"Native language identification"

Jarvis, Scott ; Paquot, Magali

Abstract

The main objective of this chapter is to examine the automatic techniques that can be applied to learner corpora to identify learners' mother tongue backgrounds from their patterns of production in the target language. The chapter focuses particularly on aspects of learner corpus design that have a direct bearing on the results (size, topic homogeneity, type and level of annotation). Special attention is also paid to the contribution of NLI research to SLA, more particularly in the form of the detection-based approach to transfer. NLI is a relatively new line of enquiry: research so far has focused on identifying the native language of writers using English as a foreign language and the chapter will thus deal with English as a target language.

Document type : *Contribution à ouvrage collectif (Book Chapter)*

Référence bibliographique

Jarvis, Scott ; Paquot, Magali. *Native language identification.* In: Granger S. ; Gilquin G. ; Meunier F., *Cambridge Handbook of Learner Corpus Research*, Cambridge University Press : Cambridge 2015

28. Learner corpora and native language identification

1. Introduction

Native language identification (NLI) is the task of automatically identifying the first language (L1) of a language user on the basis of the person's production of the target language. This research pursuit is guided by the assumption that a person's L1 background can be inferred from how frequently he or she makes use of certain features of the target language (e.g. words, word sequences, sequences of characters). The task is typically modelled as a text categorisation problem where the set of L1s is predefined and each text is assigned an L1 on account of its specific language features.

NLI offers potential practical applications in a wide variety of domains that rely on language corpora. Among other benefits, NLI appears to enhance the performance of a number of natural language processing (NLP) tasks such as speech recognition, parsing and information extraction (Mayfield Tomokiyo and Jones 2001). NLP tools and techniques are typically trained on native speaker data and are consequently often less robust when applied to non-native language (L2) (Díaz-Negrillo *et al.* 2010; Chapter 25 this volume). A second benefit of NLI is that its results may contribute to the success of machine learning approaches to author identification and profiling. These techniques are today of crucial interest for a number of web-related fields such as Internet security and cybercrime investigation (Argamon *et al.* 2009).

The results of an NLI task may also contribute to Second Language Acquisition (SLA) theory building. The ability to detect the L1 of individuals on the basis of their use of certain specific features of the target language indeed offers unprecedented opportunities for the study of transfer, i.e. 'the influence resulting from similarities and differences between the target language and any other language that has been previously (and perhaps imperfectly) acquired' (Odlin 1989: 27; see also Chapter 15 this volume). The rapprochement between NLI techniques and transfer research was first made by Tsur and Rappoport (2007) and has recently been fully articulated in the detection-based approach to transfer (Jarvis 2010, 2012). In this exploratory approach, the results of an NLI task are used as primary data to investigate the nature and extent of L1 influence in non-native language use.

The quality of the dataset used is central in NLI research and large comparable sets of texts produced by non-native speakers representing a variety of L1s are required to detect significant group differences. A limited number of NLI studies have used web-derived resources such as Wikipedia comments and news agency texts as input data (e.g. Al-Rfou' 2012; Tofighi *et al.* 2012; Brooke and Hirst 2013). However, most of the existing research has relied on multi-L1 learner corpora such as the *International Corpus of Learner English* (*ICLE*) (Granger *et al.* 2009). A major advantage of using learner corpus data is that each learner text is usually richly documented as regards the learner (e.g. proficiency) and task settings (e.g. topic) (see Chapter 2 this volume). These variables are potentially confounding and need to be carefully controlled in NLI tasks.

The main objective of this chapter is to examine the automatic techniques that can be applied to learner corpora to identify learners' mother tongue backgrounds from their patterns of production in the target language. The chapter focuses particularly on aspects of learner corpus design that have a direct bearing on the results (size, topic homogeneity, type and level of annotation). Special attention is also paid to the contribution of NLI research to SLA, more particularly in the form of the detection-based approach to transfer. NLI is a relatively new line of enquiry: research so far has focused on identifying the native language of writers using English as a foreign language and the chapter will thus deal with English as a target language.[1]

2. Core Issues

Most of the existing studies on NLI have been directed at one of the following two aims: (1) determining which machine-learning tools, techniques and procedures lead to the highest levels of NLI classification accuracy, and more recently (2) probing the nature and degree of L1 influence in language learners' L2 speech or writing within the framework of the detection-based approach to transfer. These two aims overlap but are not fully congruous. For example, researchers pursuing the second aim typically limit their analyses to a particular area of language use, such as word choice (Jarvis, Castañeda-Jiménez and Nielsen 2012), cohesion and complexity (Crossley and McNamara 2012), or error categories (Bestgen *et al.* 2012). They also tend to limit the number of variables analysed to no more than a few dozen to a few hundred and use a computer-automated classifier that provides information about which specific variables and combinations of variables are associated with which specific L1 backgrounds (e.g. Jarvis and Paquot 2012). Researchers pursuing the first aim, on the other hand, include whichever and however many variables will increase the classification accuracy of the computational model and also choose computer-automated classifiers on the basis of how accurately they can identify the L1s of L2 users rather than on the basis of how well the results illuminate the relationship between specific L1 backgrounds and specific patterns of L2 use (e.g. Tetreault *et al.* 2012; Jarvis *et al.* 2013). In the following subsections, we draw from studies of both types while discussing how machine-learning classifiers work, what types of features they use to detect L1s, the types of challenges that researchers encounter in this area of research and the strength of the evidence that NLI provides for L1 influence.

2.1 How do machine-learning classifiers identify learners' L1s?

Machine learning involves the use of computer programs designed to discover patterns in the observed data (i.e. the training set) that can be generalised and predicted to occur in data that have not yet been encountered by the program (i.e. the test set) (e.g. Alpaydin 2004). Computer programs used for such purposes are sometimes referred to as classifiers, but the term 'classifier' is often used more narrowly to refer to a computational model that the program has constructed of the relationship between classes and features (Kotsiantis 2007). The term 'class' refers to a category that a particular data sample represents, such as the L1 of the writer of a particular text;

[1] See Pepper (2012) for the first L1 identification study that is based on interlanguage data other than English. The study relies on Norwegian learner data from the *ASK corpus* and is designed as a partial replication of Jarvis, Castañeda-Jiménez and Nielsen's (2012) study of lexical features.

‘feature’ refers to an entity, pattern or property found in the data, such as a particular word that the writer has used in the text. In NLI research, machine-learning programs are used to create classifiers (or computational models or systems) that represent the relationship between L1s (classes) and characteristics of learners’ language use (features). Features are often operationalised as the relative frequencies of specific letters or combinations of letters, specific words or sequences of words, parts of speech (POS) or sequences of POS, specific errors or categories of errors, or abstract properties such as levels of cohesion, complexity and lexical diversity (see Section 2.2).

There are many different types of classifiers. As described by Kotsiantis (2007), these include decision trees, rule-based classifiers, artificial neural networks, statistical-learning classifiers such as discriminant analysis and Naïve Bayes and instance-based classifiers such as k-Nearest Neighbors and Support Vector Machines. In recent NLI studies, the most commonly used type of classifier is by far Support Vector Machines (SVM), followed by discriminant analysis (DA) and Maximum Entropy analysis (MaxEnt). In this section, we will limit our discussion to just these three.

SVM is a classification method that creates statistical models that distinguish between pairs of classes (e.g. L1s) in the training data. It does this by representing each case (i.e. each text) in the training data as a point in multidimensional space. The coordinates for each text are vectors that reflect the occurrence of the chosen features in the given text. In order to separate L1s, SVM uses mathematical means to create a hyperplane (or decision boundary) that best separates one L1’s texts from the texts of other L1s by maximising the margin (i.e. distance) between the L1s and the hyperplane that separates them (cf. Figure 1). When the model is later applied to new texts (i.e. the test set), it plots those texts in its multidimensional space and identifies the L1 for each text in accordance with which side of the hyperplane the text falls on. Because SVM is fundamentally a binary classification method, when the number of L1s exceeds two, the program that implements SVM generally learns to distinguish between either (a) one L1 versus all others, or (b) every possible pair of L1s. In both cases, the program creates multiple classifiers. In the one-versus-all method, the L1 of a new text is determined by whichever classifier produces the widest margin between the hyperplane and the text. In the one-versus-one method, on the other hand, a text’s L1 is determined as the L1 that the text has been assigned to by a plurality of competing classifiers. There are a number of computer packages that have been designed to build SVM classifiers, such as *WEKA* (Hall *et al.* 2009), *LIBSVM* (Chang and Lin 2011) and *LIBLINEAR* (Fan *et al.* 2008). Importantly, they have all been built for different purposes and do not appear to be equally useful for NLI tasks. In a recent study by Bykh and Meurers (2012), for example, the researchers compared these same three SVM packages and found that ‘the *LIBLINEAR* classifier yielded by far the best results and was in addition usually faster than the others as well’ (ibid.: 430).

