across three main dimensions. We suggest first, that interlingual explanations can rely on points of *typological* linguistic distinctiveness between any two languages. Thus, some explanations will rest in the observation of structural differences between languages, such that NLI features may arise out of differences in the way those languages implement, for example, verb-and-noun agreement or verbal aspect. Second, that explanations may arise out of how languages differ in their inventories of *lexico-grammatical* structures and operate different constraints on how these structures can be populated with lexemes. Examples would include phrasal lexemes such as the English phrasal lexemes 'strong tea' and 'powerful engine' (Halliday, 1966), which can be hard for non-native speakers to acquire and reproduce. Finally,

that explanations may rely on observation of sociolinguistic differences in language use in different settings. Thus different languages and distinct varieties within a language will develop different patterns of use, and recognition of these sociolinguistic patterns of distinctiveness can also provide a basis for explanation.

A critique of theory-light, quantitative NLID approaches

Computational linguistic approaches to NLI were arguably pioneered by Koppel *et al.* (2005). Although techniques have of course moved forward, Koppel *et al.*'s (2005) original study sets the standard for design.

In Koppel *et al.* (2005) the data is drawn solely from the ICLE corpus (International Corpus of Learner English). The ICLE corpus is of classroom essays set to be common across the sub-corpora and collected by specific teachers delivering English language classes to different L1 groups internationally. Although convenient as a data set in Koppel *et al.*'s proof-of-concept article, it is our view that such a corpus is too homogenous, in terms of the data being academic writing for a teacher within an English language class, to have broad validity in forensic tasks. A difficulty with the validity of using learner corpora in NLI work can be exemplified with one of Koppel *et al.*'s (2005) findings, *viz.* that the words 'however' and 'cannot' are characteristic of the L1 Bulgarian writers in the corpus. A critique here might be that the feature is too strongly tied to the specific corpus data.

The problem is that although 'however' and 'cannot' may have been shown to be more frequent in Bulgarian advanced English learners' essays, we do not know how well such features extend to the broader, non-student, population of Bulgarians writing in English or indeed of Bulgarians writing in other contexts and/or genres. This question can of course be investigated in two ways; either by collecting broader, more representative corpora of writers of Bulgarian influenced-English, or, as we explore in this article, through developing potential explanations for useful features. We might research, for example, whether the foreign language learning national curriculum in Bulgaria places more emphasis on cohesive discourse structuring (hence the raised frequency of 'however'), or we might set out to examine whether this particular teacher preferred the word 'however' to the word 'nevertheless' in their teaching. For these features this is speculation, but for other features Koppel *et al.* (2005) do identify some markers which hint at easier interlingual explanation. For example, they suggest that some of the spelling errors made by L1 Spanish learners relate to phonological differences between Spanish and English. They also note the difficultly their model has in distinguishing between the three Slavic languages they examine (Russian, Czech and Bulgarian). These two latter observations may indicate an interesting possibility of OLID work identifying a language family rather than a specific language and this is a theme we develop below.

Koppel *et al.* (2005) are followed by others in using the ICLE corpus or similar data sets of learner English (e.g. Horbach *et al.*, 2015 or Wong and Dras, 2011) and indeed Tetreault *et al.* (2013) report that the NLI shared task mostly used data from the TOEFL11 corpus of language test data (Blanchard *et al.*, 2013). Tetreault *et al.* also comment that competitors struggled to find enough non-TOEFL11 data to train their systems. Where they did use other corpora these were typically ICLE, as used by Koppel *et al.* (2005); FCE: First Certificate in English Corpus (Yannakoudakis *et al.*, 2011); another language test corpus, ICNALE: International Corpus Network of Asian Learners of English (Ishikawa,

2011); and Lang-8 (http://www.lang-8.com), an online language-learning resource where users post diary entries in a second language and they are corrected by native speakers of that language (Brooke and Hirst, 2013). For the forensic and intelligence applications, the use of Lang-8 may be more valid than the other options as the corpus reflects at least the online mode of production of much forensic linguistic data but even here the purpose and audience still remain firmly within the language-learning domain. Arguably, most second-language writing in English is not produced by learners performing explicitly as learners, that is to say under conditions where errors will be noted, corrected and/or marked down. Rather, most non-native use of English has communicative intent that is genuinely referential or expressive, and success of communication will trump linguistic accuracy most of the time (cf. Thorne *et al.*'s (2009) findings demonstrating how L2 writers use language for creative expression to develop and maintain identities in virtual environments).

