# Exploiting Parse Structures for Native Language Identification

**Sze-Meng Jojo Wong**
Centre for Language Technology
Macquarie University
Sydney, Australia
sze.wong@mq.edu.au

**Mark Dras**
Centre for Language Technology
Macquarie University
Sydney, Australia
mark.dras@mq.edu.au

## Abstract

Attempts to profile authors according to their characteristics extracted from textual data, including native language, have drawn attention in recent years, via various machine learning approaches utilising mostly lexical features. Drawing on the idea of contrastive analysis, which postulates that syntactic errors in a text are to some extent influenced by the native language of an author, this paper explores the usefulness of syntactic features for native language identification. We take two types of parse substructure as features—horizontal slices of trees, and the more general feature schemas from discriminative parse reranking—and show that using this kind of syntactic feature results in an accuracy score in classification of seven native languages of around 80%, an error reduction of more than 30%.

## 1 Introduction

Inferring characteristics of authors from their textual data, often termed *authorship profiling*, has seen a number of computational approaches proposed in recent years. The problem is typically treated as a classification task, where an author is classified with respect to characteristics such as gender, age, native language, and so on. This profile information is often of interest to marketing organisations for product promotional reasons as well as governments or law enforcements for crime investigation purposes. The particular application that motivates the present study is detection of phishing (Myers, 2007), the attempt to defraud through texts that are designed to deceive Internet users into giving away confidential details. One class of countermeasures to phishing consists of technical methods such as email authentication; another looks at profiling of the text's author(s) (Fette et al., 2007; Zheng et al., 2003), to find any indications of the source of the text.

In this paper we investigate classification of a text with respect to an author's native language, where this is not the language that that text is written in (which is often the case in phishing); we refer to this as *native language identification*. Initial work by Koppel et al. (2005) was followed by Tsur and Rappoport (2007), Estival et al. (2007), van Halteren (2008), and Wong and Dras (2009). By and large, the problem was tackled using various supervised machine learning approaches, with mostly lexical features over characters, words, and parts of speech, as well as some document structure.

Syntactic features, in contrast, in particular those that capture grammatical errors, which might potentially be useful for this task, have received little attention. Koppel et al. (2005) did suggest using syntactic errors in their work but did not investigate them in any detail. Wong and Dras (2009) noted the relevance of the concept of *contrastive analysis* (Lado, 1957), which postulates that native language constructions lead to characteristic errors in a second language. In their experimental work, however, they used only three manual syntactic constructions drawn from the literature; an ANOVA analysis showed a detectable effect, but they did not improve classification accuracy over purely lexical features.

In this paper, we investigate syntactic features for native language identification that are more general



*Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610,
Edinburgh, Scotland, UK, July 27–31, 2011. 

than, and that do not require the manual construction of, the above approach. Taking the trees produced by statistical parsers, we use tree cross-sections as features in a machine learning approach to determine which ones characterise non-native speaker errors. Specifically, we look at two types of parse tree substructure to use as features: horizontal slices of the trees—that is, characterising parse trees as sets of context-free grammar production rules—and the features schemas used in discriminative parse reranking. The goal of the present study is therefore to investigate the influence to which syntactic features represented by parse structures would have on the classification task of identifying an author's native language relative to, and in combination with, lexical features.

The remainder of this paper is structured as follows. In Section 2, we discuss some related work on the two key topics of this paper: primarily on comparable work in native language identification, and then on how the notion of contrastive analysis can be applicable here. We then describe the models examined in Section 3, followed by experimental setup in Section 4. Section 5 presents results, and Section 6 discussion of those results.

## 2 Related Work

### 2.1 Native Language Identification

The earliest work on native language identification in this classification paradigm is that of Koppel et al. (2005), in which they deployed a machine learning approach to the task, using as features function words, character n-grams, and part-of-speech (PoS) bi-grams, as well as some spelling mistakes. With five different groups of English authors (of native languages Bulgarian, Czech, French, Russian, and Spanish) selected from the first version of *International Corpus of Learner English* (ICLE), they gained a relatively high classification accuracy of 80%. Koppel et al. (2005) also suggested that syntactic features (syntactic errors) might be useful features, but only investigated this idea at a shallow level by treating rare PoS bigrams as ungrammatical structures.

