Contrastive Analysis and Native Language Identification

Sze-Meng Jojo Wong
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
szewong@science.mq.edu.au

Mark Dras
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
madras@science.mq.edu.au

Abstract

Attempts to profile authors based on their characteristics, including native language, have drawn attention in recent years, via several approaches using machine learning with simple features. In this paper we investigate the potential usefulness to this task of contrastive analysis from second language acquisition research, which postulates that the (syntactic) errors in a text are influenced by an author's native language. We explore this, first, by conducting an analysis of three syntactic error types, through hypothesis testing and machine learning; and second, through adding in these errors as features to the replication of a previous machine learning approach. This preliminary study provides some support for the use of this kind of syntactic errors as a clue to identifying the native language of an author.

1 Introduction

There is a range of work that attempts to infer, from some textual data, characteristics of the text's author. This is often described by the term *authorship profiling*, and may be concerned with determining an author's gender, age, or some other attributes. This information is often of interest to, for example, governments or marketing departments; the application that motivates the current work is profiling of phishing texts, texts that are designed to deceive a user into giving away confidential details (Fette et al., 2007; Zheng et al., 2003).

The particular characteristic of interest in this paper is the native language of an author, where this is not the language that the text is written in. There has been only a relatively small amount of other research investigating this question, notably Koppel et al. (2005), Tsur and Rappoport (2007), Estival et al. (2007), and van Halteren (2008). In general these tackle the problem as a text classification task using machine learning, with features over characters, words, parts of speech, and document structures. Koppel et al. (2005) also suggest syntactic features, although they do not use them in that work.

The goal of this paper is to make a preliminary investigation into the use of syntactic errors in native language identification. The research drawn on for this work comes from the field of *contrastive analysis* in second language acquisition (SLA). According to the contrastive analysis hypothesis formulated by Lado (1957), difficulties in acquiring a new (second) language are derived from the differences between the new language and the native (first) language of a language user. Amongst the frequently observed syntactic error types in non-native English which it has been argued are attributable to language transfer are subject-verb disagreement, noun-number disagreement, and misuse of determiners. Contrastive analysis was largely displaced in SLA by *error analysis* (Corder, 1967), which argued that there are many other types of error in SLA, and that too much emphasis was placed on transfer errors. However, looking at the relationship in the reverse direction and in a probabilistic manner, contrastive analysis could still be useful to predict the native language of an author through errors found in the text.

The structure of this paper is twofold. Firstly, we explore the potential of some syntactic errors derived from contrastive analysis as useful features in determining the authors' native language: in particular, the three common types of error mentioned above. In other words, we are exploring the contrastive analysis hypothesis in a reverse direction.

Secondly, our study intends to investigate whether such syntactic features are useful stylistic markers for native language identification in addition to other features from the work of Koppel et al. (2005).

The rest of the paper is structured as follows. Section 2 reviews the literature studying native language identification and contrastive analysis. Section 3 decribes the methodology adopted in our study. The experimental results obtained are organised into two separate sections: Section 4 presents the results obtained merely from syntactic features; Section 5 discusses a replication of the work of Koppel et al. (2005) and details the results from adding in the syntactic features. Finally, Section 6 concludes.

2 Literature Review

2.1 Native language identification

Koppel et al. (2005) took a machine learning approach to the task, using as features function words, character n-grams, and part-of-speech (POS) bigrams; they gained a reasonably high classification accuracy of 80% across five different groups of non-native English authors (Bulgarian, Czech, French, Russian, and Spanish), selected from the first version of *International Corpus of Learner English* (ICLE). Koppel et al. (2005) also suggest that syntactic errors might be useful features, but these were not explored in their study. Tsur and Rappoport (2007) replicate this work of Koppel et al. (2005) and hypothesise that the choice of words in second language writing is highly influenced by the frequency of native language syllables – the *phonology* of the native language. Approximating this by character bigrams alone, they achieved a classification accuracy of 66%.

Native language is also among one of the characteristics investigated in the authorship profiling task of Estival et al. (2007). Unlike the approach of Koppel et al. (2005), linguistic errors in written texts are not of concern here; rather this study focuses merely on lexical and structural features. The approach deployed yields a relatively good classification accuracy of 84% when the native language alone is used as the profiling criterion. However, it should be noted that a smaller number of native language groups were examined in this study – namely, Arabic, English, and Spanish. The work was also carried out on data that is not publicly available.

