Robust, lexicalized native language identification

Julian BROOKE *Graeme HIRST*
University of Toronto, Department of Computer Science, Toronto, Canada
jbrooke@cs.toronto.edu, gh@cs.toronto.edu

ABSTRACT
Previous approaches to the task of native language identification (Koppel et al., 2005) have been limited to small, within-corpus evaluations. Because these are restrictive and unreliable, we apply cross-corpus evaluation to the task. We demonstrate the efficacy of lexical features, which had previously been avoided due to the within-corpus topic confounds, and provide a detailed evaluation of various options, including a simple bias adaptation technique and a number of classifier algorithms. Using a new web corpus as a training set, we reach high classification accuracy for a 7-language task, performance which is robust across two independent test sets. Although we show that even higher accuracy is possible using cross-validation, we present strong evidence calling into question the validity of cross-validation evaluation using the standard dataset.

KEYWORDS: Native language identification, text classification, evaluation.

1 Introduction

Native language identification (Koppel et al., 2005) is a task in which features of the second language (L2) texts written by non-native speakers of various different native language (L1) backgrounds are used to identify those backgrounds. One potential application is as a facet of author profiling, which can be used to identify those who misrepresent themselves online (Fette et al., 2007). Another is as a preprocessing step to language learner error correction (Leacock et al., 2010): for example, Rozovskaya and Roth (2011) use L1-specific information to improve their preposition-correction system, and recent work in collocation correction relies on the specific forms present in the writer's native language (Chang et al., 2008; Dahlmeier and Ng, 2011).

As a distinct task in computational linguistics, native language identification has been reasonably well-addressed (Koppel et al., 2005; Tsur and Rappoport, 2007; Wong and Dras, 2009), and in fact there has been a flurry of recent activity (Kochmar, 2011; Golcher and Reznicek, 2011; Wong and Dras, 2011; Brooke and Hirst, 2011; Wong et al., 2012). Though a wide range of feature types has been explored—with conflicting results—the evaluation of these feature sets has been fairly uniform: training and testing in one of several small corpora of learner essays (Granger et al., 2009; Yannakoudakis et al., 2011; Lüdeling et al., 2008), which are unfortunately quite expensive to collect. A notable problem with these corpora with respect to native language identification, however, is a clear interaction between native language and essay topic. Generally speaking, the solution in previous work has been to avoid the use of lexical features that might carry topical information, limiting feature sets to syntactic and phonological phenomena. There are two reasons to be critical of this approach. First, there are almost certainly kinds of language transfer (Odlin, 1989), i.e. transfer related to lexical choice, that are being overlooked. Second, and more troubling, is that avoiding the lexicon is not fully effective as a means of countering the effects of topic: some recent work indicates that variation in topic also has significant influence on non-lexical features (Golcher and Reznicek, 2011; Brooke and Hirst, 2011), calling into question the reliability of previous results that assume otherwise.

The approach we present here resolves this tension by requiring training and test sets that are independently sampled. Although corpora may have some form of confounding variation that may artificially inflate or (in some cases) lower performance relative to other samples from the same corpus, any variation that is consistent across very distinct corpora is likely to be a true indicator of L1. Although we test on the typical essay corpora used by other researchers, we train on a very different dataset, a large but messy corpus of journal entries scraped from a language learner website. Without the distraction of (irrelevant) topic bias, we can test the efficacy of lexical features, including *n*-grams and dependencies. We also test a number of options at the level of the classifier, most notably a multiclass support vector machine (SVM) decision-tree classifier that leverages the genetic relationships among languages, and a simple but elegant method for adapting an SVM classifier to the test corpus without integrating the confounding variation found there. Our best classifier with lexical and syntactic features provides results that compare well with previously-reported single-corpus performance; we also present, however, evidence that calls into question the validity of these previous results, showing that topic bias within the corpus is having a major effect and that indeed the performance of models built in the topic-biased ICLE corpus is not robust, regardless of the features chosen.

2 Related Work

The earliest focused work on native language detection was by Koppel et al. (2005). They classified texts from the International Corpus of Learner English (ICLE) into one of five (European) native language backgrounds using support vector machines. They described their feature set as stylistic; features included the frequency of function words, rare POS bigrams, letter *n*-grams, and spelling errors. They reported a performance of just over 80% on the task using the full feature set.

Other work on the ICLE includes that of Tsur and Rappoport (2007), who were concerned with identifying phonological language transfer; they focused on the construction of character *n*-gram models, reporting 66% accuracy with just these sub-word features, with only a small drop in performance when the dominant topic words in each sub-corpus (as identified using *tf-idf*) were removed. Wong and Dras (2009) investigated particular types of syntactic error: subject-verb disagreement, noun-number disagreement, and determiner problems, relating the appearance of these errors to the features of relevant L1s. However, they reported that these features do not help with classification, and they also note that character *n*-grams, though effective on their own, are not particularly useful in combination with other features. In follow-up work, Wong and Dras (2011) attained the best results to date, 80% performance on a 7-language task, by including syntactic production rules. Recent work by Wong et al. (2012) and Swanson and Charniak (2012) has explored the use of statistical grammatical induction techniques—Adaptor Grammars in the former case, Tree Substitution Grammars in the latter—to select better syntatic features for classification.