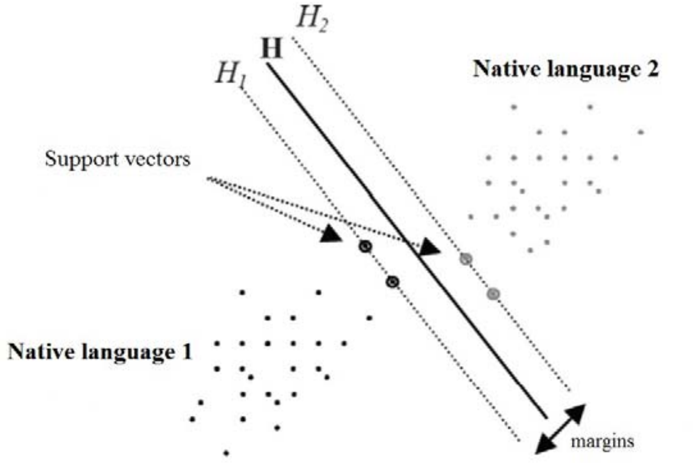

Figure 1: Principles of Support Vector Machines

DA is similar to SVM in the sense that it creates a multidimensional space and plots each text as a point in that space. As with SVM, in DA a text's coordinates in the multidimensional space are vectors reflecting the presence or absence of each target feature or its (relative) frequency, combined with weights that minimise the distances between texts from the same L1 while maximising distances between texts from different L1s. Unlike SVM, DA does not attempt to create margins to separate L1s. Instead, it uses mathematical means to determine the centre – or centroid – of all points representing a particular L1. It thus creates a separate centroid for each L1 and it classifies new texts by plotting them in this multidimensional space and by measuring their distance to each centroid (cf. Figure 2). New texts – or texts in the test set – are identified as belonging to the L1 group whose centroid they are closest to. There are a number of computer applications that perform DA classification, including popular statistical packages such as *SPSS*, *SAS* and *R*. Not all of them have the same varieties of DA, however, and researchers who use DA often supplement the program's built-in functions with their own programming (cf. studies in Jarvis and Crossley 2012).

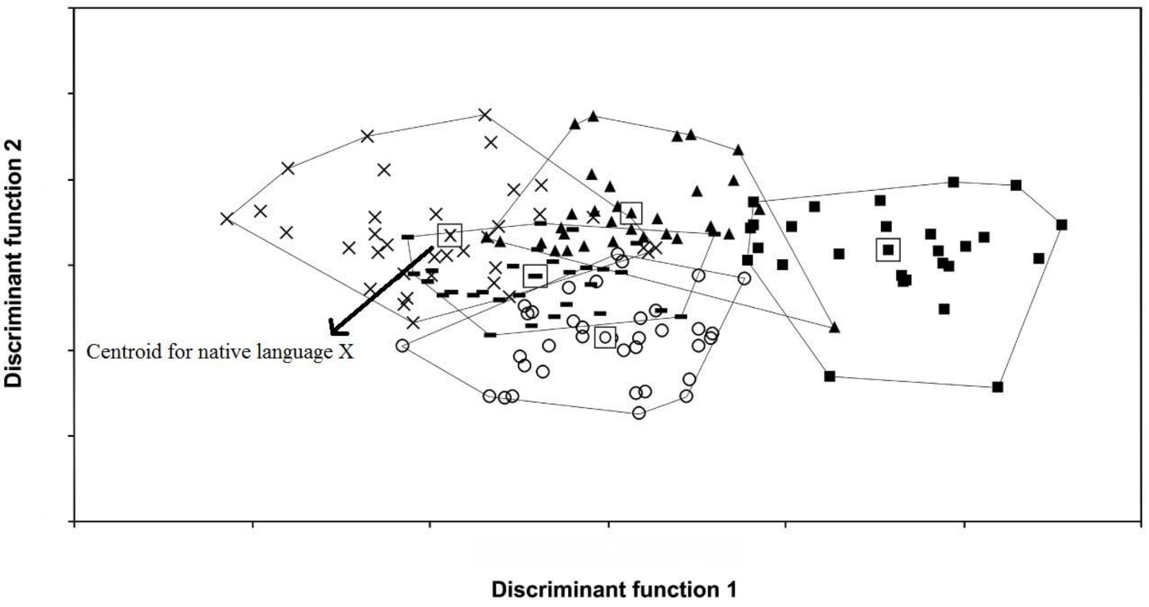

Figure 2: Principles of Discriminant Analysis

MaxEnt relies on the notion of entropy from information theory (Shannon 1948), which states that when nothing is known about a particular case, the probability of its class association is equally distributed across all of the candidate classes. This means that, when attempting to identify the L1 affiliation of a particular text, all of the candidate L1s have an equal probability of being the correct L1 until the researcher begins examining the individual features of the text and how those features are distributed across the texts representing different L1s. For example, when there are eight candidate L1s and one of them is Bulgarian, the prior probability that the writer of a particular text is Bulgarian is 1/8. However, if the text contains a feature that occurs 25% of the time in texts written by Bulgarian speakers, then the probability that the writer of the text is Bulgarian rises to 1/4 while it also changes for the other L1s in a way that is proportional to the distribution of that feature across L1 groups. When multiple features are entered into the model, MaxEnt 'allocates the probability distribution as evenly as possible, so it computes the entropy of all conditional probabilities and finds the most unconstrained distribution' (Ahn 2011: 15). MaxEnt can also be described as a specialised multinomial logistic model (Brooke and Hirst 2012a) in that it classifies texts according to L1 background by determining the probabilities of association between each text and each L1, and by identifying the L1 affiliation of a text as the L1 for which it has the highest probability of association. Computer applications that implement MaxEnt classifiers are few in number; all of the recent NLI studies that have made use of MaxEnt (e.g. Wong and Dras 2011; Wong *et al.* 2012) have relied on the *MegaM* software.[2]

Each of the three methods of classification just described has its own advantages and disadvantages. In terms of raw classification accuracy, SVM so far has achieved the highest

[2] www.umiacs.umd.edu/~hal/megam/ (last accessed on 2 April 2014).

levels of L1 identification with both the *ICLE* (Bykh and Meurers 2012; Tetreault *et al.* 2012) and the *TOEFL11* corpora (Jarvis *et al.* 2013). One of its major advantages is that it can accommodate classification models consisting of an unlimited number of features; this is often what is required to achieve the highest possible L1 identification accuracy. NLI studies that have relied on SVM have typically included more than 1,000 features in their models, and in the case of Jarvis *et al.* (2013), the number of unique features included in the final model exceeded 400,000.

When SVM is compared with other classification methods using the same number of features, however, SVM does not always produce the best results. For example, Jarvis (2011) compared DA with SVM and 18 other classification methods in relation to their ability to identify the L1 affiliations of 2,033 argumentative essays in the *ICLE* that were written by learners from twelve L1 backgrounds. The features used in the analysis included 722 of the most frequent lexical n-grams in the corpus. One of the major disadvantages of DA is that it requires a large ratio of cases to variables in order to avoid creating an unstable model where small changes in the measured variables can generate vastly different outcomes. An unstable model will not provide trustworthy results. The recommended minimum ratio of cases per variable is often said to be 20:1 (Spicer 2005: 146), although others have pointed out that a ratio of 10:1 is sometimes sufficient (Brown and Wicker 2000: 214). Jarvis (2011) adopted the latter ratio, but even so, the maximum number of features that could be used in his study of 2,033 texts was only about 200. This turned out not to be a problem, however, because while relying on only 200 features, DA achieved a higher L1 classification accuracy than SVM and all other methods that were given access to all 722 features in the feature pool.[3] The most important advantages of DA are that it performs well (sometimes better than other classifiers) when dealing with only a few dozen to a few hundred variables, and it also gives a clear indication of how learners' use of individual features helps distinguish specific L1s from one another. This is valuable for analyses that attempt to explicate the nature of L1 influence and illustrate the unique linguistic tendencies of learners from specific L1 backgrounds.