Tsur and Rappoport (2007) replicated the work of Koppel et al. (2005) to investigate the hypothesis 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 bi-grams alone, they managed to achieve a classification accuracy of 66%.

Native language is also amongst the characteristics investigated in the task of authorship profiling by Estival et al. (2007), as well as other demographic and personality characteristics. This study used a variety of lexical and document structure features. For the native language identification classification task, their model yielded a reasonably high accuracy of 84%, but this was over a set of only three languages (Arabic, English and Spanish) and against a most frequent baseline of 62.9%.

Another related work is that of van Halteren (2008), who used the Europarl corpus of parliamentary speeches. In Europarl, one original language is transcribed, and the others translated from it; the task was to identify the original language. On the basis of frequency counts of word-based n-grams, surprisingly high classification accuracies within the range of 87-97% were achieved across six languages (English, German, French, Dutch, Spanish, and Italian). This turns out, however, to be significantly influenced by the use of particular phrases used by speakers of different languages in the parliamentary context (e.g. the way Germans typically address the chamber).

To our knowledge, Wong and Dras (2009) is the only work that has investigated the usefulness of syntactic features for the task of native language identification. They first replicated the work of Koppel et al. (2005) with the three types of lexical feature, namely function words, character n-grams, and PoS bi-grams. They then examined the literature on contrastive analysis (see Section 2.2), from the field of second language acquisition, and selected three syntactic errors commonly observed in non-native English users—subject-verb disagreement, noun-number disagreement and misuse of determiners—that had been identified as being influenced by the native language. An ANOVA analysis showed that the native language identification constructions were identifiable; however, the overall classification was not improved over the lexical features by using just the three manually detected syntactic errors. The best overall accuracy re-

ported was 73.71%; this was on the second version of ICLE, across seven languages (those of Koppel et al. (2005), plus the two Asian languages Chinese and Japanese).

As a possible approach that would improve the classification accuracy over just the three manually detected syntactic errors, Wong and Dras (2009) suggested deploying (but did not carry out) an idea put forward by Gamon (2004) (citing Baayen et al. (1996)) for the related task of identifying the author of a text: to use CFG production rules to characterise syntactic structures used by authors.[1] We note that similar ideas have been used in the task of sentence grammaticality judgement, which utilise parser outputs (both trees and by-products) as classification features (Mutton et al., 2007; Sun et al., 2007; Foster et al., 2008; Wagner et al., 2009; Tetreault et al., 2010; Wong and Dras, 2010). We combine this idea with one we introduce in this paper, of using discriminative reranking features as a broader characterisation of the parse tree.

### 2.2 Contrastive analysis

*Contrastive analysis* (Lado, 1957) was an early attempt in the field of second language acquisition to explain the kinds and source of errors that non-native speakers make. It arose out of behaviourist psychology, and saw language learning as an issue of habit formation that could be inhibited by previous habits inculcated in learning the native language. The theory was also tied to structural linguistics: it compared the syntactic structures of the native and second languages to find differences that might cause learning difficulties. The Lado work postulated the Contrastive Analysis Hypothesis (CAH), claiming that "those elements which are similar to [the learner's] native language will be simple for him, and those elements that are different will be difficult"; the consequence is that there will be more errors made in those difficult elements.

While contrastive analysis was influential at first, it was increasingly noticed that many errors were common across all language learners regardless of native language, which could not be explained under contrastive analysis. Corder (1967) then described an alternative, *error analysis*, where contrastive analysis-style errors were seen as only one type of error, 'interlanguage' or 'interference' errors; other types were 'intralingual' and 'developmental' errors, which are not specific to the native language (Richards, 1971).