The work of Kochmar (2011) is distinct from those above in a number of ways: she used a different corpus of essays, derived from the Cambridge Learner Corpus[1], and concentrated on pairwise (SVM) classification within two European language sub-families. An exhaustive feature analysis indicated that character *n*-gram frequency is the most useful feature type for her task; unlike Wong and Dras (2011), syntactic production rules provided little benefit. With respect to lexical features, Kochmar presented some results using word *n*-grams, but regarded them as attributable to topic bias in the corpus. Error-type features (e.g. spelling, missing determiner) as provided by the corpus annotation offered little improvement over the high performance offered by the distributional features (e.g. POS/character *n*-grams).

Golcher and Reznicek (2011) used a string-distance metric to identify the native language of German learners in the Falko corpus (Lüdeling et al., 2008), and contrasted this with a topic classification task in the same corpus. Even after taking steps to mitigate topic bias (removing the influence of the words in the title), the usefulness of the three feature types that they investigated (word token, word lemma, and POS) was remarkably similar across the two tasks, with the word features dominating in both cases. Surprisingly, the effect of POS was higher in topic classification than it is on L1-classification. Earlier work of ours (Brooke and Hirst, 2011) also tested the confounding effect of topic in the context of native language identification. To motivate the use of new corpora for future research, we segregated a portion of the ICLE by topic and found that all the core features used by Koppel et al. for L1-identification showed significant drops in performance when topic-segregated 2-fold cross-validation is compared to standard (randomized) 2-fold cross-validation. This was particularly true of character *n*-grams.

[1] http://www.cup.cam.ac.uk/gb/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus

Finally, we note that native language identification has also been included as an element of larger author profiling studies (Estival et al., 2007; Garera and Yarowsky, 2009). A closely related task is the identification of translated texts and/or their language of origin (Baroni and Bernardini, 2006; van Halteren, 2008; Koppel and Ordan, 2011), though the tasks are distinct because the learners included in native language identification studies are usually at a level of linguistic proficiency below that of a professional translator (who in any case may be writing in his or her L1, rather than an L2) and are not operating under the requirement of faithfulness to some original text. Distinguishing whether or not a text is non-native (Tomokiyo and Jones, 2001) is also a related task, but most work in the area of L1 identification, including ours, assumes that we already know that a text was produced by a non-native speaker.

3 Corpora

Our training corpus is a set of 154,702 English journal entries collected from the Lang-8 language learner website.[2] In all, 65 L1s are represented, but only 14 languages have more than 1000 entries, with Asian languages being overrepresented (the website is based in Japan). Users may write whatever they want in their journal, and there are a variety of text types (some learners post assignments or translation questions, for instance), though the majority can be described as short personal narratives. The English proficiency of these learners varies widely; other users of the site can comment on journal entries and offer suggestions to improve the text. In the corpus, the entries are tagged for username, title, native language (which is provided by the user when they register), and date and time of posting. The average length of an entry is about 150 word tokens after our pre-processing, which strips HTML tags and non-ASCII characters prior to parsing—words with non-ASCII characters are ultimately disregarded during feature selection. Since these texts are relatively short compared to our test sets, for our purposes here we append consecutive short texts of writers with the same L1 (often the same author) until they are at least 250 tokens in length, which results in an average length of 431 tokens.

The International Corpus of Learner English (Granger et al., 2009), or ICLE, is a set of 6085 non-native speaker essays collected from university students at institutions around the world; v2.0 represents includes 16 L1 backgrounds, mostly European. Other variables tagged in the corpus are topic, genre (argumentative or literary), setting (timed writing or not), age, gender, and educational institution, all of which vary in unpredictable ways throughout the corpus. As already mentioned, a major problem with the ICLE is topic variation, which is both unnaturally strong and often arbitrary; for practical reasons, the different parts of the corpus were collected by EFL instructors in different countries, who chose a small, often fairly distinct set of topics for their students (of a particular L1 background) to write about. To investigate the proficiency level of students, the creators tested a sample of each native language for English level using the *Common European Framework* (CEF), showing that while learners in the corpus are generally at least of intermediate proficiency, the percentage of advanced learners is very different for different L1 backgrounds, another potential confound. The average text length in the ICLE is 617 words.

Our third learner corpus is a small sample of the First Certificate in English (FCE) portion of the Cambridge Learner Corpus, which has recently been released for the purposes of

[2]The URL is http://lang-8.com. We do not have permission to distribute the corpus directly; following Sharoff (2006), we will release a list of web URLs and software which can be used to recreate the corpus.

L1	Corpus		
	Lang-8	ICLE	FCE
Japanese	59156	366	81
Chinese	38044	982	66
French	1414	347	146
Spanish	3080	251	200
Italian	1072	392	76
Polish	1549	365	76
Russian	7159	276	83

Table 1: Number of texts in learner corpora, by L1.

essay scoring evaluation (Yannakoudakis et al., 2011); 16 different L1 backgrounds are represented. Each of the 1244 texts consists of two short answers in the form of a letter, a report, an article, or a short story, each tagged with the score provided by a trained examiner. The texts are also marked for specific usage errors, though we stripped this information in our pre-processing step. The average length of the texts in the FCE corpus is 428 words, or about 200 words less than the ICLE.