In short, Koppel *et al.* (2005), along with those that follow them, provide useful groundwork and reliable results in a constrained experimental context but may lack broader validity at both the theoretical level and, crucially for us, in forensic application. The principal issue here is that current approaches are almost entirely theory-free and almost entirely empirical and in such a context a good feature is judged to be one that allows for reliable categorisation of the L1.

The strengths and weaknesses of n-gram features

Koppel *et al.*'s (2005) experiment is based in many variables, which include letter n-grams and the distribution of a standard list of 400 function words. In the 2013 NLI shared task more than 60% of teams used some form of word n-gram as part of their analysis (see Table 1 below), with character and part-of-speech (POS) n-grams also being significantly used. The highest scoring teams all used some form of n-gram features in their approaches. Where reliable prediction is the only criterion for a useful OLID system, it would be foolish to ignore such results. However, n-gram analysis is in the first instance opaque and explanation-free. Examination of specific highly predictive n-grams is rare to find in the literature and any such analyses provide no explanations of themselves as to *why* a particular feature predicts a specific description of a text. This resistance to explanation is particularly problematic when an OLID analysis needs to be integrated with other complex and possibly conflicting forensic evidence in the context of a courtroom. In these contexts OLID analyses, as currently designed, do not easily provide a basis for reasoned decision-making.

Types and levels of explanation

Explanations will of course come in different forms, depending on the type of feature. Koppel *et al.*'s (2005) feature of learners' spelling errors in English based in Spanish pronunciation may be explained by a higher-level observation of the languages having different phonotactic rules and this explanation would then result from a different sort of empirical linguistic research. In both this case and in the use of 'however' by Bulgarian learners, the fact that there exists a potential explanation can help with evaluation and, importantly, predict where a feature will be useful and where it might fail as a predictor. If, for example, the Bulgarian heavy use of 'however' resolves to a specific English language textbook in use in Bulgaria, then we might conclude that it will not be useful for identifying Bulgarian writers who learnt through an alternative textbook. If the

Feature	Description / n (for n-gram features)	Number of teams / 29	Expressed as percentage
Word n-grams	1	16	55%
	2	18	62%
	3	9	31%
	4	3	10%
	5	1	3.5%
POS n-grams	1	11	38%
	2	15	52%
	3	12	41%
	4	6	21%
	5	2	7%
Character n-grams	1	10	35%
	2	15	52%
	3	16	55%
	4	9	31%
	5	6	21%
	6	3	10%
	7	2	7%
	8-9	1	3.5%
Function n-grams		2	7%
Syntactic features	Dependencies	6	5%
	TSG	3	10%
	CF Productions	1	3.5%
	Adaptor grammars	1	3.5%
Spelling features		3	10%

Table 1. Frequency of feature use in 2013 NLI shared task calculated from Tetreault *et al.*, 2013, Table 8.

pronunciation feature in Spanish differs between European Spanish and South American Spanish, then we might allow that the feature is useful in predicting an L1 Spanish writer but only in one geographical region. Without potential explanation no such judgements can be made. To be sure, there have been a number of attempts since Koppel *et al.* (2005) to apply explanation-based selections of features for classification tasks. Two notable examples are Brooke and Hirst (2013) and Bykh *et al.* (2013). In the former, explicit references are made to problems with extra-linguistic contextual constraints such as, for example, the fact that proper nouns, which the authors include in their set of features, may 'not directly indicate language transfer from the L1 but rather reflect real-world correlations between native language and country of residence' (2013: 400). Bykh *et al.* (2013), in turn, achieved a higher classification accuracy in the 2013 NLI shared task with linguistically-informed features (e.g. parts of speech, lemma realisations and use of derivational and inflectional suffixes) rather than surface-based n-grams. Their study also shows that linguistic explanation, apart from being useful to the forensic linguist, can also improve the performance of automated systems.