Another relevant piece of research is that of van Halteren (van Halteren, 2008), which has demonstrated the possibility of identifying the source language of medium-length translated texts (between 400 and 2500 words). On the basis of frequency counts of word-based n-grams, surprisingly high classification accuracies from 87% to 97% are achievable in identifying the source language of *European Parliament* (EUROPARL) speeches. Six common European languages were examined – English, German, French, Dutch, Spanish, and Italian. In addition, van Halteren also uncovered salient markers for a particular source language. Many of these were tied to the content and the domain (e.g. the greeting to the European Parliament is always translated a particular way from German to English in comparison with other languages), suggesting a reason for the high classification accuracy rates.

2.2 Contrastive analysis

The goal of contrastive analysis is to predict linguistic difficulties experienced during the acquisition of a second language; as formulated by Lado (1957), it suggests that difficulties in acquiring a new (second) language are derived from the differences between the new language and the native (first) language of a language learner. In this regard, errors potentially made by learners of a second language are predicted from interference by the native language. Such a phenomenon is usually known as *negative transfer*. In error analysis (Corder, 1967), this was seen as only one kind of error, *interlanguage* or *interference errors*; other types were *intralingual* and *developmental* errors, which are not specific to the native language (Richards, 1971).

To return to contrastive analysis, numerous studies of different language pairs have already been carried out, in particular focusing on learners of English. Dušková (1969) investigated Czech learners of English in terms of various lexical and syntactical errors; Light and Warshawsky (1974) examined Russian learners of English (and French learners to some extent) on their improper usage of syntax as well as semantics; Guilford (1998) specifically explored the difficulties of French learners of English in various aspects, from lexical and syntactical to idiosyncratic; and Mohamed et al. (2004) targeted

grammatical errors of Chinese learners in English. Among these studies, commonly observed syntactic error types made by non-native English learners include subject-verb disagreement, noun-number disagreement, and misuse of determiners.

There are many other studies examining interlanguage errors, generally restricted in their scope of investigation to a specific grammatical aspect of English in which the native language of the learners might have an influence. To give some examples, Granger and Tyson (1996) examined the usage of connectors in English by a number of different native speakers – French, German, Dutch, and Chinese; Vassileva (1998) investigated the employment of first person singular and plural by another different set of native speakers – German, French, Russian, and Bulgarian; Slabakova (2000) explored the acquisition of telicity marking in English by Spanish and Bulgarian learners; Yang and Huang (2004) studied the impact of the absence of grammatical tense in Chinese on the acquisition of English tense-aspect system (i.e. telicity marking); Franck et al. (2002) and Vigliocco et al. (1996) specifically examined the usage of subject-verb agreement in English by French and Spanish, respectively.

3 Methodology

3.1 Data

The data used in our study is adopted from the *International Corpus of Learner English* (ICLE) compiled by Granger et al. (2009) for the precise purpose of studying the English writings of non-native English learners from diverse countries. All the contributors to the corpus are believed to possess similar English proficiency level (ranging from intermediate to advanced English learners) and are of about the same age (all in their twenties). This was also the data used by Koppel et al. (2005) and Tsur and Rappoport (2007), although where they used the first version of the corpus, we use the second.

The first version contains 11 sub-corpora of English essays contributed by students of different native languages – Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, and Swedish; the second has been extended to additional 5 other native languages – Chinese, Japanese, Norwegian, Turkish, and Tswana. In this work, we use the five languages of Koppel et al. (2005) – Bulgarian, Czech, French, Russian, Spanish – as well as Chinese and Japanese, based on the work discussed in Section 2.2. For each native language, we randomly select from among essays with length of 500-1000 words: 70 essays for training, 25 essays for testing, and another 15 essays for development. By contrast, Koppel et al. (2005) took all 258 texts from their version for each language and evaluated by ten-fold cross validation. We used fewer with a view to reserving more for future work. From our sample, the average text length broken down by native language is given in Table 1.