For this study, we selected the seven languages which had sufficient numbers in all three corpora, i.e. at least 1000 texts in the Lang-8 corpus, 200 texts in the ICLE, and 50 texts in the FCE. Table 1 shows, for each L1, the number of texts present in each corpus. For testing in the ICLE, we use 200 from each set, and a separate set of 50 per L1 is used for our bias adaptation method. For testing in the FCE, we use 50 texts, and 15 texts for bias adaptation.

4 Classifier Experiments

We split our main experiments into two parts. In our initial investigation, we found that using the full set of feature types, to be described later in Section 5, provided near-optimal results. Given that exploring the exhaustive set of combinations is not feasible in this space, we elect to first take the full feature set as fixed and turn our attention to higher-level classifier options, establishing the best among those options before we proceed with a feature analysis.

4.1 Classifier Options

Our experiments included testing the following options:

Balanced training (bal) vs. cost weight (cost) Statistical classifiers generally depend on having similar class distributions in training and testing sets, an assumption which is violated here. There are two simple ways to handle this problem: either balancing the training sets by discarding extra training data, or training the classifier with using different cost weights for different classes, promoting classification of rarer classes to the level expected in the (balanced) test data. We use the cost weight equation from Morik et al. (1999).

Binary (bin) vs. frequency (freq) features Previous work has mostly used normalized frequency rather than binary occurrence in a text as the feature value used for classification; Wong and Dras (2011) are an exception, but they do not justify that choice.

SVM vs. MaxEnt classifier Support vector machines were a popular option in previous work, but Wong and Dras (2011) report better performance with a Maximum Entropy (MaxEnt) classifier. A full discussion of these two machine learning methods is omitted here, though we note that (pairwise) SVMs are generally conceptualized as a hyperplane which maximizes the margin between classes in the feature space, while MaxEnt is a multinomial logistic model built by constrained maximization of the probability of the training data. For SVM classification (see below), we use LIBLINEAR (Fan et al., 2008), which is optimized for linear kernel classification of large datasets; except as explicitly mentioned below, we present results using default parameter settings (which were found to give good results). Feature vectors are normalized to the unit circle (Graf and Borer, 2001). For MaxEnt we follow Wong and Dras (2011) in using MegaM.[3]

Regularization parameters In the context of building a robust classifier for cross-corpus classification, the regularization of the model (Alpaydin, 2010), i.e. the degree to which the classifier increases in complexity to fit the training data, is of obvious relevance. For SVMs, the key parameter is C, which controls the penalty for misclassified examples in the training set: a large value of C means these errors have a higher influence on the objective function, promoting more complex models that minimize error but may result in overfitting. For the MaxEnt classifier, the λ parameter controls the influence of a Gaussian prior on the feature weights: low values of λ correspond to an imprecise prior, allowing the feature weights to fit the data. We tuned the corresponding parameter for each classifier configuration using 7-class task performance in the development set for each test corpus.[4]

Multiclass SVM type While MaxEnt has a natural multiclass interpretation, an SVM decision plane is appropriate only for binary choice. A standard approach to multiclass SVM is to combine multiple pairwise SVM classifiers (Hsu and Lin, 2002). Two general options in this vein are *one vs. one* (1v1), where $n(n-1)/2$ individual classifiers (for n classes), each trained on one pair of classes, are combined, and *one vs. all* (1va), where n classifiers are trained by separating one class from all the others. The winner of 1va is obviously the class with the highest margin (distance from the decision plane), but for 1v1 it is typically the class which is chosen by the most classifiers (ties are broken in favor of the highest margins). A third, novel option is made possible by the genetic relationships among languages in our test set: an SVM binary decision tree (tree), presented in figure 1.[5] Note that tree classifiers have a significant performance advantage over both 1va and 1v1 classifiers with respect to the number of classifiers required ($n-1$), and an advantage over the 1va classifiers with respect to the average size of the training sets used to build those classifiers. Finally, Crammer and Singer (2002) have proposed a multiclass SVM classifier based on class prototypes (pro) rather than hyperplane boundaries, and we also test this option (as implemented in LIBLINEAR).

Bias adaption, pairwise (adS) Since there are significant differences in the genre, domain, and quality of texts across our training and test corpora, some form of domain adaption (Daumé and Marcu, 2006; Bruzzone and Marconcini, 2010) would almost certainly be

[3] http://www.cs.utah.edu/~hal/megam/

[4] Since the C parameter is selected once for each configuration based on the 7-class task, some results that we would otherwise expect to be equivalent, e.g. the 2-class SVM classifers, actually vary slightly.

[5] There is some controversy in the literature about the genetic relationship amongst Romance languages; see the discussion by Kochmar (2011).