A general call for explanation, or for the mining of the output of shallow n-gram analysis for explanation is not sufficient, particularly in forensic contexts. What is re-

quired is an understanding of what types of explanation might arise and how best these could be approached by the human analyst; this is what we turn to next.

Our current OLID project has collected corpora of non-learner[1], non-native writers of English and we work with native speakers of the L1 to identify features in the English for which we can provide explanation. The informants' task is to identify features in a text and also to provide explanations as to why those features might have interlingual validity in predicting the L1. As an example, we can here provide some of the features identified for the Polish-English language pair along with the informant's explanations (see Table 2). The analyst did not follow a specific classificatory framework and labelled them as 'punctuation', 'typographic', 'grammatical' and 'style' ones. Similarly, the informants for the other language pairs had discretion in coding the features and suggesting explanations; this bottom-up approach is meant to ensure that as many features are captured as possible at the initial stage even if some of them later prove to have no predictive power. For the purposes of this article we looked at the individual features suggested by the informants across the language pairs and attempted to group the informants' labels within a classificatory framework that would respond to the various practical considerations mentioned earlier, not least of which was the need to make the explanations accessible to the end-user (which might include linguistically naïve analysts, e.g. solicitors). The framework thus could not be overly complex, which precluded one based around Odlin's (1989) comprehensive description of cross-linguistic influences in language learning comprehensive, or on an existing system of language description (e.g. Systemic Functional Grammar). The result has been a tripartite classification including (1) typological, (2) lexico-grammatical and (3) sociolinguistic types of explanation.

Feature name	Category	Informant's explanation
Unnecessary commas	Punctuation	There are very strict and complex rules in Polish about comma use, and this carries over into English text written by L1 (native) Polish speakers. The feature is even more prevalent when the speaker has a lower level of fluency in English.
Non-capitalisation of adjectives	Typographic	In Polish, adjectives are never capitalized but this is not always true in English. For example, nationalities in English are always capitalized.
Article errors	Grammatical	Polish does not use articles and therefore standard use is difficult for L1 Polish speakers to master in English. Four groups of problems can be distinguished here: • Omission of the indefinite article (*'a' and 'an'*) – very frequent • Omission of the definite article (*'the'*) – also frequent • Unnecessary insertion of the indefinite article – relatively rare • Unnecessary insertion of the definite article – relatively frequent
Use of negation	Style	There are two principal choices when it comes to negations in English, *cf.* *'I had no choice'* and *'I didn't have any choice'*. L1 Polish authors tend to overwhelmingly use the latter version, whereas L1 English authors tend to prefer the former.

Table 2. Selected interlingual features from Polish with explanations.

Explanations based in typological distinctions

The grammatical feature exemplified in Table 2 refers to a structural property of Polish (that it does not use articles) and because articles *are* used in English this creates some difficulties for the Polish learner of English. This is a level of explanation that is useful in itself, but it can become more useful when considered in the context of a broader group of languages. Thus the World Atlas of Language Structures (WALS; http://www.wals.info) identifies 198 languages that use neither definite nor indefinite articles (Dryer, 2013). This in itself is already useful as it narrows the possibilities from the more than 6000 world languages, and from this we can build a list of languages which are co-predicted by this individual feature. This idea of co-predicted languages may be useful in a model that accommodates independent non-linguistic information about a text's origin. With regard to the 198 languages, we can see using WALS that beyond Europe there are concentrations of languages with this feature in East Africa, the native languages of northern South America and in Asian languages such as Hindi, Punjabi and Pashto. Within Europe this feature is largely restricted to the Slavic language family; the only other languages with this feature are the Baltic languages, and Finnish and Saami (as spoken in northern Finland by the Lapp peoples). Within the Slavic languages the lack of articles is a feature which is common across the language family: of the 10 main Slavic languages[2], five (Russian, Ukrainian, Polish, Czech and Serbo-Croatian) use neither definite nor indefinite articles; two (Macedonian and Bulgarian) do have definite words distinct from demonstratives (i.e. in English 'the' is a distinct word from 'this' or 'those') but do not use indefinite articles; and for the remaining three main Slavic languages (Belorussian, Slovak, Slovene) WALS provides no usable information in this regard.