Bulgarian	668
Czech	747
French	639
Russian	692
Spanish	621
Chinese	570
Japanese	610

Table 1: Mean text length of native language (words)

3.2 Tools

As in the work discussed in Section 2.1, we use a machine learner. Since its performance in classification problems and its ability in handling high dimensional feature spaces have been well attested (Joachims, 1998), the support vector machine (SVM) is chosen as the classifier. We adopt the online SVM tool, *LIBSVM*[1] (Version 2.89) by Chang and Lin (2001). All the classifications are first conducted under the default settings, where the radial basic function (RBF) kernel is used as it is appropriate for learning a non-linear relationship between multiple features. The kernel is tuned to find the best pair of parameters (C, γ) for data training.

In addition to the machine learning tool, we require a grammar checker that help in detecting the syntactic errors. *Queequeg*,[2] a very small English grammar checker, detects the three error types that are of concern in our study, namely subject-verb disagreement, noun-number disagreement, and misuse of determiners (mostly articles).

4 Syntactic Features

Given that the main focus of this paper is to uncover whether syntactic features are useful in determining

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm

[2] http://queequeg.sourceforge.net/index-e.html

the native language of the authors, syntactic features are first examined separately. Statistical analysis is performed to gain an overview of the distribution of the syntactic errors detected from seven groups of non-native English users. A classification with SVM is then conducted to investigate the degree to which syntactic errors are able to classify the authors according to their native language.

4.1 Features

For the present study, only the three major syntactic error types named above are explored and are used as the syntactic features for classification learning.

Subject-verb disagreement: refers to a situation in which the subject of a sentence disagrees with the verb of the sentence in terms of number or person. An excerpt adopted from the training data that demonstrates such an error: **If the situation become worse . . . /If the situation becomes worse . . . .*

Noun-number disagreement: refers to a situation in which a noun is in disagreement with its determiner in terms of number. An excerpt adopted from the training data that demonstrates such an error: **They provide many negative image . . . /They provide many negative images . . . .*

Misuse of determiners: refers to situations in which the determiners (such as articles, demonstratives, as well as possessive pronouns) are improperly used with the nouns they modify. These situations include missing a determiner when required as well as having an extra determiner when not needed. An excerpt adopted from the training data that demonstrates such an error: **Cyber cafes should not be located outside airport. /Cyber cafes should not be located outside an airport*.[3]

Table 2 provides an overview of which of these grammatical phenomena are present in each native language. All three exist in English; a '-' indicates that generally speaking it does not exists or exists to a much lesser extent in a particular native language (e.g. with Slavic languages and determiners). A '+' indicates that the phenomenon exists, but not that it coincides precisely with the English one. For example, Spanish and French have much more extensive use of determiners than in English; the presence or absence of determiners in Bulgarian has no effect on aspectual interpretation, unlike in English; and as for Chinese and Japanese, the usage of determiners is far less frequent than that of the other languages and generally more deictic in nature. Conjugations (and consequently subject-verb agreement), on the other hand, are more extensive in the European languages than in English.

[3]Such an error may also be recognised as noun-number disagreement in which the grammatical form is . . . *outside airports*; but *Queequeg* identifies this as misuse of determiners.

Language	Subject-verb agreement	Noun-number agreement	Use of determiners
Bulgarian	+	+	+
Czech	+	+	-
French	+	+	+
Russian	+	+	-
Spanish	+	+	+
Chinese	-	-	+
Japanese	-	-	+

Table 2: Presence or absence of grammatical features

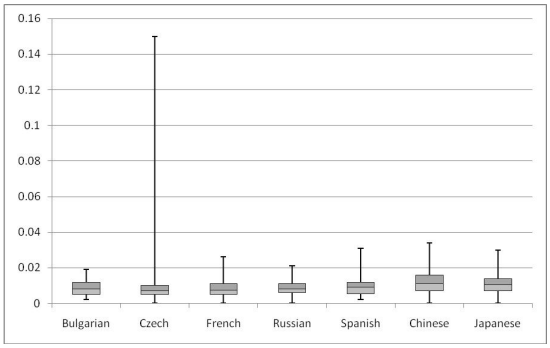

Figure 1: Boxplot: subject-verb disagreement errors

4.2 Data analysis

Boxplots: Figures 1 to 3 depict the distribution of each error type as observed in the training data – 490 essays written by 7 distinct groups of non-native English users. The frequencies of each error type presented in these figures are normalised by the corresponding text length (i.e. the total number of