In an overview of contrastive analysis after the emergence of error analysis, Wardhaugh (1970) noted that there were two interpretations of the CAH, termed the strong and weak forms. Under the strong form, all errors were attributed to the native language, and clearly that was not tenable in light of error analysis evidence. In the weak form, these differences have an influence but are not the sole determinant of language learning difficulty. Wardhaugh noted claims at the time that the hypothesis was no longer useful in either the strong or the weak version: "Such a claim is perhaps unwarranted, but a period of quiescence is probable for CA itself". This appears to be the case, with the then-dominant error analysis giving way to newer, more specialised theories of second language acquisition, such as the *competition model* of MacWhinney and Bates (1989) or the *processability theory* of Pienemann (1998). Nevertheless, smaller studies specifically of interlanguage errors have continued to be carried out, generally restricted in their scope 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. There are also a few teaching resources

[1] It is not entirely clear how this might work for authorship identification: would the Brontë sisters, the corpus Gamon worked with, have used a significant number of different syntactic constructions from each other? In the context of native language identification, however, constrastive analysis postulates that this is exactly the case for the different classes.

for English language teachers that collate such phenomena, such as that of Swan and Smith (2001).

NLP techniques and a probabilistic view of native language identification now let us revisit and make use of the weak form of the CAH. Interlanguage errors, as represented by differences in parse trees, may be characteristic of the native language of a learner; we can use the occurrence of these to come up with a revised likelihood of the native language. In this paper, we use machine learning in a prediction task as our approach to this.

## 3 Models

This section describes the three basic models investigated: the lexical model, based on Koppel et al. (2005), as the baseline; and then the two models that exploit syntactic information. In Section 5 we look at the performance of each model independently and also in combination: to combine, we just concatenate feature vectors.

**Lexical** As Wong and Dras (2009), we replicate the features of Koppel et al. (2005) to produce our LEXICAL model. These are of three types: function words,[2] character n-grams, and PoS n-grams. We follow Wong and Dras (2009) in resolving some unclear issues from Koppel et al. (2005). Specifically, we use the same list of function words, left unspecified in Koppel et al. (2005), that were empirically determined by Wong and Dras (2009) to be the best of three candidates; we used character bi-grams, as the best performing n-grams, although this also had been left unspecified by Koppel et al. (2005); and we used the most frequently occurring PoS bi-grams and tri-grams, obtained by using the Brill tagger provided in NLTK (Bird et al., 2009) being trained on the Brown corpus. In total, there are 798 features of this class with 398 function words, 200 most frequently occurring character bi-grams, and 200 most frequently occurring PoS bi-grams. Both function words and PoS bi-grams have feature values of binary type; while for character bi-grams, the feature value is the relative frequency. (These types of feature value are the best performing one for each lexical feature.)

We omitted the 250 rare bi-grams used by Koppel et al. (2005), as an ablative analysis showed that they contributed nothing to classification accuracy.

**Production Rules** Under this model (PRODRULE), we take as features horizontal slices of parse trees, in effect treating them as sets of CFG production rules. Feature values are binary. We look at all possible rules as features, but also present results for subsets of features chosen using feature selection. For each language in our dataset, we identify the $n$ rules most characteristic of the language using Information Gain (IG). For $m$ classes, we use the formulation of Yang and Pedersen (1997):

$$\begin{aligned} IG(r) = & -\sum_{i=1}^{m} \Pr(c_i) \log \Pr(c_i) \\ & + \Pr(r) \sum_{i=1}^{m} \Pr(c_i|r) \log \Pr(c_i|r) \\ & + \Pr(\bar{r}) \sum_{i=1}^{m} \Pr(c_i|\bar{r}) \log \Pr(c_i|\bar{r}) \end{aligned} \quad (1)$$

We also investigated simple frequencies, frequency ratios, and pointwise mutual information; as in much other work, IG performed best, so we do not present results for the others. Bi-normal separation (Forman, 2003), often competitive with IG, is only suitable for binary classification.

It is worth noting that the production rules being used here are all non-lexicalised ones, except those lexicalised with function words and punctuation, to avoid topic-related clues.

**Reranking Features** As opposed to the horizontal parse production rules, features used for discriminative reranking are cross-sections of parse trees that might capture other aspects of ungrammatical structures. For these we use the 13 feature schemas described in Charniak and Johnson (2005), which were inspired by earlier work in discriminative estimation techniques, such as Johnson et al. (1999) and Collins (2000). Examples of these feature schemas include tuples covering head-to-head dependencies, preterminals together with their closest maximal projection ancestors, and subtrees rooted in the least common ancestor.