The advantages and disadvantages of MaxEnt are somewhat less clear because this method of classification has been used in only a few NLI studies and has only rarely been compared with SVM or other classification methods on the same task. There are two studies we are aware of that have compared MaxEnt with SVM. The first, Wong and Dras (2011), found that MaxEnt performed much better than SVM on a classification task involving 26,284 unique features. A subsequent study by Brooke and Hirst (2012a) compared eleven configurations of SVM and two configurations of MaxEnt in relation to their ability to identify the L1 affiliations of texts included in the *ICLE* and the First Certificate in English (FCE) portion of the *Cambridge Learner Corpus.* The researchers found that MaxEnt performed better than SVM when no bias

[3] This does not mean that DA will always outperform SVM when put on equal footing. In fact, SVM in this particular study was instantiated through the R statistical application, and it is possible that the results would have been more favourable toward SVM if it had been implemented through *LIBLINEAR* using more advanced procedures and settings (cf. Bykh and Meurers 2012). Also, SVM undoubtedly would have outperformed DA if it had been fed a substantially larger number of features, such as 1,000 or more.

adaptation[4] was used, but was worse than almost all SVM options when bias adaptation was used. Importantly, both studies show that MaxEnt produces promising results and carries some of the benefits of both SVM and DA: like SVM, it can be applied to an unlimited number of features, and like DA, it provides detailed results that can elucidate which features are most indicative of which L1s (see, e.g., Ahn 2011).

2.2 What types of language features are most useful for L1 identification?

The vast majority of existing NLI studies have included words (i.e. lexical features) in their classifiers. Depending on the nature of the task, the number of L1s and the relationship between the L1s, relatively high levels of L1 classification accuracy can be achieved with just a few dozen words. For example, in an examination of written descriptions of a silent film produced by adolescent foreign-language learners of English from five L1 backgrounds (Danish, Finnish, Portuguese, Spanish, Swedish), Jarvis, Castañeda-Jiménez and Nielsen (2012) found that the learners' L1s could be identified with 76.9% accuracy on the basis of their use of just 53 of the most frequent words in the corpus (e.g. *a*, *away*, *be*, *come*, *into*, *the*). This is because the learners from the different L1 backgrounds used these words with predictably different frequencies. For example, the Finnish speakers used *a*, *the* and *be* significantly less frequently than all other groups, the Portuguese and Spanish groups used *away* and *come* significantly less frequently than the other three groups, the Danish speakers used *into* significantly more frequently than all other groups, and so on. There was no single word that distinguished all groups from all others, but because the classifier included several different words that separated groups in different ways, it was able to identify most texts' L1 affiliation correctly on the basis of how the 53 target words patterned together within any given text.

When examining the usefulness of lexical features for NLI, it is important to recognise that lexical items often consist of multiple elements (e.g. *train station, to break down*; see Chapter 10 this volume on such multi-word expressions). Language data are also replete with recurring multi-word sequences (e.g. *as far as*, *on the other hand, it is important*) that may or may not be regarded as lexical items, but which may nevertheless reflect the preferences of speakers from particular L1 backgrounds in ways that go beyond what can be found when examining words individually (Paquot 2013). Multiword sequences – or lexical n-grams (e.g. bigrams, trigrams, 4-grams, 5-grams) – have been included as features in a number of NLI studies and they often contribute more to the classifier's ability to identify L1s than any other category of features. This was certainly the case in Brooke and Hirst (2012a), who ran NLI analyses using several categories of features, including not only lexical n-grams, but also character n-grams, individual function words and various types of grammatical and syntactic features. The researchers noted that lexical n-grams 'alone account for almost all of the accuracy we see when all features are combined' (Brooke and Hirst 2012a: 401).

Error categories are another type of feature that has figured prominently in past NLI studies. In most cases, researchers have used automated tools to tag and quantify errors by type (e.g. Koppel

[4] Bias adaptation as defined by Brooke and Hirst (2012a) is a matter of domain adaptation. It involves the adjustment of a single parameter to allow the classifier to adapt a model learnt in one domain (e.g. in one corpus) to another domain (another corpus).

et al. 2005; Wong and Dras 2009; Tetreault *et al.* 2012; Chapter 26 this volume), but the typical accuracy rates for automated error tagging are not high and this has led to disappointing results. To address this weakness, Bestgen *et al.* (2012) manually tagged the errors in their corpus (see Section 3) and in a follow-up study, Jarvis, Bestgen, Crossley, Granger, Paquot, Thewissen and McNamara (2012) found that those errors, by themselves, were slightly but not significantly more effective than lexical n-grams in identifying the L1 affiliations of argumentative essays written in English by native speakers of French, German and Spanish. The authors furthermore showed that the best model included a combination of different types of features and that manually tagged error categories contributed significantly and substantially to the strength of that model (see also Kochmar 2011).

Two other types of features have figured prominently in past NLI studies: POS n-grams (e.g. PREP + ART + NOUN) and letter n-grams (e.g. *pr*, *pl*, *pla*, *ple*). The use of POS n-grams requires that the words in each text be tagged for the POS they represent (e.g. *in the house* >> *in*[PREP] *the*[ART] *house*[NOUN]). This is generally done with automated tagging programs, such as the TreeTagger (Schmid 1995). Automated POS tagging is not perfectly accurate, but Schmid found that the TreeTagger identifies the POS of words with an accuracy rate between 96% and 98% (see Chapter 5 this volume for a discussion of POS-tagging of learner corpus data). POS n-grams – and particularly POS bigrams – have been found to be very useful for NLI, often in combination with other types of features. For Mayfield Tomokiyo and Jones (2001), the features underpinning the most accurate NLI classifier included words and mixed bigrams made up of a noun (i.e. the actual word) and the POS tag for whatever word co-occurred with the noun. Later studies by Koppel *et al.* (2005) and Tsur and Rappoport (2007) included a set of 250 rare POS bigrams in their analyses, but did not report that the POS bigrams contributed much to their classifiers. Wong and Dras (2009), on the other hand, expanded their set of POS n-grams to 650 and found that they could achieve their highest level of classification accuracy with a pool of features that included only function words and POS n-grams.

POS n-grams are often used in studies that also include letter n-grams and some of these studies have shown that letter n-grams are valuable for NLI. The study by Koppel *et al.* (2005) showed that their highest accuracy rate was achieved with a combination of 400 function words, 200 letter n-grams, 185 error types and 250 rare POS bigrams. However, they also mentioned that they were able to achieve nearly as high an accuracy rate with the 400 function words alone, and also with the 200 letter n-grams alone. In a later study by Kochmar (2011), the researcher found that letter n-grams and POS n-grams were the most powerful features of all in that they contributed more to successful NLI than any of their lexical, syntactic or error features. No other study we are aware of has found letter or POS n-grams to be more useful for NLI than lexical features (including single words and lexical n-grams), but POS and letter n-grams usually do play an important role in the most successful NLI models (Tetreault *et al.* 2012; Jarvis *et al.* 2013).

Beyond the standard types of features just discussed, a number of studies have explored the potential usefulness of various complex and abstract features. These include – but are not limited to – syntactic constituency structures (Wong and Dras 2011), syntactic dependencies (Kochmar 2011), Tree Substitution Grammar features (Post and Gildea 2009), measures of textual cohesion, lexical sophistication, syntactic complexity and conceptual knowledge (Crossley and McNamara

2012) and psychological indices (e.g. sadness, negative emotion, overall affect) (see Torney *et al.* 2012). The general finding from the studies that have used these features is that they do indeed tend to improve classification accuracy when combined with more traditional features. On the other hand, the highest L1 classification accuracy achieved so far on the *TOEFL11* corpus used only traditional features consisting of lexical and POS n-grams (Jarvis *et al.* 2013). The fact that a classifier built only on traditional features outperformed systems such as that of Tetreault *et al.* (2012) – which included not only traditional features but also language-specific models, Tree Substitution Grammar features and syntactic dependencies (see Section 3) – suggests that learners' L1-related patterns of L2 performance are to be found mainly in their choice and placement of words and word categories.

2.3 What types of challenges do researchers encounter in this area of research?

The most serious challenges that NLI researchers face have to do with the corpora on which they run their analyses. An ideal training corpus for NLI research would be one where (1) all of the texts in the corpus are written on the same topic or at least in the same genre with a symmetrical distribution of topics across L1 groups, (2) the texts are all of similar lengths or have similar length means and standard deviations across L1 groups, (3) all of the learners are at precisely the same level of L2 proficiency or at least the levels of proficiency are evenly balanced across L1 groups, (4) the learners within and across groups have similar educational, socio-economic and psychological profiles, and (5) the learners within and across groups have had comparable amounts and types of instruction in and exposure to the target language. Due to the practical realities of corpus construction, however, no learner corpus has yet met these criteria. The problem that this creates for an NLI analysis is that, unless the only non-random difference between L1 groups is their native language, then the classification program will construct a classifier by relying on whatever predictable differences between groups it can identify, even if the differences have nothing to do with the learners' native languages. As a result, the confounding factors will artificially boost classification accuracy if the training and testing sets come from the same corpus (see Brooke and Hirst 2013 for a discussion of topic bias as a confounding factor on word n-grams as well as characters n-grams, POS n-grams and function words).