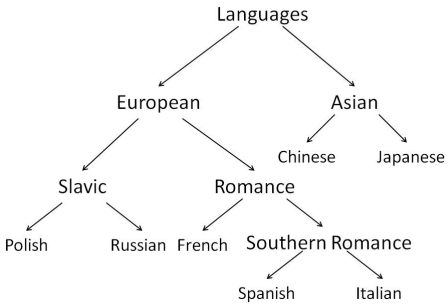

Figure 1: Binary decision tree for SVM experiments

helpful. However, even unsupervised forms of transfer learning (Pan and Yang, 2010) are likely to take advantage of those confounding factors that prompted us to reject within-corpus evaluation; we believe that any change to the feature weights based on samples from the same corpus that our test set is drawn from is ultimately self-defeating in this context. However, there is one key parameter to these that is not a feature weight: the bias. In pairwise SVM, changing the bias slides the hyperplane, changing only the total number of positive (or negative) features required to make a classification, not the individual influence of a particular feature (i.e. the sign of a feature weight). With respect to its effect (changing the balance of classes), it is closely related to our cost factor option above; however, whereas the cost factor is a parameter used during training, we shift the bias using our own iterative process after the model is built, using a sample from the same corpus as the test set (a development set).[6] Our algorithm is as follows: we first initialize our step size to the absolute value of the original bias, and then we iteratively modify the bias, adding or subtracting the present step size such that we are moving in the direction of a distribution where the ratio of classes predicted in our development set is the same as in the final test set, reclassifying the data after each step.[7] If we overshoot the desired ratio, we halve the step size, and continue until we reach the desired ratio or the predicted ratio does not change for 10 iterations. We do this separately for each test set with the corresponding development set.

Bias adaption, multi (adM) The MaxEnt and SVM prototype classifiers also have bias terms that can be optimized, but unlike the pairwise classifiers they cannot be dealt with one at a time; optimizing the bias for one class will affect the others in unpredictable ways. We proceed with the same basic algorithm as the pairwise classifier, but we do this for all bias terms simultaneously, i.e. all biases are adjusted in a single step. Each bias has a separate step size, and the optimization ends when the entire distribution is correct or nothing has changed in 10 iterations. We also implemented this for SVM 1va, i.e. interpreting it as a single multiclass classifier rather than a set of pairwise classifiers.

[6]Admittedly, we could accomplish this with additional parameter tuning, but there are both practical and principled reasons for doing it this way: it is much faster to modify the biases directly rather than retraining the model, and, more importantly, we want to preserve the original feature weights; we require that they do not reflect exposure to the confounds of the testing corpus in any way.

[7]This requires knowledge of that distribution. However, it is otherwise unsupervised in that we are only concerned with the distribution of predictions: we do not use the true class values except to create the appropriate subsets for the SVM 1v1 and SVM tree classifiers.

Configuration	Asian		European		All	
	ICLE	FCE	ICLE	FCE	ICLE	FCE
Chance baseline	50.0	50.0	20.0	20.0	14.3	14.3
(1) SVM *1v1* cost bin	95.2	86.0	50.0	40.4	58.7	50.3
(2) SVM *tree* cost bin	95.2	86.0	48.7	41.3	59.4	49.4
(3) SVM 1va cost bin	**96.5**	86.0	54.8	44.0	61.6	50.8
(4) SVM *pro* cost bin	95.0	85.0	55.6	42.8	62.4	50.8
(5) *MaxEnt* cost bin	95.0	85.0	56.6	44.8	63.7	42.3
(6) SVM *tree*-adS cost bin	95.2	**88.0**	64.4	57.2	73.7	57.4
(7) *MaxEnt*-adM cost bin	95.0	86.0	68.2	64.4	74.0	60.8
(8) SVM *1v1*-adS cost bin	95.5	88.0	67.9	66.8	74.2	65.7
(9) SVM 1va-*adS* cost bin	95.0	88.0	71.6	67.6	77.8	**66.5**
(10) SVM *pro*-adM cost bin	95.7	87.0	71.1	66.4	77.3	64.0
(11) SVM 1va-adM cost bin	95.0	86.0	**71.7**	**68.0**	**78.0**	65.7
(12) SVM 1va-adM *bal* bin	79.8	75.0	63.1	60.4	66.8	59.1
(13) SVM 1va-adM cost *freq*	95.2	83.0	66.8	57.6	74.9	53.1

Table 2: Native language classification accuracy (%) for varying classifier options. Bold indicates best result in column, italics indicates difference from the pivot classifier (11).

4.2 Results

Table 2 shows the results of our experiments. In addition to the full 7-language task accuracy (the 'All' columns), we also present results classifying the two major subgroups; note that these are distinct tasks, e.g. for European it is the accuracy of a 5-language task, not the accuracy of the classification of those 5 languages within the 7-language task (see Figure 1 for our language classification schema). However, in our discussion, we focus on results for the full 7-language task. The upper part of Table 2 includes various key classifier options, ordered by their 7-way ICLE accuracy, while the bottom includes other options; the best classifier (11) is used as a pivot between the two.[8] The aspect(s) of the configuration that are different from the pivot are in italics, and the best results in each column are in bold. For each classifier, we report the results using the best C or λ values from an initial series of runs using the development set.

Unsurprisingly, we see better results when we use all the data at our disposal (11), rather than forcing balanced test cases (12). This result is useful, though, because it indicates that our consistently high performance in distinguishing Chinese and Japanese elsewhere in Table 2 is a result of that extra data, and not other factors, i.e. the fact that unlike our other language groupings, Chinese and Japanese do not belong to a single genetic language family (Comrie, 1987). Also clear is the preference for binary (11) rather than frequency-based (13) feature values: one possible explanation is that, in these relatively short texts, there is high variability in normalized frequencies, and a simpler metric, by having less variability, is easier for the classifier to leverage. In general, slightly less regularization (high C, low λ) values were preferred, though most were reasonably close to the default values; tuning made little difference, particularly for the SVM classifiers.