This kind of contextual information for a particular interlingual feature provides the basis for rich decision-making in forensic contexts. We have therefore set ourselves the task of examining which of the features identified by our informants will be susceptible to such an analysis. One recognized difficulty in pursuing this line of research is that typological work, including resources such as WALS, focuses mostly on phonological, morphological and syntactic features[3]. This means that typological research is unlikely to assist in providing this richness of explanation for the punctuation, typographic and style features in Table 2 above. For these more stylistic features, lexico-grammatical and sociolinguistic explanations are likely to prevail, perhaps in terms of cultural conventions and communities of practice. Explanations that can be derived for such lexico-grammatically and sociolinguistically based features are explored further below.

The real power of typological explanations comes to the fore in considering contrasting hypotheses about a text's origins within a small group of candidate languages. Table 3 sets out how just three typological distinctions (word order, type of general morphology and the path focus of motion verbs) vary across four languages: Russian, Mandarin, Persian and Arabic, and then English. Each of the core languages here is classified according to just the three typological distinctions and already it can been seen how they might be used in understanding and explaining individual features.

Language/ classification	Linguistic Tree	Word Order	Morphology (Isolating/analytic /fusional)	Motion verbs Path-focussed= verb-framed Manner-focussed= satellite-framed.
Russian	Indo-European>Balto-Slavic>Slavic>East Slavic> Russian	SVO	Fusional	Satellite framed
Chinese (Mandarin)	Sino-Tibetian> Sinitic>Chinese> Mandarin	SVO	Isolating Analytic	Typically considered satellite-framed although this is disputed and sometimes referred to as 'complex verb' framing
Persian	Indo-Iranian	SOV (but also displays NRel order)	Synthetic - Agglutinative	Largely satellite-framed verbs with some verbs of manner
Arabic	Afro-Asiatic>Semitic>Central Semitic>Arabic Languages>Arabic	VSO (SVO)	Fusional	Verb-framed
English	Indo-European	SVO	Moderately Analytic (more so than most - but not Isolating)	Satellite-framed

Table 3. Core languages typology by word order, morphology and motion verb path.

In brief, 'word order' here refers to one of the primary distinctions in typological work and is best understood as the typical order of subject (S), object (O) and verb (V) in a simple declarative sentence. Morphological typologies at this high level focus on how word agreement can vary between languages. With regard to 'motion verbs' different languages tend to focus on either the path or the manner of motion.

Examination of our feature sets as derived by our informants suggests that in general the SVO order of an L1 gives rise to very few features in an L2 that would appear because of this distinction. This is in spite of the fact that typologically on this dimension English is SVO, Persian is SOV and Arabic is VSO. Thus examining Persian and Arabic feature lists we might expect some features occurring because of the typological distinction but these features empirically seem not to exist in our data or at least have not been identified by our informants for Persian and Arabic. In contrast, as Swan and Smith (2001) suggest, the distinction between morphological language types can be seen to produce specific language errors. Thus not all typological distinctions will give rise directly to interlingual features that can help in the NLI task. Typological distinctions, however, are likely to produce structural features for the NLI/OLID tasks.

Tetreault *et al.* (2013) report a number of the consortia (HAI, LIM, MQ etc.) in the 2013 shared task as using structural features. These include part-of-speech (POS) n-grams, morpho-syntactic features and grammatical errors. A focus on structural features such as these is likely to include discriminating features between L1s which may have

typological explanations. The use of POS n-grams in particular is effectively a dragnet for discovery of points of grammatical distinctiveness between writers of different L1s. Closer examination of the specific discriminating parts of speech might therefore allow the generation of rich explanations.