Most NLI studies to date however have used the *ICLE* as both their training and test corpus. This is because the *ICLE* is one of the largest multi-L1 learner corpora (commercially) available, consisting of 6,085 essays written in L2 English by speakers of sixteen different L1s. Researchers have nevertheless noted two major problems regarding the makeup of the *ICLE*. These relate to the distribution of topics and proficiency levels across L1 groups. The topic problem has been discussed by several researchers, most recently by Tetreault *et al.* (2012: 2589-90), who pointed out that 'there are many topics for which all the authors are native speakers of a single L1… . For example, only Chinese authors responded to the prompt "Discuss the advantages and disadvantages of using credit cards"'. Consequently, Chinese speakers can be expected to be far more likely than speakers of other L1s to use words such as *credit card*, *money*, *pay*, *buy* and *purchase* in their essays – irrespective of any L1 influence (but see Section 3). The typical solutions to this problem are (a) to limit the training data to texts written on topics that do not show an L1 sampling bias (Tetreault *et al.* 2012) and (b) to omit from the analysis any words, phrases or other features that come directly from the writing prompt or which are narrowly related to it (cf. Jarvis and Paquot 2012). These solutions appear to have worked well in the studies that have implemented them, but because of the nature of the *ICLE*, it has not been possible to create a subcorpus that is completely uniform in the distribution of topics across L1s.

The topic problem is also compounded by the proficiency problem. Bestgen *et al.* (2012) examined the latter problem carefully in an NLI investigation of *ICLE* argumentative essays written by L1 speakers of French, German and Spanish, which were also assigned proficiency ratings by two professional raters (cf. Section 3). The researchers noted that, in their NLI analysis, essays written by French and German speakers were rarely misidentified as having been written by Spanish speakers. However, they also noted that the Spanish speakers' proficiency ratings were significantly lower than those of the other two groups. This means that the classifier could have distinguished between Spanish speakers and the speakers of the other two groups on the basis of proficiency-related differences rather than on the basis of differences related to their L1 backgrounds. In order to mitigate this problem, the researchers narrowed their data to those texts that had the same proficiency rating and they followed this up with a qualitative analysis of the texts written by the different L1 groups in order to show which features of the learners' L2 performance clearly did resemble the L1. As far as the quantitative analysis was concerned, the researchers' solution was appropriate, but they noted that the number of texts in each group that represented the same proficiency level was quite small, and that this may have rendered the results unreliable. Presumably, the sample size would have become completely untenable if the researchers had tried to limit the data to texts that were both at the same proficiency level and written on the same topic. In recognition of these problems, Tetreault *et al.* (2012) have recommended the use of the *TOEFL11* corpus for future NLI research. This corpus is larger and better balanced for topic and proficiency than the *ICLE* is, though it is not perfectly symmetrical either (see Jarvis *et al.* 2013). However, the *ICLE* still has an important role to play in future research of this type (see Section 4).

Aside from problems related to the type of learner corpus data available, other challenges that researchers encounter in this area of research include problems related to the selection and extraction of features and technical challenges related to using the classification software and performing the statistical analysis. Regarding challenges related to the selection of features, it is perhaps best when the selection of features is driven by well-motivated hypotheses and theoretical assumptions, including contrastive analyses between the L1s in question (cf. Odlin 2006). Most NLI studies so far, however, have been exploratory and have included an unrestrained assortment of features, including but not limited to those that worked well in earlier studies, such as words, lexical n-grams and POS n-grams. In such a young area of research, it may indeed make sense to try as many different feature types as possible and see what works best. However, even after the researcher has chosen which features to include in the analysis, important decisions need to be made about how to extract them from the data and in what form. For example, when using words as features, the researcher needs to decide whether different forms of the same word (*go*, *goes*, *going*, *gone*, *went*) will be treated as different words or the same word (lemma). Alternatively, it might be useful to do both; Jarvis *et al.* (2013) included both word forms and lemmas in their NLI analysis of the *TOEFL11*, and their system has produced the highest L1 classification accuracies so far for the *TOEFL11*.

Concerning technical challenges related to using the classification software, some classification tools, such as DA, are often used in conjunction with automated protocols for selecting the best set of features from the feature pool the researcher has assembled and made available to the classification software. In the case of DA, it is possible to use a stepwise feature-selection procedure, which adds features to the classifier one by one in accordance with how much they contribute to the classifier's ability to discriminate L1s. This is useful, but the researcher must set criteria, such as *p* values, that determine which features and how many features will be added to the classifier. Often, the best way to determine this is through experimentation with different values.

As mentioned earlier, certain classification tools, such as SVM, do not function optimally unless parameter tuning is carried out on test data. *LIBLINEAR* offers eight different varieties of SVM (Fan *et al.* 2008), and apparently all of them make use of a parameter referred to as a cost parameter. The cost parameter allows SVM to create a soft margin between classes that permits some misclassifications (i.e. texts on the wrong side of the margin). The higher the value of this parameter, the fewer misclassifications of the training set it allows. However, as we describe in the following paragraph, fitting the classifier too rigidly to the training data can render it less generalisable to new cases. *LIBLINEAR* has a built-in program that optimises this parameter by testing multiple values of it until the program arrives at a value that gives the best results. *LIBLINEAR*'s optimisation program is not necessarily always the most efficient, however, and Jarvis *et al.* (2013) used their own program to find the best value of the cost parameter. Programming skills are a major asset in this area of research.

Of final concern in this section are challenges associated with the generalizability of predictive models. If a classifier is only trained on a dataset of known data (i.e. training dataset), its classification accuracy will be overly optimistic because it will have been tailored to account specifically for the texts it has already encountered, and it might not generalise well to unknown data. An analogy is useful for understanding how this works: if a man goes to a tailor to have a custom shirt made, the tailor will attempt to design the shirt so that it conforms to the specific contours of the man's body. This is essentially what classification software does with the training data: it builds a model that conforms to the specific contours of the data. This may seem like an advantage until one recognises that when a shirt has been designed to fit a particular person perfectly, it might not fit anyone else well at all. A good classifier needs to do what a clothing manufacturer does: it needs to capture what is generalizable and not what is idiosyncratic in the training data.

In order to determine how generalizable a classifier is, it must be tested on a separate test dataset. This can be done in a number of ways. One way is to apply the trained classifier to a completely new set of data to see how accurately it identifies the L1s in the new dataset. Using a completely new set of data as the test set is preferable if both the training set and test set are very large, each containing several hundred if not several thousand texts per L1. However, in most cases, researchers do not have access to such large amounts of data and practical constraints mean that they need to use the same corpus for both training and testing. In such cases, whatever corpus they are using could be split into a single training set and a single test set. Because different ways of splitting the data can lead to different results (cf. Molinaro *et al.* 2005), however, most NLI studies have divided the corpus into equal size subsets (typically 10) and have cycled through a series of steps where each set is given its own turn to serve as the test set while all other sets are combined into the training set. This is called k-fold cross-validation, or 10-fold cross-validation when the number of sets is 10. The average classification accuracy across folds is used as an indication of how well the classifier will be able to identify the L1s of new texts in the future (see Jarvis 2011: 135). Most classification software performs cross-validation by using the same classifier (i.e. the same features and parameters) during each stage of the cross-validation. However, some researchers have argued that this can lead to overly optimistic estimates of the generalizability of the model and that what is really needed is to allow the classification software to construct a different classifier for each stage of the cross-validation (e.g. Lecocke and Hess 2006). Doing so, however, often requires researchers to create their own programs to facilitate this type of analysis, and it also causes problems for researchers to interpret and report exactly what the final solution is since each stage of the cross-validation might have used a slightly different set of features and classifier parameters (see Section 3 and the studies in Jarvis and Crossley 2012 for examples of how researchers have dealt with these challenges).

2.4 How strong is the evidence that this type of research provides for L1 influence?

Jarvis (2000) has argued that L1 influence has three outward manifestations: (a) similar L2 patterns produced by speakers of the same L1, (b) different L2 patterns produced by speakers of different L1s and (c) congruence between the patterns learners produce in their L1 and L2 (see also Chapter 15 this volume). Jarvis emphasised that all three effects need to be examined, and that a researcher is not justified in claiming to have found L1 influence unless more than one effect has been demonstrated and unless the researcher has adequately controlled for other variables (e.g. L2 proficiency, types and amounts of L2 instruction) that can produce similar effects. In NLI tasks, classifiers are built with software that specifically seeks out the first two effects in the training data in order to minimise within-group differences and maximise between-group differences. Thus, when a classifier is successful in accurately identifying the L1 affiliations of learner texts significantly beyond the level of chance, the classifier has essentially met Jarvis' criteria for confirming the presence of L1 influence in the data as long as the potentially confounding effects of other variables have been adequately ruled out. Nearly all NLI studies so far have successfully met the first criterion (i.e. finding at least two of the three effects), but few if any have successfully met the second criterion (i.e. eliminating potential confounds) – due mainly to problems inherent in existing learner corpora, as we discussed in the preceding section.