[8]The effects of the options in each of the two parts of the table are fairly independent, so for simplicity of presentation we test them separately.

Between the two main classifier types, the MaxEnt classifier was, with the appropriate choice of λ (5), the best performing classifier in the ICLE when no bias adaption was used; it was, however, worse than almost all of our SVM options in the main 7-language classification task when bias tuning was allowed (7). This does not appear to be a failure of the adaption algorithm, but rather a real distinction between the two classifiers: our experience is that the SVM classifiers are less robust, i.e. more prone to errors when training and test sets differ significantly, but they can be easily recalibrated for optimal performance with a relatively small amount of information. Here, we show that changing the bias alone is enough for major gains across all the SVM types (6,8–11), results which are statistically significant.

Our novel binary tree classifier (2,6) is competitive but ultimately performs poorly compared than other options, suggesting that the simplicity of the classifier does come with a trade-off in performance. The 1va classifiers (3,9,11) are consistently better than 1v1 (1,8), while the performance of the prototype-based SVM (4,10) is nearly indistinguishable from 1va. This is somewhat surprising, since we might expect a 1v1 or prototype approach to be able to better deal with the commonalities and differences among languages than the 1va, which lumps diverse languages into a single 'other' category. With respect to the 1va classifier, it does not seem to matter much whether pairwise (9) or single classifier (11) bias tuning is used; the latter gave us the best 7-class performance in the ICLE (and we use it as our best classifier), but the former gave slightly better performance in the FCE. In the ICLE, the difference between the best bias-adapted 1va classifier and the 1v1, tree, and MaxEnt classifiers is statistically significant (χ^2 test, $p < 0.001$).

5 Feature Analysis

5.1 Features

Our model includes the following feature types:

Function words A common feature in stylistic analysis. Our list of 416 common English words comes from the LIWC (Pennebaker et al., 2001).

Character *n*-grams (unigrams, bigrams, and trigrams) For bigrams and trigrams, the beginning and end of a word are treated as special characters.

Word *n*-grams (unigrams and bigrams) Note that word *n*-grams are a superset of function words. Punctuation is included.

POS *n*-grams (unigrams, bigrams, and trigrams) POS tagging is provided by the Stanford Parser V1.6.9 (Klein and Manning, 2003), also used by Wong and Dras (2011).

POS/function mixture *n*-grams (bigrams and trigrams) Wong et al. (2012) report better results with POS *n*-grams that retain the identity of individual function words rather than using their part of speech.

CFG productions Context-free grammar production rules, as provided by the Stanford parser. Lexical production rules are not included.

Dependencies Dependencies consist of two lexical items and the syntactic relationship between them. Also produced by the Stanford parser (de Marneffe et al., 2006).

Features	Asian		European		All	
	ICLE	FCE	ICLE	FCE	ICLE	FCE
Chance baseline	50.0	50.0	20.0	20.0	14.3	14.3
(1) Function words	72.7	71.0	40.3	37.2	35.6	36.0
(2) Character *n*-grams	78.3	63.0	37.5	28.8	37.4	22.6
(3) POS *n*-grams	86.8	78.0	47.9	50.0	52.9	44.3
(4) POS/function *n*-grams	93.3	85.0	60.6	56.8	67.4	59.4
(5) CFG productions	78.5	72.0	46.9	43.2	49.7	41.1
(6) Dependencies	94.0	79.0	49.8	46.8	61.4	45.1
(7) Word *n*-grams	94.3	89.0	71.1	66.8	77.1	68.3
(8) Syntactic Features	94.3	87.0	60.1	61.2	68.1	65.1
(9) Lexical Features	95.2	86.0	71.0	67.6	77.8	67.1
(10) Lexical+Syntactic	**96.0**	90.0	72.3	66.4	78.4	68.2
(11) All features	95.0	86.0	71.7	68.0	78.0	65.7
(12) (4)+(7)	95.5	90.0	**72.5**	66.8	**79.3**	**70.0**
(13) (4)+(7), no proper nouns	94.5	87.0	69.6	67.2	76.5	65.7
(14) (4)+(7), df ≥ 20	95.0	86.0	71.3	**68.4**	77.3	65.4
(15) (4)+(7), IG > 0	89.5	**93.0**	69.5	66.4	76.5	65.7

Table 3: Native language classification accuracy (%), by feature set. Bold indicates best result in column.

Syntactic Features POS *n*-grams, POS/function mixture *n*-grams, and CFG productions.

Lexical Features Word *n*-grams and dependencies.

Proper Nouns Not actually a separate feature, proper nouns are included by default in character and word *n*-grams as well as dependencies. They are obviously relevant to the task, but there are applications (e.g. forensic profiling) where they might not be appropriate, since they do not directly indicate language transfer from the L1 but rather reflect real-world correlations between native language and country of residence, etc. Here, we report results with all proper nouns excluded from consideration for all relevant features.