Typological explanation is not, however, limited to structural features. As Koppel *et al.*'s (2005) original article attests, pronunciation features, for example, can have phonological explanations. In the NLI 2013 shared task a few consortia report using spelling errors as features (Goutte *et al.*, 2013; Lavergne *et al.*, 2013; Nicolai *et al.*, 2013) and for some of these at least typological explanations might be developed. Not all structural or spelling features will be amenable to typological explanation, however, and we will have to look further through our typology of explanation to understand what causes these effects in L2 texts.

Explanations based in lexico-grammatical distinctions

Given the wealth of language typology research and access to readily available resources like WALS, explanations based on typological differences seem to be the most practicable to develop. However, because of the large number of natural languages in existence, in its search for 'common denominators' traditional typological research of necessity has had to operate at relatively high levels of description and as such does not have the capacity to capture and explain finer differences. Lexico-grammatical descriptions, with their roots in Systemic Functional Linguistics, can help deliver a different type of explanation. Table 4 below illustrates two features from our data within this category.

Feature	L1	Description	Examples of feature
Deviant phrasal lexeme	AR	The phrasal lexeme 'equal parts X and Y' is modified.	'I should warn you: this blog will be equal parts food, equal parts fashion'
Violation of inversion rule	RU	Certain lexemes trigger inversion in standard English but not in and Slavic languages, where inversion is mostly used to form questions and rarely triggered by lexical items.	'I think Colombian girls are about the same level of beauty but no way they can compare in personality. '

Table 4. Example explanations for lexico-grammatical features.

The focus within this category is predominantly on the idea of the formulaic sequence, i.e. 'a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar' (Wray, 2002: 9). EFL literature as well as our data suggest that three categories of formulaic language can be particularly interesting for our purposes: idioms, collocations and phrasal lexemes.

Idioms are expressions whose meaning cannot be decoded by analysing their individual constitutive lexemes. A speaker expressing the intention 'to paint the town red'

as part of their weekend entertainment plans does not likely refer to using a brush and a bucket of red paint to decorate the town's buildings. Idioms are thus metaphorical but they are also culture-specific, as exemplified below with the ways of conceptualising the idea of expending effort unnecessarily in, respectively, English, Spanish, Polish and Arabic:

> Carry coals to Newcastle
> Throw water into the sea
> Carry wood to the forest
> Carry dates to Basra

Collocation in turn is the habitual co-occurrence of two or more words that cannot be predicted in traditional transformational models of grammar; for example there is no reason why in English the noun *bath* is typically preceded by the verb *run*, rather than *prepare*. 'I prepared him a bath' (rather than 'I ran him a bath') may thus be the result of an L1 interference. Similarly, the English phrasal lexemes 'strong tea' and 'powerful engine' might be realized as 'powerful tea' and 'strong engine' by a non-native speaker, or 'mum and dad' can become 'dad and mum' (examples from Halliday, 1966).

We find that lexico-grammatical features like these are relatively easy to spot. However, explanations may be difficult to obtain if the forensic linguist is monolingual or has no familiarity with the languages in question. In other words while s/he may identify 'dad and mum' as marked and thus potentially non-native, the explanation may not impose itself readily. Having identified likely or possible non-native features it is necessary and possible to conduct research to generate explanations.

The most used type of feature in the NLI 2013 shared task was the word n-gram and whilst word n-grams can be mined for explicable structural interlingual features, this is not true of all discriminating word n-grams. Conversely, looking for word n-grams will certainly miss some potentially useful phrasal lexemes. For example the lexeme from the L1 Arabic writer based on the formula 'equal parts X and Y' is unlikely to be captured by such a method.

Explanations based in sociolinguistic variation

Sociolinguistic explanations relate to features which do not seem to be based in recognisable phonological, morphological, syntactic or lexico-grammatical distinctions between languages and will include patterns of punctuation, typography and ways of conceptualising natural or cultural phenomena. We need to recognize that we will not be able to provide explanation of those features in terms of typological or lexico-grammatical distinctions between languages. For these features the explanation will be broadly sociolinguistic and potentially cultural, and we need to engage in further work exploring how to express explanations for these features. Such explanations arise from rates of occurrence in stratified corpora but we also will look to explain differential use socially and culturally. Table 5 below shows examples of features with sociolinguistic explanations.