Researchers interested in using NLI methods to investigate L1 influence have so far relied on two types of solutions to compensate for the potential threats of confounding variables. The first solution, as we discussed in the preceding section, is to reduce the training data so that the texts from each L1 group are balanced for genre, topic, length and L2 proficiency. In principle, this is an entirely appropriate solution, but as we pointed out earlier, in practice it often results in such a small subsample of the original data as to make its generalisability suspect. The second solution is to go beyond the automated NLI analysis (which, again, produces evidence of the first two effects) in search for the third effect – i.e. qualitative and/or quantitative (e.g. frequency-related) evidence of L2 patterns of performance that resemble those of the L1. Augmenting L1 identification with an analysis of cross-language congruities in learners' patterns of performance not only makes the argument for L1 influence stronger, but it also helps clarify precisely how, where and to what extent learners' L2 performance is affected by their L1s. This is the approach taken in all of the empirical studies included in the volume edited by Jarvis and Crossley (2012). Although the researchers in these studies were not able to determine precisely the degree to which L1 influence (versus L2 proficiency and other confounding factors) accounted for the strength of their L1 classification models, they were able to show that the learners' use of many L2 features did indeed reflect patterns of the learners' L1s and were almost certainly the result of L1 influence. Importantly, even when the limitations of the corpus make it impossible to verify L1 influence through NLI analysis alone, NLI analysis allows researchers to cast their nets broadly over large sets of features in order to identify in a very efficient manner which specific L2 features are most likely to be affected by the L1. NLI is in some respects analogous to using a metal detector: it does not settle the question of what is buried underfoot, but it does show where to dig. As described by Jarvis (2010, 2012), the detection-based approach (i.e. NLI analysis) and the comparison-based approach (i.e. traditional group-based comparisons of learners' use of individual L2 features) can be used on a complementary basis as the exploratory and confirmatory phases of a large-scale investigation of L1 influence.

3. Representative studies

The last few years have witnessed a boom in the amount of research and literature devoted to the task of automatically identifying the first language of a language speaker on the basis of his or her production in the target language. In this section, we provide a detailed presentation of three recent NLI studies which address different research questions and accordingly represent a variety of research designs: Tetreault *et al.* (2012), Bykh and Meurers (2012) and Bestgen *et al.* (2012) fed three very different sets of features (from a small set of error categories to a large combination of features) into three different classifiers and relied on various learner corpora. Despite their many differences, they are all highly successful. As such, we believe that they offer three valuable and informative examples for any researcher who would like to carry out an NLI task in the future.

3.1.Tetreault, J., Blanchard, D., Cahill, A. and Chodorow, M. 2012. 'Native tongues, lost and found: resources and empirical evaluations in native language identification', in *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. December 2012, Mumbai, India: The COLING 2012 Organizing Committee, pp. 2585-602. http://aclweb.org/anthology//C/C12/C12-1158.pdf (last accessed on 2 April 2014).

Tetreault *et al.*'s (2012) study is representative of research which aims to evaluate the effectiveness of different types of features to provide the highest levels of NLI classification accuracy. As shown in Table 1, it experiments with many of the features commonly used in the field (e.g. letter n-grams, words) as well as other more novel features (e.g. syntactic dependencies, 5-gram language models). In addition to building individual classifiers for all the features, Tetreault *et al.* also experimented with two ways of combining them: (1) they combined the features into a large set and built a unique model from this combination and (2) they used the predictions for each individual feature in a final ensemble model. They employed an NLI system using a logistic regression model as implemented in the *LIBLINEAR* library (Fan *et al.* 2008) and k-fold cross validation on four datasets (cf. Section 2.3). The first dataset comes from the *ICLE* and consists of 770 argumentative essays evenly spread across the seven L1s that, following Wong and Dras (2009), have been used in quite a few NLI tasks based on the *ICLE*: Bulgarian, Chinese, Czech, French, Japanese, Russian and Spanish. Three independent datasets were also compiled out of TOEFL tests: (1) *TOEFL7* is larger than the *ICLE* sample although it does include the same seven languages; (2) *TOEFL11* consists of 11,000 written compositions and covers eleven L1s; (3) *TOEFL11-Big* consists of 87,502 exams from speakers of the same L1s as found in *TOEFL11*.

Features	Description
Letter n-grams	These five types of features roughly correspond to the features described in the work by Koppel *et al.* (2005) and which subsequently served as a baseline in several NLI studies.
Function words	
Part-of-speech bigrams	
Spelling errors	
Writing quality features	
Word n-grams	Correctly spelled content words and bigrams.
Tree substitution grammar fragments	The use of grammatical relations in NLI tasks is less widespread but they have been used as features in quite a few recent studies (e.g. Wong and Dras 2011; Kochmar 2011). Tree substitution grammar fragments were first used in Swanson and Charniak (2012).
Stanford dependencies	
N-gram language model	Perplexity scores from 5-gram language models, one for each language in the corpus.

Table 1: A summary of the features used in Tetreault *et al.* (2012)

Results showed that combining the predictions of each individual type of feature in an ensemble model performed best in the four datasets. On the *ICLE*, the best ensemble model achieved an accuracy of 90.1% and a close look at the results for the individual types of features revealed that single words and bigrams, syntactically-based features and perplexity scores from 5-gram language models were the most powerful predictors.

Tetreault *et al.*'s (2012) work is also of particular interest to any new researcher in the field as they addressed important issues in NLI such as corpus design and corpus size, the impact of proficiency and topic bias as well as cross-corpus evaluation. For example, they showed that a system trained on a small corpus did not yield good results when tested on a larger corpus whereas using a larger corpus for training and a smaller one for testing produced a better outcome.

3.2. Bykh, S. and Meurers, D. 2012. 'Native language identification using recurring n-grams – Investigating abstraction and domain dependence', in *Proceedings of the 24th International Conference on Computational Linguistics* (COLING 2012), December 2012, Mumbai, India: The COLING 2012 Organizing Committee, pp. 425-40. http://aclweb.org/anthology//C/C12/C12-1027.pdf (last accessed on 2 April 2014).

Like Tetreault *et al.* (2012), a number of other studies have highlighted the effectiveness of n-grams for NLI. However, most studies have so far reported results for n-grams of a specified length and different types of n-grams (lexical vs. POS n-grams) have rarely been compared. Bykh and Meurers's (2012) study was thus particularly timely in that it experiments with recurring n-grams of various lengths and degrees of abstraction to determine which features ensure the highest NLI accuracy values. The authors conducted ten experiments based on ten samples from the *ICLE*. Each dataset consisted of a randomly selected set of essays between 500 and 1,000 words in length representing seven native languages frequently used in NLI tasks: Bulgarian, Chinese, Czech, French, Japanese, Russian and Spanish. For each of the languages, the dataset amounts to 70 essays for training and 25 essays for testing, thus resulting in a total of 490 essays for training and 175 for testing. The essays were tokenised and cleaned of all punctuation marks, special characters and capitalisation.

Three types of n-grams were used as features: word-based n-grams, POS n-grams and Open-Class-POS-based (OCPOS) n-grams where only the nouns, verbs, adjectives and cardinal numbers are replaced by the corresponding POS tags. All n-grams that occur in at least two different essays of the training set were selected with a view to representing the whole range of n values (unigrams, bigrams, trigrams, etc.) as well as possible [1, n] intervals (e.g. unigrams, unigrams and bigrams together, unigrams, bigrams and trigrams together) and thus investigating 'up to which value of n the n-grams may be worth considering despite being rare' (Bykh and Meurers 2012: 429).

Recurring n-grams were fed into a machine learning setup employing an SVM classifier as available in the *LIBLINEAR* library. The highest mean accuracy value achieved by the n-gram approach on the ten *ICLE* samples was 89.37% using word-based n-grams with the [1,2] interval. OCPOS n-grams yielded about 9% lower accuracy than word-based n-grams, and POS n-grams performed about 13% worse than OCPOS n-grams. However, even POS n-grams yielded a reasonably high result (about 67%) considering that seven different native languages were used as classes and the random baseline against which to evaluate the results was thus 14.29%. In the ten experiments, using a combination of n-grams of various lengths always led to better results than using n-grams of just one particular size, and the accuracy benefited from n-grams up to a length of five words.