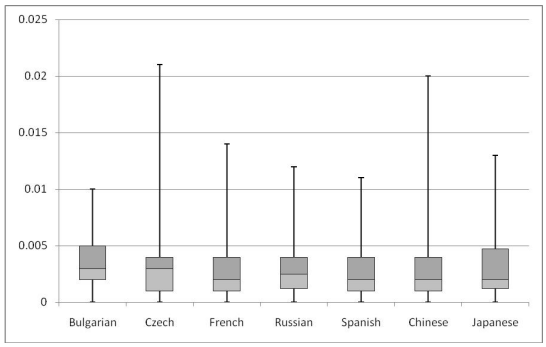

Figure 2: Boxplot: noun-number disagreement errors

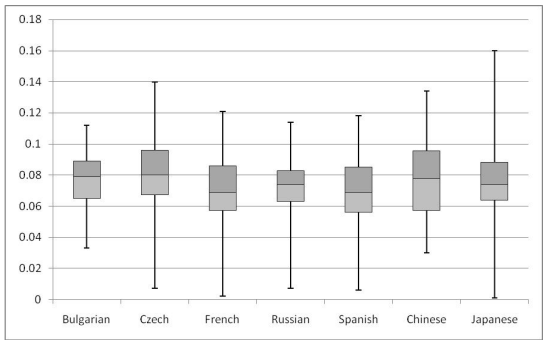

Figure 3: Boxplot: determiner misuse errors

Frequency type	Subject-verb disagreement	Noun-number disagreement	Misuse of determiners
Absolute	0.038	0.114	5.306E-10
Relative	0.178	0.906	0.006

Table 3: P-value of ANOVA test per error type

words). The boxplots present the median, quartiles and range, to give an initial idea of the distribution of each error type.

These boxplots do show some variability among non-native English users with different native languages with respect to their syntactic errors. This is most obvious in Figure 3, with the distribution of errors concerning misuse of determiners. This could possibly be explained by the interference of native language as indicated in the contrastive analysis. Czech and Chinese seem to have more difficulties when dealing with determiners as compared to French and Spanish, since determiners (especially articles) are absent from the language system of Czech and are less frequently used in Chinese, while the usage of determiners in French and Spanish is somewhat different from (and generally more extensive than) in English.

ANOVA tests: The boxplots do not suggest an extremely non-Gaussian distribution, so we use ANOVA tests to determine whether the distributions do in fact differ. A single-factor ANOVA, with the language type being the factor, was carried out for each syntactic error type, for both absolute frequency and relative frequency (normalised by text length). The results are presented in Table 3. Tables 4 to 6 present some descriptive statistics for each of the error types in terms of mean, standard deviation, median, first quartile, and third quartile.

The most interesting result is for the case of determiner misuse. This is highly statistically significant

Language	Mean	Std. Dev.	Median	Q1	Q3
Bulgarian	5.829	3.074	6	4	7
	0.0088	0.0042	0.008	0.005	0.012
Czech	5.414	3.268	5	3	7
	0.0106	0.0213	0.007	0.005	0.01
French	5.243	3.272	4	3	6
	0.0083	0.0048	0.0075	0.005	0.011
Russian	6.086	3.247	6	3	8
	0.0088	0.0045	0.008	0.006	0.011
Spanish	5.786	3.438	5	3	8
	0.0093	0.0051	0.009	0.0053	0.012
Chinese	6.757	3.617	6	4	9
	0.0118	0.0063	0.011	0.007	0.016
Japanese	6.857	4.175	6	4	8
	0.0112	0.0063	0.0105	0.007	0.014

Table 4: Descriptive statistics of subject-verb disagreement errors (first row – absolute frequency; second row – relative frequency)

Language	Mean	Std. Dev.	Median	Q1	Q3
Bulgarian	2.086	1.576	2	1	3
	0.0033	0.0025	0.003	0.002	0.005
Czech	2.457	2.250	2	1	4
	0.0033	0.0033	0.003	0.001	0.004
French	1.814	1.6	1	1	3
	0.003	0.0028	0.002	0.001	0.004
Russian	2.157	1.968	2	1	3
	0.003	0.0024	0.0025	0.0013	0.004
Spanish	1.7	1.376	1.5	1	2
	0.0027	0.0023	0.002	0.001	0.004
Chinese	1.671	1.791	1	1	2
	0.003	0.0032	0.002	0.001	0.004
Japanese	1.971	1.810	1	1	3
	0.0033	0.0029	0.002	0.0013	0.0048

Table 5: As Table 4, for noun-number disagreement

for both absolute and relative frequencies (with the p-values of 5.306E-10 and 0.006 respectively). This seems to be in line with our expectation and the explanation above.