All three features are clearly culturally based. *Inshalla* has its roots in Islam and the phatic expression 'Have you eaten?' can be traced back to historical contexts of famine and their implications for the cultural significance of food in modern-day Korea. However, the phrasal lexeme 'bad habits' would be less obvious to spot and also more difficult to explain.

Feature	L1	Description	Examples of feature
Inshallah	Ar	L1 Arabic speakers tend to refer to God much more frequently when speaking about the future (Inshallah, or 'God willing'). In Arabic, sentences look like this: Next week, God willing (Inshallah) my friends and I want to go to Lebanon.	'Really miss you and wonder how kind is life being to u...inshallah all is going well'
Phatic expression	All	Language-specific conventions for phatic communication	'Have you eaten?' (Korean greeting/welcome)
Phrasal lexeme 'bad habits'	Ru	'(without) bad habits' in Russian means the person is a non-smoker and does not drink. It may be used in dating ads and other types of self-promotion discourse.	'I'm responsible, polite and patient with kids. Without bad habits (no smoking, drinking *etc*).'

Table 5. Example explanations for sociolinguistic features.

Some features at this explanatory level will then be relatively self-evident but others may be potentially confusing. Similar to the lexico-grammatical level, explanations will be difficult to obtain if the analyst is monolingual or has no familiarity with the languages and cultures in question. Nevertheless, as in the case of the previous category, upon identifying a marked, apparently interlingual feature, the forensic linguist could research it using search engines and/or a tailor-made geo-locating tool. What is important to note is that sociolinguistic explanations will likely be useful to the analyst in narrowing down the list of candidate L1s in cases where typological explanations point to a language family. In researching the possible sociolinguistic explanations a variety of online resources can be used, for example to obtain information on EFL national curricula and/or the most popular textbooks in a given country of interest. In addition, online traffic data could help identify culturally salient topics/frames for the country; the next step could be to see how these are habitually lexicalised in the L1. For example, in the discussion of internal political controversies, 'corruption' and 'nepotism' could be among the most salient content words in one country but 'theft' and 'money-laundering' could dominate in another. A member of an English-speaking online forum could feasibly apply these culturally imposed frames when discussing the current political affairs of a country other than his/her own.

Conclusions and paths forward

This article makes one key assertion; that if NLI or OLID, or any other computationally based analysis for that matter, is to be used in forensic work, then features used in such analyses must be explicable. Further to this assertion, we articulate a framework for types of explanation and show how different types of explanation are required to explain the breadth of different NLI language features. The framework enables a rich explanation of the features identified. Classifying the descriptions enables not only a better understanding of the features already collected, but can indicate areas which might benefit from a more systematic search for features. We are not advocating a strict and exclusive typology of feature explanation such that offering a typological explanation for a particular feature prevents there also being an explanation based in sociolinguistic variation. On the contrary, we are pursuing richness of explanation, which might include a variety of levels of analysis. An explanation-based approach allows the forensic

linguist to understand the implications of the presence or absence of features as they vary across the contexts and situations s/he might encounter.

The three types of feature explanation that we use above are typological, lexico-grammatical and sociolinguistic. We acknowledge that there may be further levels of explanation not considered here. Bykh *et al.* (2013) indicated that explanations can help support computational work and we posit that not only could such a framework as outlined here support and deepen the utility of existing computational features, but it might also indicate computational approaches to feature finding, such as e.g. developing search techniques to identify phrasal lexemes or parallel idioms across languages.

Notes

[1]By 'non-learners' we mean individuals whose primary motivation in writing in online environments seems to be creative expression rather than language learning. It is of course often difficult to separate the two, which is why when in doubt we used contextual information (as well as the texts in question themselves) to decide if a text should be included in our corpus.

[2]Those with more than a million speakers.

[3]There are a few exceptions e.g. WALS chapters on Ordinal and Distributive Numerals.

References

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A. and Chodorow, M. (2013). *TOEFL11: A Corpus of Non-Native English.* Rapport interne, Educational Testing Service.

Brooke, J. and 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 from http://www.aclweb.org/anthology/W13-1725.