To explore the generalizability of the results, the authors conducted a second set of similar experiments with three additional learner corpora representing three native languages, i.e. Spanish, Swedish and Chinese: the *Non-Native Corpus of English* (*NOCE*, Díaz-Negrillo 2009), the *Uppsala Student English Corpus* (*USE*, Axelsson 2000) and the *Hong Kong University of Science and Technology English Examination Corpus* (*HKUST*, Milton and Chowdhury 1994). The Spanish, Swedish and Chinese *ICLE* corpora were used to compile ten separate training sets of 420 randomly selected essays each, with 140 essays per native language. Two separate test sets were compiled. To compile the first test set, Bykh and Meurers (2012) randomly selected 70 essays per L1 from the *ICLE*. The second test set consisted of 70 randomly selected essays per native language from the *USE* and *HKUST* and 140 shorter essays from the *NOCE* that were merged into pairs to obtain 70 texts of comparable size to the essays in the other learner corpora. The setup made it possible to conduct ten single-corpus evaluations (i.e. training and testing on the same data) on the *ICLE* samples alone and ten cross-corpus evaluations using *ICLE* data for training and data from *NOCE, USE* and *HKUST* for testing. As in the first set of experiments, models using word-based n-grams performed best on average, with a best mean accuracy value of 96.48% for the single-corpus evaluation setup using *ICLE* only and 87.57% in the cross-corpus evaluation. Contrary to Brooke and Hirst's (2013) claim that surface-based models trained on the *ICLE* do not generalise well to other learner corpora (see Section 2.3), Bykh and Meurers's (2012) results thus rather suggest that training on *ICLE* and testing on other learner corpora still yield reasonably high accuracy values for an NLI task with three native languages.

3.3. Bestgen, Y., Granger, S. and Thewissen, J. 2012. 'Error patterns and automatic L1 identification', in Jarvis and Crossley (eds), pp. 127-53.

Bestgen *et al.*'s (2012) work differs from the two studies described above in a number of ways. Unlike in Tetreault *et al.* (2012) and Bykh and Meurers (2012), the features used were not automatically extracted; the main objective was to explore whether manually annotated error patterns could be relied on for NLI. As a result, the dataset was also much smaller and consisted of 223 argumentative essays from the *ICLE* written by French-, German- and Spanish-speaking learners, amounting to c. 50,000 tokens per L1 group. The learner essays were carefully annotated for errors in accordance with the latest version of the Louvain Error Tagging Manual (Dagneaux *et al.* 2008).

The classification technique used in this study was DA with a stepwise feature-selection procedure. The features fed into the DA classifier were the relative frequencies of 48 error sub-categories representing the following seven main error domains: formal errors, grammatical errors, lexical errors, lexico-grammatical errors, punctuation errors, style errors and errors involving redundant, missing or misordered words. The DA was implemented with LOOCV (i.e. leave-one-out cross-validation) which classifies each text based on the model constructed from the DA for the other 222 texts. This is the same as a k-fold cross-validation with K being equal to the number of texts in the original dataset.

The approach yielded an overall classification accuracy of 65%, which is surprisingly good when considering that, unlike in Tetreault *et al.* (2012) or Bykh and Meurers (2012), the model consisted of a very small set of features. In fact, only 12 of the 48 error categories served to differentiate texts written by speakers of the three L1s and were selected in more than half of the folds of the CV. For example, seven error types were strongly associated with the Spanish learner group: spelling errors, lexical errors on single words, lexical errors on phrases, erroneous article use, erroneous dependent prepositions used with verbs, unclear pronominal references and erroneous demonstrative determiners.

After presenting the results, the authors provided detailed analysis of the effects of L2 proficiency and developmental factors on the error patterns in the dataset that made their NLI classification task successful. They also discussed potential transfer effects on the errors that best discriminate between the different L1s and demonstrated that a number of the errors could safely be attributed to L1 influence.

4. Critical assessment and future directions

NLI is a new area of research whose potential has only just barely begun to be tapped. As mentioned in Section 1, NLI seems to hold promise for future applications in speech recognition, information extraction, authorship attribution, author profiling and crime investigation. NLI results may also be used to inform language pedagogy, language assessment and intelligent computer-assisted language learning systems (see Chapter 24 this volume). Learners experience various difficulties and make different errors as the result of interference from their L1s (Díez-Bedmar and Papp 2008; Bestgen and Granger 2011). A writing tutoring system which includes information about potential transfer-induced errors (e.g. false cognates) from a variety of L1s and can detect the native language of the learner with high precision will be able to determine whether an error is likely to be L1-induced and prioritise L1-L2 contrastive feedback accordingly (Amaral and Meurers 2008).

The most obvious value of NLI for SLA is its potential contribution to the exploration and verification of L1 influence but it presents a number of challenges. As described in Section 2.4, successful L1 identification confirms only two of the three effects of L1 influence mentioned by Jarvis (2000). The effect it is blind to is whether learners' L2 performance reflects patterns of their L1s. A seemingly promising future direction for NLI research would therefore involve the use of L1 corpora as part of the training data for an L1 classifier. Exactly how this would work is not entirely clear, but it would need to involve ways of matching L1 features – including words and sequences of words – with their counterparts in the target language. Brooke and Hirst (2012b) have made a preliminary attempt at doing this through the use of bilingual dictionaries to create lists of all possible translations of words and bigrams in their L1 corpora, which were used as the training data. The researchers then created a way to characterise the overall prevalence of these features in L1 texts and also applied the formula to L2 texts. Each L2 text was then classified as representing the L1 whose prevalence profile it most closely resembled. The researchers found that their results were somewhat disappointing, but clearly they have taken an important first step toward incorporating L1-L2 congruity into NLI. We hope that other researchers will continue working on this important challenge.

As mentioned, the framework for investigating transfer that was proposed by Jarvis (2000) highlights three important effects of L1 influence. Jarvis (2010) introduced yet a fourth effect, and this is something that might also profitably be incorporated into future NLI research. The fourth effect is referred to as *intralingual contrasts* because it involves a class of features of the L2 where learners' performance can be predicted to vary (from one feature to the next) because the corresponding features of the L1 do not form a unitary class. Prepositions are a good example because they form a coherent syntactic category in English, but their counterparts in another language might consist of a variety of linguistic elements. For example, the prepositions *near*, *under* and *in* function in a syntactically similar manner in English, but their counterparts in Finnish are a preposition (e.g. *lähellä taloa* = "near the house"), a postposition (e.g. *talon alla* = "under the house") and a locative case suffix (e.g. *talossa* = "in the house"). On the basis of L1 influence, one could predict that Finnish-speaking learners of English will have little trouble using *near*, more difficulty in the proper use of *under* and the most difficulty with the use of *in*. This is indeed what the evidence shows (Jarvis 2010) and future NLI research could probably benefit substantially from contrastive analyses of this type that take into consideration the unique configuration of similarities, differences and zero relationships (cf. Ringbom 2007) that exist between a given L1 and L2. Such a principled, theory-driven and a priori approach to selecting features for L1 identification is difficult to find in any existing NLI studies.

Of the features that have been used so far, learners' errors are perhaps the most problematic and yet they offer a great deal of potential promise. The available automatic error taggers are not sufficiently accurate to allow automatically tagged errors to contribute much to NLI research (e.g. Koppel *et al.* 2005; Wong and Dras 2009). Manually tagged errors, on the other hand, are useful (Kochmar 2011; Bestgen *et al.* 2012), but since NLI researchers usually aim for full automation, the future of manually tagged errors in NLI research is dubious. Manual error tagging is expensive and time-consuming, so unless automatic error tagging eventually becomes fully viable, we do not expect to find this category of features in many future studies.

Solving technical problems like this as well as discovering the best sets and combinations of features, and the best classifiers, parameters and procedures for conducting NLI research will require coordinated efforts by a large community of dedicated and competent researchers. Fortunately, researchers within the field of computational linguistics regularly organise shared tasks that involve numerous teams of researchers competing to solve the same problem by applying their own innovative solutions to the same set of data. This allows researchers to determine efficiently and on a large scale which methods, techniques and procedures are the most promising for addressing a specific problem. In 2013, an NLI shared task was organised by Joel Tetreault, Aoife Cahill and Daniel Blanchard.[5] Twenty-nine teams from around the world responded to the call to participate, and each was given access to an early release of the *TOEFL11* corpus to create their L1 classification models. The data included not only the original 11,000 *TOEFL11* essays written by L2 learners of English, but also information about the L1 backgrounds and L2 proficiency levels of the writers of each essay, as well as the topic that each essay was written about. Later, the teams participating in the shared task were given 1,100 additional essays written by learners from the same L1 backgrounds, but this time the researchers were not given information about which L1 was associated with which specific text. The researchers' task was thus to predict the L1 of the writer of each text. Nearly half of the teams achieved an overall classification accuracy rate higher than 80% and the highest rate was 83.6% (for an overview of the results, see Tetreault *et al.* 2013). The most successful methods in the shared task were similar to or the same as many of the methods we have described earlier. The main value of a shared task is that it allows apples to be compared with apples; because each of the teams used exactly the same data, differences in their NLI accuracy rates can be ascribed specifically to differences in the features, classifiers and parameters they relied on rather than to differences in their data or sampling procedures. The organizers of the first NLI shared task have already begun preparing for a second shared task involving a broader range of genres. This will be an important next step in advancing this line of research.