Feature Selection Wong and Dras (2011) tested feature selection based on information gain, but it provided no improvement in performance. For practical reasons, we have included by default a simple frequency-based feature selection; only features that appear in 5 different texts in the training set are included. Even with this restriction, our feature set has almost 800,000 features. Here, we test the effect of a higher frequency cutoff (at 20), and limiting our set to features with positive information gain.

5.2 Analysis

Again, we focus on the results of the full 7-language task (the 'All' columns). Clearly, all the feature types can be used to distinguish native language: each of the results in Table 3 is well above a chance baseline, though function words (1) and character *n*-grams (2) give a fairly modest performance individually. Compared to these, production (5) rules are markedly more useful, a result which is compatible with the conclusions of Wong and

Dras (2011). Nonetheless POS (3) and in particular mixed POS/function words *n*-grams (4) offer even better performance, despite being somewhat simpler. Compared to the latter of these, the usefulness of lexical dependencies (6) is muted, and shows a very inconsistent performance across the two test sets. Word *n*-grams (7), however, alone account for almost all of the accuracy we see when all features are combined.

Adding the POS features and CFG productions (8) generally boosts performance, suggesting that the syntactic features may not be entirely redundant, while the combination of the lexical features also provides a small improvement in the 7-language ICLE task, though the FCE is worse (9). Further adding the syntactic features to the lexical features increases performance for most of the tasks (10), while including character *n*-grams tends to degrade performance (11). Finally, we exhaustively tested feature combinations and found that the best performing for the 7-language task used only the two best individual feature types, POS/function word mixtures and lexical *n*-grams, though the differences among all the options containing lexical *n*-grams are not statistically significant (12).

When we remove proper nouns (13), there is a modest drop in performance, indicating that they had some positive role in the classification, but the benefits of using lexical features goes well beyond proper nouns. Additional frequency-based feature selection (14) has a small, mostly negative effect, as does restricting features to those with positive information gain (15). In general, we see no evidence that a simpler model is preferred in this case, though if speed is a concern one can be used without too much loss.

We also looked briefly at the individual lexical features that were useful based on their information gain in the training set. One thing that was immediately evident is that some common, entirely correct English words and expressions were extremely helpful for distinguishing native languages. For example, the phrase *decide to* was ranked high: we note that in at least one language in our set (French), a closely analogous cognate construction *decider de* exists, whereas another language, Chinese, has no analogous construction, since the verb that most closely means *decide to* (*jueding*) is phonetically dissimilar, has no element corresponding to *to*, is more common as a noun, and in fact is pragmatically associated only with major decisions, often in a legal context (closer to the English *make a decision to*). By default, learners will prefer forms that correspond to those from their L1 (Odlin, 1989), and lexical features are key to identifying this kind of language transfer.

6 ICLE-training Experiments

One of the primary motivations for our cross-corpus approach to NLI is the confounding variation found in the ICLE corpus. In this section, we turn to using the ICLE as a training corpus in order to highlight these problems, particularly those relevant to 'stylistic' features, which have been thought of as immune to these effects. The first experiment, the results of which are presented in Table 4, consists of two types of 2-fold cross-validation in the ICLE corpus: the first is standard, randomized cross-validation, while in the second, the two folds (of 700 texts each) are segregated by essay prompt; essays based on a given prompt are in one fold or the other.[9] For this we use the 1va classifier without any bias adaption, which is unnecessary in the case of cross-validation.

Within the ICLE, we see in the 'Difference' column of Table 4 the consistent effects of essay

[9]This experiment is possible only in the ICLE, since titles in the Lang-8 are freely chosen by each writer, and there is little variety of prompts in the FCE.

Features	Random	Segregated	Difference
Chance baseline	14.3	14.3	–
(1) Function words	58.0	46.7	−11.3
(2) Character *n*-grams	51.2	48.2	−3.0
(3) POS *n*-grams	83.3	72.2	−11.1
(4) POS/function *n*-grams	87.6	79.2	−10.4
(5) CFG productions	86.1	79.7	−6.4
(6) Dependencies	89.1	77.1	−12.0
(7) Word *n*-grams	94.3	81.3	−13.0
(8) All (1–7)	90.4	81.6	−8.8

Table 4: ICLE within-corpus experiment classification accuracy (%), by feature set.

prompt on classification, across all kinds of features. The effects on lexical features (6,7) are, not surprisingly, most pronounced, but other popular features are also implicated to varying degrees. The effectiveness of various features under both conditions roughly mirrors the results in the previous section, though there are a few notable exceptions: for instance, production rules (5) were more useful here than in the Lang-8 trained cross-corpus experiments; this is interesting since many of the most recent results in the ICLE (Wong and Dras, 2011; Swanson and Charniak, 2012) make use of these grammatical features. Surprisingly, character *n*-grams were the least affected, a contrast from our earlier work on ICLE topic bias (Brooke and Hirst, 2011), though there remains little doubt that they are inferior features for this task. Lexical *n*-grams are ultimately the most preferred feature (7), even when topic effects are partially[10] controlled for.