As for subject-verb disagreement, significant differences are only observed in absolute frequency (with a p-value of 0.038). The inconsistency in results could be attributed to the differences in text length. We therefore additionally carried out another single-factor ANOVA test on the text length from our sample (mean values are given in Table 1), which shows that the text lengths are indeed different. The lack of a positive result is a little surprising, as Chinese and Japanese do not have subject-verb

Language	Mean	Std. Dev.	Median	Q1	Q3
Bulgarian	51.471	16.258	47.5	40.25	63.75
	0.0771	0.0169	0.079	0.065	0.089
Czech	61.529	23.766	59.5	44	73
	0.082	0.0253	0.08	0.0673	0.096
French	44.286	14.056	45	34	52
	0.0689	0.0216	0.069	0.0573	0.086
Russian	49.343	15.480	48.5	40.25	59
	0.072	0.0182	0.074	0.063	0.083
Spanish	43.9	15.402	43	31.75	53.75
	0.0706	0.0214	0.069	0.056	0.085
Chinese	44.686	15.373	45	33	54.75
	0.0782	0.0252	0.078	0.0573	0.0958
Japanese	46.243	16.616	43.5	36.25	55.75
	0.0768	0.0271	0.074	0.064	0.0883

Table 6: As Table 4, for determiner misuse

agreement, while the other languages do. However, we note that the absolute numbers here are quite low, unlike for the case of determiner misuse.

Noun-number disagreement, however, does not demonstrate significant differences amongst the seven groups of non-native English users (neither for the absolute frequency nor for the relative frequency), even though again the native languages differ in whether this phenomenon exists. Again, the absolute numbers are small.

Perhaps noun-number disagreement is just not an interference error. Instead, it may be regarded as a developmental error according to the notion of error analysis (Corder, 1967). Developmental errors are largely due to the complexity of the (second) language's grammatical system itself. They will gradually diminish as learners become more competent.

We also note at this point some limitations of the grammar checker *Queequeg* itself. In particular, the grammar checker suffers from false positives, in many cases because it fails to distinguish between count nouns and mass nouns. As such, the checker tends to generate more false positives when determining if the determiners are in disagreement with the nouns they modify. An example of such false positive generated by the checker is as follows: *It could help us to save some money* . . . , where *some money* is detected as ungrammatical. A manual evaluation of a sample of the training data reveals a relatively high false positive rate of 48.2% in determiner misuse errors. (The grammar checker also records a false negative rate of 11.1%.) However, there is no evidence to suggest any bias in the errors with respect to native language, so it just seems to act as random noise.

4.3 Learning from syntactic errors

Using the machine learner noted in Section 3.2, the result of classification based on merely syntactic features is shown in Table 7 below. The majority class baseline is 14.29%, given that there are 7 native languages with an equal quantity of test data. Since only three syntactic error types being examined, it is not unreasonable to expect that the accuracy would not improve to too great an extent. Nevertheless, the classification accuracies are somewhat higher than the baseline, approximately 5% (prior tuning) and 10% (after tuning) better when the relative frequency of the features is being examined. The improvement in classification accuracy after tuning is significant at the 95% confidence level, based on a z-test of two proportions.

Baseline	Presence/ absence	Relative frequency (before tuning)	Relative frequency (after tuning)
14.29% (25/175)	15.43% (27/175)	19.43% (34/175)	24.57% (43/175)

Table 7: Classification accuracy for error features

5 Learning from All Features

The second focus of our study is to investigate the effects of combining syntactic features with lexical features in determining the native language of the authors. To do this, we broadly replicate the work of Koppel et al. (2005) which used a machine learning approach with features commonly used in authorship analysis – function words, character n-grams, and POS n-grams. Koppel et al. (2005) also used spelling errors as features, although we do not do that here. Spelling errors would undoubtedly improve the overall classification performance to some extent but due to time constraints, we keep it for future work.