Finally, as the new *TOEFL11* corpus gains popularity for NLI research, we caution researchers not to assume that it is ideal or even the best or only corpus that should be used for this type of research. Although it does seem better suited for research on NLP than most or all previous corpora, it is less well documented than the *ICLE*, which might remain more attractive for research dealing with the multifaceted nature of second language acquisition. Certainly, future research on NLI should be applied to a wider variety of corpora, genres and conditions in order

[5] For more detail, see https://sites.google.com/site/nlisharedtask2013/home (last accessed on 15 April 2014).

to determine how extensive and reliable L1 influence is, through which combinations of features it manifests itself and how well its manifestations can be detected through the proper combination of scholarship and technology.

Key readings

Mayfield Tomokiyo, L. and Jones, R. 2001. 'You're not from 'round here, are you? Naive Bayes detection of non-native utterance text', in *Proceedings of NAACL2001: The Second Meeting of the North American Chapter of the Association for Computational Linguistics*. Carnegie Mellon University, Pittsburgh, PA, USA, 2-7 June 2001.

The first NLI task was to the best of our knowledge reported by Mayfield Tomokiyo and Jones (2001). This exploratory study made use of a limited set of 45 transcripts to investigate whether spontaneous speech produced by native versus non-native speakers of English could be distinguished automatically with a view to customising the acoustic and language models of a speech recognition system and thus improving handling of non-native input.

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*, Chicago, IL, USA, pp. 624-8.

Koppel *et al.* (2005) focused on a classification task involving five L1 backgrounds and used SVM. The features fed into the classifier included 400 function words, 200 frequent letter n-grams, 185 error types and 250 rare POS bigrams. Quite a few follow-up studies have adopted the same sets of features as first proposed in Koppel *et al.* (2005), which is often presented as a seminal work in NLI.

Jarvis, S. 2011. 'Data mining with learner corpora: Choosing classifiers for L1 detection', in Meunier, F., De Cock, S., Gilquin, G. and Paquot, M. (eds.), *A Taste for Corpora: In Honour of Sylviane Granger*. Amsterdam: Benjamins, pp. 127-54.

This work will be of interest to all researchers who need to decide which classifier to use for a particular NLI task. After a presentation of the different types of classifiers and their major characteristics, Jarvis (2011) compares the performance of twenty classifiers on an NLI task that uses a set of 722 n-grams of length one to four as features and relies on samples of the *ICLE* as a dataset for training and testing.

Jarvis, S. and Crossley, S. A. (eds.) 2012. *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*. Bristol: Multilingual Matters.

This edited volume has its origins in the fields of SLA and learner corpus research and focuses on the use of NLI tools and techniques to investigate the nature and extent of L1 influence. In the introductory chapter, Jarvis sets the scene for the detection-based approach to transfer. The core of the book consists of five empirical studies that experiment with different sets of features (words, n-grams, error categories and automated indices of cohesion, lexical sophistication and syntactic complexity) for L1 automatic identification. The plus point of this collection of chapters is that special attention is paid to the interpretation of results in terms of transfer effects and how to tease them apart from other confounding variables.

Tetreault, J., Blanchard, D. and Aoife, C. 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*, June 2013, Atlanta, GA, USA: Association for Computational Linguistics.

Tetreault *et al.* (2013) report on the first NLI Shared Task whose goal was to evaluate different machine-learning tools and procedures on the same dataset (*TOEFL11*) and determine which NLI system would maximise classification accuracy. The focus is thus on the technical details of the different systems and the report is illustrative of the NLP paradigm to native language identification. More details about the different contributions to the Shared Task are available in separate chapters of the workshop proceedings.

References

Ahn, C. 2011. *Automatically Detecting Authors' Native Language*. Unpublished Master Dissertation, Naval Postgraduate School, Monterey, CA. http://edocs.nps.edu/npspubs/scholarly/theses/2011/March/11Mar_Ahn.pdf (last accessed on 2 April 2014).

Al-Rfou', R. 2012. 'Detecting English writing styles for non-native speakers', *Computing Research Repository* abs/1211.0498. http://arxiv.org/abs/1211.0498 (last accessed on 2 April 2014).

Alpaydin, E. 2004. *Introduction to Machine Learning*. Cambridge, MA: The MIT Press.

Argamon, S., Koppel, M., Pennebaker, J. and Schler, J. 2009. 'Automatically profiling the author of an anonymous text', *Communications of the ACM* 52(2): 119-23. http://dl.acm.org/citation.cfm?id=1461959 (last accessed on 2 April 2014).

Amaral, L. and Meurers, D. 2008. 'From recording linguistic competence to supporting inferences about language acquisition in context: Extending the conceptualization of student models for intelligent computer-assisted language learning', *Computer-Assisted Language Learning* 21(4): 323-38.

Axelsson, M. W. 2000. 'USE – The Uppsala student English corpus: an instrument for needs analysis', *ICAME Journal* 24: 155-7.

Bestgen, Y. and Granger, S. 2011. 'Categorising spelling errors to assess L2 writing', *International Journal of Continuing Engineering Education and Life-Long Learning* 21(2/3): 235-52.

Bestgen, Y., Granger, S. and Thewissen, J. 2012. 'Error patterns and automatic L1 identification', in Jarvis and Crossley (eds), pp. 127-53.

Brooke, J. and Hirst, G. 2012a. 'Robust, lexicalized native language identification', in *Proceedings of the 24th International Conference on Computational Linguistics* (COLING-

2012), December 2012, Mumbai, India: The COLING 2012 Organizing Committee, pp. 391-407. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.379.4451 (last accessed on 2 April 2014).

Brooke, J. and Hirst, G. 2012b. ‘Measuring interlanguage: native language identification with L1-influence metrics’, in *Proceedings of the 8th ELRA Conference on Language Resources and Evaluation* (LREC 2012), Istanbul. www.lrec-conf.org/proceedings/lrec2012/pdf/129_Paper.pdf (last accessed on 2 April 2014).

Brooke, J. and Hirst, G. 2013. ‘Native language detection with ‘cheap’ learner corpora’, in Granger, S., Gilquin, G. and Meunier, F. (eds.), *Twenty Years of Learner Corpus Research. Looking Back, Moving Ahead. Proceedings of the First Learner Corpus Research Conference (LCR2011)*. Louvain-la-Neuve: Presses universitaires de Louvain, pp. 37-47.

Brown, M. T. and Wicker, L. R. 2000. ‘Discriminant analysis’, in Tinsley, H. E. A. and Brown, S. D. (eds.), *Handbook of Applied Multivariate Statistics and Mathematical Modelling*. New York: Academic Press, pp. 209-35.

Bykh, S. and Meurers, D. 2012. ‘Native language identification using recurring n-grams – Investigating abstraction and domain dependence’, in *Proceedings of the 24th International Conference on Computational Linguistics* (COLING 2012), December 2012, Mumbai, India: The COLING 2012 Organizing Committee, pp. 425-40. http://aclweb.org/anthology//C/C12/C12-1027.pdf (last accessed on 2 April 2014).

Chang, C.-C. and Lin, C.-J. 2011. ‘LIBSVM: A library for support vector machines’, *ACM Transactions on Intelligent Systems and Technology,* 2(3): 27:1–27:27. Software available at www.csie.ntu.edu.tw/~cjlin/libsvm/ (last accessed on 2 April 2014).

Crossley, S. A. and McNamara, D. 2012. ‘Detecting the first language of second language writers using automated indices of cohesion, lexical sophistication, syntactic complexity and conceptual knowledge’, in Jarvis and Crossley (eds.), 106-26.

Dagneaux, E., Denness, S., Granger, S., Meunier, F., Neff van Aertselaer, J. and Thewissen, J. 2008. *Error Tagging Manual version 1.3*. Université catholique de Louvain, Centre for English Corpus Linguistics.

Díaz Negrillo, A. 2009. *EARS: A User’s Manual*. Munich: LINCOM Academic Reference Books.

Díaz-Negrillo, A., Meurers, D., Valera, S. and Wunsch, H. 2010. ‘Towards interlanguage POS annotation for effective learner corpora in SLA and FLT’, *Language Forum* 36(1/2): 1-15.