We also present cross-corpus experiments with the FCE and a language-balanced 150-text portion of the Lang-8 corpus as test sets. As with our training set, this test set consists of combined texts, this time with a minimum length of 500, making the texts of comparable length to those in the ICLE. We create another set of 50 texts for bias adaption. In the latter experiment, we also include a special set of features: the POS/function mixture 5-grams which were selected by the adaptor grammars of Wong et al. (2012), providing superior performance over exhaustive enumerations. Since these features were derived from the ICLE, they could not be defensibly used in other experiments (i.e. with the ICLE as a test set), but we can test their usefulness here. Since the original experiment involved cross-validation, there are in fact 5 different sets; our set consists of the union of these sets.[11]

The cross-corpus results in Table 5 are strikingly lower than the within-ICLE results. They also compare poorly to our earlier cross-corpus results in this paper. Part of this difference is, of course, the effect of the much-larger Lang-8 dataset, though the balanced result in Table 2 (12), uses a very similar amount of data (as measured in tokens) from the Lang-8 but attains a much better FCE classification accuracy (roughly 20% better). The POS/function mixture

[10]There are more pervasive topic and genre effects that segregating by prompt does not resolve. For instance, a large number of the Japanese texts are personal narratives, each with a different title, while in the Russian texts there is a particular focus on the literature of various authors, and in the Chinese texts there is a discussion of the advantages or disadvantages associated with certain government policies.

[11]We originally intended to take the intersection, but in fact the intersection of the feature sets is empty; no single feature was useful in every fold.

Features	Lang-8	FCE
Chance baseline	14.3	14.3
(1) Function words	27.6	20.0
(2) Character *n*-grams	29.7	24.0
(3) POS *n*-grams	37.0	32.8
(4) POS/function *n*-grams	40.2	33.4
(5) CFG productions	32.5	31.4
(6) Dependencies	30.7	25.1
(7) Word *n*-grams	50.8	35.7
(8) All (1–7)	46.8	39.1
(9) Adaptor grammar *n*-grams	40.9	30.8

Table 5: ICLE-training cross-corpus classification accuracy (%), by feature set.

features (9) derived using adaptor grammars do reasonably well, but are only marginally better than exhaustive mixture features (4) in the Lang-8 test set, and are markedly worse than a number of other features in the FCE. Again, lexical *n*-grams (7) are obviously the best individual feature type.

7 Discussion

The results in the previous section highlight the problematic nature of within-corpus evaluation in general, and the inadequacy of the ICLE as a training corpus in particular. It is unclear to what extent previous results on this task are influenced by these effects, but we believe there is at least reason to be skeptical of some of the conclusions. In particular, sophisticated feature selection techniques which have been the focus of recent work may result in models which perform better in the ICLE, but which have little or no benefit beyond that particular corpus. We believe more attention should be paid to the overall validity of NLI experiments, rather than to specific technical approaches. One interesting open question is whether features such as proper nouns, which are of obvious but somewhat trivial benefit, should be excluded. Certainly, we would argue that lexical features in general are far too important to the task to simply be discarded; our experiments here suggest that their usefulness goes well beyond proper nouns and is not simply a reflection of topic.

Though higher performance is clearly possible using cross-validation, our Lang-8 trained model does reasonably well in both our testing corpora; the results are fairly consistent, and the difference can be attributed to the smaller size of the FCE texts. It is clear that factors such as the choice of classifier and the size of the dataset play some role, though the most obvious improvement came from the use of our bias adaptation technique, which uses a small amount of data from a test corpus to improve the model; this was particularly effective for SVMs. Importantly, this method keeps the feature weights constant, a necessary precondition when the testing corpus has known arbitrary biases. Given the variation in text size, genres, and learner proficiency, some kind of adaption is clearly necessary to get competitive results, though our experiments using it with the ICLE as training data suggest the method cannot overcome a problematic training set.

We note that our still sizable error rate on this task may in fact be due to a learner proficiency effect; on inspection, one of the authors (a native speaker of English) found that some of

the European texts were nearly indistinguishable from native writing. As suggested by the statistics provided in the ICLE manual (Granger et al., 2009), many of these learners are highly proficient, and thus they might have completely integrated the norms of their L2, making them legitimately indistinguishable. We also tested the correlation between essay scores in the FCE and our classification accuracy, and found a small negative correlation, suggesting that those who scored better were harder to classify; text length, though, was a confounding factor, since longer texts got better scores and are also easier to classify. Finally, we also noticed that French was the most consistently misclassified language, by a significant margin; this could be due, in part, to the historical connection between French and English that makes French L1 transfer somewhat less distinct, whereas distant languages like Chinese and Japanese are easy discerned, an effect we saw even when the training sets were balanced. In general, we think the relationship between proficiency, distances between languages, and L1 classification merits further study.

One important strength of the current work is the training dataset, which, unlike many learner corpora resources, is fully accessible via the web (and growing!). The coverage of European languages is poor, however, and since large amounts of data are necessary to fully leverage the potential of lexical features, one future direction would be to look for even more inexpensive ways of finding learner texts, perhaps by collecting English texts that appear on otherwise non-English websites. Armed with larger datasets, we would like to move beyond classification of a handful of L1s, moving towards a system that can identify influence from a full range of common L1 backgrounds.