5.1 Features

Function words: Koppel et al. (2005) did not specify which set of function words was used, although they noted that there were 400 words in the set. Consequently, we explored three sets of function words. Firstly, a short list of 70 function words was examined; these function words were used by Mosteller and Wallace (1964) in their seminal work where they successfully attributed the twelve disputed Federalist papers. Secondly, a long list of 363 function words was adopted from Miller et al. (1958) from where the 70 function words used by Mosteller and Wallace (1964) were originally extracted. Considering that Koppel et al.(2005) made use of 400 function words, we then searched for some stop words commonly used in information retrieval to make up a list of close to 400 words – where our third list consists of 398 function words with stop words[4].

Character n-grams: As Koppel et al. (2005) did not indicate which sort of character n-grams

[4]Stop words were retrieved from Onix Text Retrieval Toolkit. `http://www.lextek.com/manuals/onix/stopwords1.html`

was used, we examined three different types: unigram, bi-gram, and tri-gram. The 200 most frequently used character bi-grams and tri-grams were extracted from our training data. As for unigrams, only the 100 most frequently used ones were extracted since there were fewer than 200 unique unigrams. Space and punctuation were considered as tokens when forming n-grams.

POS n-grams: In terms of POS n-grams, Koppel et al. (2005) tested on 250 rare bi-grams extracted from the Brown corpus. In our study, in addition to 250 rare bi-grams from the Brown corpus, we also examined the 200 most frequently used POS bi-grams and tri-grams extracted from our training data. We used the Brill tagger provided by NLTK for our POS tagging (Bird et al., 2009). Having trained on the Brown corpus, the Brill tagger performs at approximately 93% accuracy.

For each of the lexical features, four sets of classification were performed. The data was examined without normalising, with normalising to lowercase, according to their presence, as well as their relative frequency (per text length). (Note that since both the classification results with and without normalising to lowercase are similar, only the results without normalising will be presented.)

5.2 Results

Individual features: The classification results (before tuning) for each lexical feature – function words, character n-grams, and POS n-grams – are presented in Table 8, 9, and 10, respectively. Each table contains results with and without integrating with syntactic features (i.e. the three syntactic error types as identified in Section 4). It is obvious that function words and POS n-grams perform with higher accuracies when their presence is used as the feature value for classification; whereas character n-grams perform better when their relative frequency is considered. Also note that the best performance of character n-grams (i.e. bi-grams) before tuning is far below 60%, as compared with the other two lexical features. It, however, achieves as high as 69.14% after tuning where both function words and POS bi-grams are at 64.57% and 66.29%, respectively.

The classification results for the 250 rare bi-grams from the Brown corpus are not presented here since the results are all at around the baseline (14.29%).

Function words	Presence/ absence (- errors)	Presence/ absence (+ errors)	Relative frequency (- errors)	Relative frequency (+ errors)
70 words	50.86% (89/175)	50.86% (89/175)	40.57% (71/175)	42.86% (75/175)
363 words	60.57% (106/175)	61.14% (107/175)	41.71% (73/175)	43.43% (76/175)
398 words	65.14% (114/175)	65.14% (114/175)	41.71% (73/175)	43.43% (76/175)

Table 8: Classification accuracy for function words

Character n-grams	Presence/ absence (- errors)	Presence/ absence (+ errors)	Relative frequency (- errors)	Relative frequency (+ errors)
Character unigram	56.57% (99/175)	56.57% (99/175)	50.29% (88/175)	42.29% (74/175)
Character bi-gram	22.86% (40/175)	22.86% (40/175)	50.29% (88/175)	41.71% (73/175)
Character tri-gram	28.57% (50/175)	28.57% (50/175)	43.43% (76/175)	30.29% (53/175)

Table 9: Classification accuracy for character n-grams

Combined features: Table 11 presents both before and after tuning classification results of all combinations of lexical features (with and without syntactic errors). Each lexical feature was chosen for combination based on their best individual result. The combination of all three lexical features results in better classification accuracy than combinations of two features, noting however that character n-grams make no difference. In summary, our best accuracy thus far is at 73.71%. As illustrated in the confusion matrix (Table 12), misclassifications occur largely in Spanish and the Slavic languages.