Díez-Bedmar, M.B. and Papp, S. 2008. ‘The use of the English article system by Chinese and Spanish learners’, in Gilquin, G., Papp, P. and Díez-Bedmar, M. B. (eds.), *Linking up Contrastive and Learner Corpus Research*. Amsterdam and Atlanta: Rodopi, pp. 147-75.

Fan, R., Chang, K., Hsieh, C., Wang, X. and Lin, C. 2008. ‘Liblinear: A library for large linear classification’, *The Journal of Machine Learning Research* 9: 1871-4. Software available at www.csie.ntu.edu.tw/~cjlin/liblinear/ (last accessed on 2 April 2014).

Granger, S., Dagneaux, E., Meunier, F. and Paquot, M. 2009. *The International Corpus of Learner English. Version 2. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. 2009. ‘The weka data mining software: An update’, *The SIGKDD Explorations* 11: 10-18. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.148.3671 (last accessed on 2 April 2014).

Jarvis, S. 2000. ‘Methodological rigor in the study of transfer: Identifying L1 influence in the interlanguage lexicon’, *Language Learning* 50: 245-309.

Jarvis, S. 2010. ‘Comparison-based and detection-based approaches to transfer research’, in Roberts, L., Howard, M., Laoire, M. Ó and Singleton, D. (eds.), *EUROSLA Yearbook 10*. Amsterdam: John Benjamins, pp. 169-92.

Jarvis, S. 2011. ‘Data mining with learner corpora: Choosing classifiers for L1 detection’, in Meunier, F., De Cock, S., Gilquin, G. and Paquot, M. (eds.), *A Taste for Corpora: In Honour of Sylviane Granger*. Amsterdam/Philadelphia: Benjamins, pp. 127-54.

Jarvis, S. 2012. ‘The detection-based approach: an overview’, in Jarvis, S. and Crossley, S. (eds.), pp. 1-33.

Jarvis, S., Bestgen, Y., Crossley, S. A., Granger, S., Paquot, M., Thewissen, J. and McNamara, D. 2012. ‘The comparative and combined contributions of n-grams, Coh-Metrix indices and error types in the L1 classification of learner texts’, in Jarvis and Crossley (eds.), pp. 154-77.

Jarvis, S., Castañeda-Jiménez, G. and Nielsen, R. 2012. ‘Detecting L2 writers’ L1s on the basis of their lexical styles’, in Jarvis and Crossley (eds.), pp. 34-70.

Jarvis, S. and Crossley, S. A. (eds.) 2012. *Approaching Language Transfer through Text Classification. Explorations in the Detection-based Approach*. Bristol: Multilingual Matters.

Jarvis, S. and Paquot, M. 2012. ‘Exploring the role of n-grams in L1 identification’, in Jarvis and Crossley (eds.), pp. 71-105.

Jarvis, S., Bestgen, Y. and Pepper, S. 2013. ‘Maximizing classification accuracy in native language identification’, in *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, June 2013, Atlanta, GA, USA: Association for Computational Linguistics, pp. 111-18. http://aclweb.org/anthology//W/W13/W13-1714.pdf (last accessed on 2 April 2014).

Kochmar, E. 2011. *Identification of a Writer’s Native Language by Error Analysis*. Unpublished MA dissertation. University of Cambridge, UK. www.cl.cam.ac.uk/~ek358/Native_Language_Detection.pdf (last accessed on 2 April 2014).

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, Chicago, IL, USA, pp. 624-8. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.113.7470 (last accessed on 2 April 2014).

Kotsiantis, S. 2007. ‘Supervised machine learning: A review of classification techniques’, *Informatica Journal* 31: 249-68.

Lecocke, M. and Hess, K. 2006. ‘An empirical study of univariate and genetic algorithm-based feature selection in binary classification with microarray data’, *Cancer Informatics* 2: 313-27.

Mayfield Tomokiyo, L. and Jones, R. 2001. ‘You’re not from ‘round here, are you? Naive Bayes detection of non-native utterance text’. *Proceedings of NAACL2001: The Second Meeting of the North American Chapter of the Association for Computational Linguistics*. Carnegie Mellon University, Pittsburgh, PA, USA, 2-7 June 2001. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.8099 (last accessed on 2 April 2014).

Milton, J. C. P. and Chowdhury, N. 1994. ‘Tagging the interlanguage of Chinese learners of English’, in *Proceedings Joint Seminar on Corpus Linguistics and Lexicology*, Guangzhou and Hong Kong, 19-22 June 1993, Language Centre, HKUST, pp. 127-43. http://repository.ust.hk/dspace/handle/1783.1/1087 (last accessed on 2 April 2014).

Molinaro, A.M., Simon, R. and Pfeiffer, R.M. 2005. ‘Prediction error estimation: A comparison of resampling methods’, *Bioinformatics* 21: 3301-7.

Odlin, T. 1989. *Language Transfer: Cross-linguistic Influence in Language Learning.* Cambridge: Cambridge University Press.

Odlin, T. 2006. ‘Could a Contrastive Analysis Ever Be Complete?’, in Arabski, J. (ed.) *Cross-Linguistic Influences in the Second Language Lexicon*. Clevedon, UK: Multilingual Matters, pp. 22-35.

Paquot, M. 2013. ‘Lexical bundles and L1 transfer effects’, *New Frontiers in Learner Corpus Research, Special Issue of International Journal of Corpus Linguistics* 18(3): 391-417.

Pepper, S. 2012. *Lexical Transfer in Norwergian Interlanguage*. Unpublished Master dissertation. University of Oslo. https://www.duo.uio.no/handle/10852/34792 (last accessed on 2 April 2014)

Post, M. and Gildea, D. 2009. Language modeling with tree substitution grammars. *Proceedings of NIPS workshop on Grammar Induction, Representation of Language, and Language Learning*. Whistler, BC. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.156.3063 (last accessed on 2 April 2014)

Ringbom, H. 2007. *Cross-linguistic Similarity in Foreign Language Learning*. Clevedon: Multilingual Matters.

Schmid, H. 1995. *Tree Tagger – a Language Independent Part-of-Speech Tagger*. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Shannon, C. 1948. ‘A mathematical theory of communication’, *Bell System Technical Journal* 27: 379-423. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Spicer, J. 2005. *Making Sense of Multivariate Data Analysis: An Intuitive Approach*. Thousand Oaks, CA: Sage Publications.

Swanson, B. and Charniak, E. 2012. ‘Native language detection with tree substitution grammars’, in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Republic of Korea, 8-14 July 2012, pp. 193-7. http://aclweb.org/anthology//P/P12/P12-2038.pdf (last accessed on 2 April 2014).

Tetreault, J., Blanchard, D., Cahill, A. and Chodorow, M. 2012. ‘Native tongues, lost and found: resources and empirical evaluations in native language identification’, in *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*. December 2012, Mumbai, India: The COLING 2012 Organizing Committee, pp. 2585-602. http://aclweb.org/anthology//C/C12/C12-1158.pdf (last accessed on 2 April 2014).

Tetreault, J., Blanchard, D. and Aoife, C. 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*, June 2013, Atlanta, GA, USA: Association for Computational Linguistics. http://aclweb.org/anthology//W/W13/W13-1706.pdf (last accessed on 2 April 2014).

Tofighi, P., Kŏse, C. and Rouka, L. 2012. ‘Author’s native language identification from Web-based texts’, *International Journal of Computer and Communication Engineering* 1(1): 47-50.

Torney, R., Vamplew, P. and Yearwood, J. 2012. ‘Using psycholinguistic features for profiling first language of authors’, *Journal of the American Society for Information Science and Technology* 63(6): 1256-69.

Tsur, O. and Rappoport, A. 2007. ‘Using classifier features for studying the effect of native language on the choice of written second language words’, in *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*. Prague: Association for Computational Linguistics, pp. 9-16. http://dl.acm.org/citation.cfm?id=1629797 (last accessed on 2 April 2014).

Wong, S.-M. J. and Dras, M. 2009. ‘Contrastive analysis and native language identification’, in *Proceedings of the Australasian Language Technology Association*. Sydney, Australia: Association for Computational Linguistics, pp. 53-61. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.383.3713 (last accessed on 2 April 2014).

Wong, S.-M. 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*, Edinburgh, Scotland, UK, July 27–31, 2011, pp. 1600-10. http://aclweb.org/anthology//D/D11/D11-1148.pdf (last accessed on 2 April 2014).

Wong, S.-M. J., Dras, M. and Johnson, M. 2012. ‘Exploring adaptor grammars for native language Identification’, in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea, 12-14 July 2012, pp. 699-709. http://aclweb.org/anthology//D/D12/D12-1064.pdf (last accessed on 2 April 2014).