8 Conclusion

In this paper, we have demonstrated the feasibility of a cross-corpus approach to the development of native language identification systems, sidestepping the problem of within-corpus confounds to test the efficacy of relevant options. The most striking result is the dominance of shallow lexical features in our best classifier, even in comparison to high-performing, sophisticated feature types such as syntactic productions; also important is some degree of domain adaption, and we present a simple but highly effective method. We have also highlighted the not-insignificant role that other choices play in the classifier performance; for instance, the *one vs. all* classifier, which has been somewhat maligned in SVM multiclass comparisons (Hsu and Lin, 2002; Duan and Keerthi, 2005), provides the best performance in both test corpora when our simple bias adaption method is applied. We have also presented new evidence that within-corpus evaluations techniques are problematic, and that the validity of results that use the ICLE in this manner need to be re-evaluated.

Acknowledgements

Thanks to Gabriel Murray for bringing the Lang-8 website to our attention, and Jojo Wong and Mark Dras for providing their adaptor grammar features. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada.

References

Alpaydin, E. (2010). *Introduction to Machine Learning*. The MIT Press, 2nd edition.

Baroni, M. and Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21:259–274.

Brooke, J. and Hirst, G. (2011). Native language detection with 'cheap' learner corpora. Presented at the 2011 Conference of Learner Corpus Research (LCR2011).

Bruzzone, L. and Marconcini, M. (2010). Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:770–787.

Chang, Y.-C., Chang, J. S., Chen, H.-J., and Liou, H.-C. (2008). An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning*, 21(3):283–299.

Comrie, B., editor (1987). *The World's Major Languages*. Oxford University Press, Oxford.

Crammer, K. and Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292.

Dahlmeier, D. and Ng, H. T. (2011). Correcting semantic collocation errors with L1-induced paraphrases. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pages 107–117, Edinburgh, Scotland, UK.

Daumé, H. and Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126.

de Marneffe, M.-C., MacCartney, B., and Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*, Genova, Italy.

Duan, K.-B. and Keerthi, S. S. (2005). Which is the best multiclass SVM method? An empirical study. In *Proceedings of the Sixth International Workshop on Multiple Classifier Systems*, pages 278–285.

Estival, D., Gaustad, T., Pham, S. B., Radford, W., and Hutchinson, B. (2007). Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING '07)*, pages 263–272.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Fette, I., Sadeh, N., and Tomasic, A. (2007). Learning to detect phishing emails. In *Proceedings of the 16th International World Wide Web Conference (WWW '07)*, pages 649–656, Banff, Alberta, Canada.

Garera, N. and Yarowsky, D. (2009). Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP '09)*, pages 710–718, Singapore.

Golcher, F. and Reznicek, M. (2011). Stylometry and the interplay of title and L1 in the different annotation layers in the Falko corpus. In *Proceedings of Quantitative Investigations in Theoretical Linguistics 4*, Berlin.

Graf, A. B. A. and Borer, S. (2001). Normalization in support vector machines. In *Proceedings of the 23rd DAGM-Symposium on Pattern Recognition*, pages 277–282.

Granger, S., Dagneaux, E., Meunier, F., and Paquot, M. (2009). *International Corpus of Learner English (Version 2)*. Presses Universitaires de Louvain, Louvain-la-Neuve.

Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13:415–425.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430.

Kochmar, E. (2011). Identification of a writer's native language by error analysis. Master's thesis, University of Cambridge.

Koppel, M. and Ordan, N. (2011). Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, Portland, Oregon.

Koppel, M., Schler, J., and Zigdon, K. (2005). Determining an author's native language by mining a text for errors. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05)*, pages 624–628, Chicago, Illinois, USA.

Leacock, C., Chodorow, M., Gamon, M., and Tetreault, J. (2010). *Automated Grammatical Error Detection for Language Learners*. Morgan & Claypool.

Lüdeling, A., Doolittle, S., Hirschmann, H., Schmidt, K., and Walter, M. (2008). Das Lernerkorpus Falko. *Deutsch als Fremdsprache*, 42:67–73.

Morik, K., Brockhausen, P., and Joachims, T. (1999). Combining statistical learning with a knowledge-based approach — a case study in intensive care monitoring. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 268–277.

Odlin, T. (1989). *Language Transfer*. Cambridge University Press.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10).

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC2001 Manual*. Erlbaum Publishers, Mahwah, NJ.

Rozovskaya, A. and Roth, D. (2011). Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*, Portland, Oregon.

Sharoff, S. (2006). Open-source corpora: using the net to fish for linguistic data. *International Journal of Corpus Linguistics*, 11(4):435–462.

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 (ACL '12)*, pages 193–197, Jeju, Korea.

Tomokiyo, L. M. and Jones, R. (2001). You're not from 'round here, are you?: naïve Bayes detection of non-native utterance text. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL '01)*, pages 1–8, Pittsburgh, Pennsylvania.

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 (CACLA '07)*, pages 9–16, Prague, Czech Republic.

van Halteren, H. (2008). Source language markers in EUROPARL translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING '08)*, pages 937–944, Manchester, UK.

Wong, S.-M. J. and Dras, M. (2009). Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 53–61.

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 (EMNLP '11)*, pages 1600–1610, Edinburgh, Scotland, UK.

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 (EMNLP-CoNLL '12)*, Jeju, Korea.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 180–189, Portland, Oregon.