5.3 Discussion

Comparisons with Koppel et al. (2005): Based on the results presented in Table 8 and 9, our classification results prior to tuning for both function words and character n-grams (without considering the syntactic features) appear to be lower than the results obtained by Koppel et al. (2005) (as presented in Table 13). However, character n-grams performs on par with Koppel et al. after tuning. The difference in classification accuracy (function words in particular) can be explained by the corpus size. In our study, we only adopted 110 essays for each native language. Koppel et al. made use of 258 essays for each native language. A simple analysis (extrapo-

POS n-grams	Presence/ absence (- errors)	Presence/ absence (+ errors)	Relative frequency (- errors)	Relative frequency (+ errors)
POS bi-gram	62.86% (110/175)	63.43% (111/175)	58.29% (102/175)	48.0% (84/175)
POS tri-gram	57.71% (101/175)	57.14% (100/175)	48.0% (84/175)	37.14% (65/175)

Table 10: Classification accuracy for POS n-grams

Combinations of features	prior tuning (- errors)	prior tuning (+ errors)	after tuning (- errors)	after tuning (+ errors)
Function words + character n-grams	58.29% (102/175)	58.29% (102/175)	64.57% (113/175)	64.57% (113/175)
Function words + POS n-grams	73.71% (129/175)	73.71% (129/175)	73.71% (129/175)	73.71% (129/175)
Character n-grams + POS n-grams	63.43% (111/175)	63.43% (111/175)	66.29% (116/175)	66.29% (116/175)
Function words + char n-grams + POS n-grams	72.57% (127/175)	72.57% (127/175)	73.71% (129/175)	73.71% (129/175)

Table 11: Classification accuracy for all combinations of lexical features

	BL	CZ	FR	RU	SP	CN	JP
BL	[16]	4	-	5	-	-	-
CZ	3	[18]	-	3	1	-	-
FR	1	-	[24]	-	-	-	-
RU	3	4	3	[14]	-	-	1
SP	1	2	4	3	[14]	-	1
CN	1	1	1	-	-	[20]	2
JP	-	-	-	-	-	4	[21]

Table 12: Confusion matrix based on both lexical and syntactic features (BL:Bulgarian, CZ:Czech, FR:French, RU:Russian, SP:Spanish, CN:Chinese, JP:Japanese)

lating from a curve fitted by a linear regression of the results for variously sized subsets of our data) suggests that our results are consistent with Koppel et al.'s given the sample size. (Note that the results of POS n-grams could not be commented here since Koppel et al. had considered these features as errors and did not provide a separate classification result.)

Usefulness of syntactic features: For the best combinations of features, our classification results of integrating the syntactic features (i.e. syntactic error types) with the lexical features do not demonstrate any improvement in terms of classification accuracy. For the individual feature types with results in Table 8 to Table 10, the syntactic error types sometimes in fact decrease accuracies. This could be due to the small number of syntactic error types being considered at this stage. Such a small number of features (three in our case) would not be sufficient to add much to the approximately 760 features used in our replication of the Koppel et al.'s work. Furthermore, error detection may be flawed as the result of the limitations noted in the grammar checker.

Other issues of note: Character n-grams, as seen in our classification results (see Table 11) do not seem to be contributing to the overall classification.

Types of lexical feature	Koppel et al.	Our best result (prior tuning)	Our best result (after tuning)
Function words	~71.0%	~65.0%	~65.0%
Character n-grams	~68.0%	~56.0%	~69.0%

Table 13: Comparison of results with Koppel et al.

It is noticeable when character n-grams are combined with function words and when combined with POS n-grams separately. Both combinations do not exhibit any improvement in accuracy. In addition, with character n-grams adding to the other two lexical features, the overall classification accuracy does not seem to be improved either. Nevertheless, as mentioned in Section 5.2 (under individual features), character n-grams alone are able to achieve an accuracy close to 69%. It seems that character n-grams are somehow a useful marker as argued by Koppel et al. (2005) that such feature may reflect the orthographic conventions of individual native language. Furthermore, this is consistent with the hypothesis put forward by Tsur and Rappoport (2007) in their study. It was claimed that the choice of words in second language writing is highly influenced by the frequency of native language syllabus (i.e. the *phonology* of the native language) which can be captured by character n-grams. For example, confusion between phonemes /l/ and /r/ is commonly observed in Japanese learners of English.