Bykh, S., Vajjala, S., Krivanek, J. and Meuers, D. (2013). Combining 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*, 197–206, Atlanta, Georgia: NAACL/HLT.

Cheng, E. K. (2013). Being Pragmatic About Forensic Linguistics. *Journal of Law and Policy*, 212, 541–550.

Dryer, M. (2013). Definite Articles. In M. Dryer and M. Hapelmath, Eds., *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institue for Evolutionary Anthropology.

Goutte, C., Léger, S. and Carpuat, M. (2013). Feature space selection and combination for native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 96–100.

Grosjean, F. (1982). *Life with Two Languages.* Cambridge, MA.: Harvard University Press.

Grosjean, F. (1999). Studying bilinguals: Methodological and conceptual issues. *Bilingualism: Language and Cognition*, 1, 117–130.

Halliday, M. A. K. (1966). Lexis as a Linguistic Level. *Journal of Linguistics*, 21, 57–67.

Hammarberg, B. (2001). Roles of L1 and L2 in L3 production and acquisition. In J. Cenoz, B. Hufeisen and U. Jessner, Eds., *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives.* Clevedon: Multilingual Matters, 21–41.

Horbach, A., Poitz, J. and Palmer, A. (2015). Using shallow syntactic features to measure influences of L1 and proficiency level in EFL writings. In *Proceedings of the 4th workshop on NLP for Computer Assisted Language Learning at NODALIDA 2015*, 21–34, Vilnius.

Ishikawa, S. I. (2011). A new horizon in learner corpus studies: The aim of the ICNALE project. In *Corpora and language technologies in teaching, learning and research*, 3–11.

Jarvis, S. (2012). Detection-Based Approaches: Methods, Theories and Applications. In S. Jarvis and S. Crossley, Eds., *Approaching Language Transfer through Text Classification*. Bristol: Multilingual Matters, 178–188.

S. Jarvis and S. Crossley, Eds. (2012). *Approaching Language Transfer through Text Classification Second Language Acquisition Kindle Edition*. Bristol: Multilingual Matter Ltd.

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: ACM Press.

Lavergne, T., Illouz, G., Max, A. and Nagata, R. (2013). LIMSI's participation in the 2013 shared task on native language identification. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, 260–265, Atlanta, Georgia, USA.

Malmasi, S., Evanini, K., Cahill, A., Tretreault, J., Pugh, R., Hamill, C., Napolitano, D. and Qian, Y. (2017). A Report on the 2017 Native Language Identification Shared Task. In *12th Workshop on Innovative Use of NLP for Building Educational Applications*, 62–75.

Nicolai, G., Hauer, B., Salameh, M., Yao, L. and Kondrak, G. (2013). Cognate and misspelling features for natural language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 140–145.

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. Cambridge: Cambridge University Press.

Perkins, R. and Grant, T. (2018). Native Language Influence Detection for Forensic Authorship Analysis. Identifying L1 Persian Bloggers. *International Journal of Speech Language and the Law*, 25(1), 1–20.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(209-231).

M. Swan and B. Smith, Eds. (2001). *Learner English: A Teacher's Guide to Interference and other Problems*. Cambridge: Cambridge University Press, 2 ed.

Tetreault, J., Blanchard, D. and Cahill, A. (2013). A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, 48–57.

Thorne, S. L., Black, R. W. and Sykes, J. M. (2009). Second language use, socialization, and learning in Internet interest communities and online gaming. *The Modern Language Journal*, 93(Focus Issue), 802–821.

Weinreich, U. (1953). *Languages in Contact. Findings and Problems*. New York, NY: Linguistic Circle of New York.

Wong, S. J. and Dras, M. (2011). Exploiting Parse Structures for Native Language Identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing 1600–1610*, Edinburgh.

Wray, A. (2002). Formulaic language in computer-supported communication: theory meets reality. *Language Awareness*, 112, 114–131.

Yannakoudakis, H., Briscoe, T. and Medlock, B. (2011). A New Dataset and Method for Automatically Grading ESOL Texts. In *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 180–189.