6 Conclusion

We have found some modest support for the contention that contrastive analysis can help in detecting the native language of a text's author, through a statistical analysis of three syntactic error types and through machine learning using only features based on those error types. However, in combining these with features used in other machine learning approaches to this task, we did not find an improvement in classification accuracy.

An examination of the results suggests that using more error types, and a method for more accurately identifying them, might result in improvements. A still more useful approach might be to use an automatic means to detect different types of syntactic errors, such as the idea suggested by Gamon (2004) in which context-free grammar production rules can be explored to detect ungrammatical structures based on long-distance dependencies. Furthermore, error analysis may be worth exploring to uncover non-interference errors which could then be discarded as irrelevant to determining native language.

Acknowledgments

The authors would like to acknowledge the support of ARC Linkage grant LP0776267, and thank the reviewers for useful feedback.

References

Stephen Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Stephen P. Corder. 1967. The significance of learners' errors. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 5(4):161–170.

Libuše Dušková. 1969. On sources of error in foreign language learning. *International Review of Applied Linguistics (IRAL)*, 7(1):11–36.

Dominique Estival, Tanja Gaustad, Son-Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 263–272.

Ian Fette, Norman Sadeh, and Anthony Tomasic. 2007. Learning to detect phishing emails. In *Proceedings of the 16th International World Wide Web Conference*.

Julie Franck, Gabriella Vigliocco, and Janet Nicol. 2002. Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, 17(4):371–404.

Michael Gamon. 2004. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 611–617.

Sylviane Granger and Stephanie Tyson. 1996. Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, 15(1):17–27.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, , and Magali Paquot. 2009. *International Corpus of Learner English (Version 2)*. Presses Universitaires de Louvain, Louvain-la-Neuve.

Jonathon Guilford. 1998. English learner interlanguage: What's wrong with it? *Anglophonia French Journal of English Studies*, 4:73–100.

Thorsten Joachims. 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In *Machine Learning: ECML-98*, pages 137–142.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. In *Intelligence and Security Informatics*, volume 3495 of *Lecture Notes in Computer Science*, pages 209–217. Springer-Verlag.

Robert Lado. 1957. *Linguistics Across Cultures: Applied Linguistics for Language Teachers*. University of Michigan Press, Ann Arbor, MI, US.

Richard L. Light and Diane Warshawsky. 1974. Preliminary error analysis: Russians using English. Technical report, National Institute of Education, USA.

George A. Miller, E. B. Newman, and Elizabeth A. Friedman. 1958. Length frequency statistics for written English. *Information and Control*, 1(4):370–389.

Abdul R. Mohamed, Li-Lian Goh, and Eliza Wan-Rose. 2004. English errors and Chinese learners. *Sunway College Journal*, 1:83–97.

Frederick Mosteller and David L. Wallace. 1964. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA, US.

Jack C. Richards. 1971. A non-contrastive approach to error analysis. *ELT Journal*, 25(3):204–219.

Roumyana Slabakova. 2000. L1 transfer revisited: the L2 acquisition of telicity marking in English by Spanish and Bulgarian native speakers. *Linguistics*, 38(4):739–770.

Oren Tsur and Ari Rappoport. 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*, pages 9–16.

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

Irena Vassileva. 1998. Who am I/how are we in academic writing? A contrastive analysis of authorial presence in English, German, French, Russian and Bulgarian. *International Journal of Applied Linguistics*, 8(2):163–185.

Garbriella Vigliocco, Brian Butterworth, and Merrill F. Garrett. 1996. Subject-verb agreement in Spanish and English: Differences in the role of conceptual constraints. *Cognition*, 61(3):261–298.

Suying Yang and Yue-Yuan Huang. 2004. The impact of the absence of grammatical tense in L1 on the acquisition of the tense-aspect system in L2. *International Review of Applied Linguistics in Language Teaching (IRAL)*, 42(1):49–70.

Rong Zheng, Yi Qin, Zan Huang, and Hsinchun Chen. 2003. Authorship analysis in cybercrime investigation. In *Intelligence and Security Informatics*, volume 2665 of *Lecture Notes in Computer Science*, pages 59–73. Springer-Verlag.