These feature schemas are not the only possible ones—they were empirically selected for the specific purpose of augmenting the Charniak parser. However, much subsequent work has tended to use

[2] As with most work in authorship profiling, only function words are used, so that the result is not tied to a particular domain, and no clues are obtained from different topics that different authors might write about.

these same features, albeit sometimes with extensions for specific purposes (e.g. Johnson and Ural (2010) for the Berkeley parser (Petrov et al., 2006), Ng et al. (2010) for the C&C parser (Clark and Curran, 2007)). We also use this standard set, specifically the set of instantiated feature schemas from the parser from Charniak and Johnson (2005) as trained on the Wall Street Journal (WSJ), which gives 1,333,837 potential features.

## 4 Experimental Setup

### 4.1 Data

We use 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 claimed to possess similar English proficiency levels (ranging from intermediate to advanced learners) and are in the same age group (all in their twenties at the time of corpus collection.) 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 version 2.

Briefly, the first version contains 11 sub-corpora of English essays contributed by second-year and third-year university students of different native language backgrounds (mostly European and Slavic languages) — Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, and Swedish; the second version has been extended to additional 5 other native languages (including Asian languages) — Chinese, Japanese, Norwegian, Turkish, and Tswana.

As per Wong and Dras (2009), we examine seven languages, namely Bulgarian, Czech, French, Russian, Spanish, Chinese, and Japanese. For each native language, we randomly select from amongst essays with length of 500-1000 words. For the purpose of the present study, we have 95 essays per native language. For the same reason as highlighted by Wong and Dras (2009), we intentionally use fewer essays as compared to Koppel et al. (2005)[3] with a view to reserving more data for future work. We divide these into training sets of 70 essays per language, with a held-out test set of 25 essays per language. There are 17,718 training sentences and 6,791 testing sentences.

### 4.2 Parsers

We use two parsers: the Stanford parser (Klein and Manning, 2003) and the Charniak and Johnson (henceforth C&J) parser (Charniak and Johnson, 2005). Both are widely used, and produce relatively accurate parses: the Stanford parser gets a labelled f-score of 85.61 on the WSJ, and the C&J 91.09.

With the Stanford parser, there are 26,284 unique parse production rules extractable from our ICLE training set of 490 texts, while the C&J parser produces 27,705. For reranking, we use only the C&J parser—since the parser stores these features during parsing, we can use them directly as classification features. On the ICLE training data, there are 6,230 features with frequency $>10$, and 19,659 with frequency $>5$.

### 4.3 Classifiers

For our experiments we used a maximum entropy (MaxEnt) machine learner, MegaM[4] (fifth release) by Hal Daumé III. (We also used an SVM for comparison, but the results were uniformly worse, and degraded more quickly as number of features increased, so we only report the MaxEnt results here). The classifier is tuned to obtain an optimal classification model.

### 4.4 Evaluation Methodology

Given our relatively small amount of data, we use $k$-fold cross-validation, choosing $k = 5$. While testing for statistical significance of classification results is often not carried out in NLP, we do so here because the quantity of data could raise questions about the certainty of any effect. In an encyclopedic survey of cross-validation in machine learning contexts, Refaeilzadeh et al. (2009) note that there is as yet no universal standard for testing of statistical significance; and that while more sophisticated techniques have been proposed, none is more widely accepted than a paired t-test over folds. We therefore use this paired t-test over folds, as formulated of Alpaydin

[3]Koppel et al. (2005) took all 258 texts per language from ICLE Version 1 and evaluated using 10-fold cross valiadation.

[4]MegaM is available on `http://www.cs.utah.edu/~hal/megam/`.

(2004). Under this cross-validation, 5 separate training feature sets are constructed, excluding the test fold; 3 folds are used for training, 1 fold for tuning and 1 fold for testing.

We also use a held-out test set for comparison, as it is well-known that cross-validation can overestimate prediction error (Hastie et al., 2009). We do not carry out significance testing here—with this held-out test set size ($n = 125$), two models would have to differ by a great deal to be significant. We only use it as a check on the effect of applying to completely new data.

## 5 Results

Table 1 presents the results for the three models individually under cross-validation. The first point to note is that PROD-RULE, under both parsers, is a substantial improvement over LEXICAL when (non-lexicalised) parse rules together with rules lexicalised with function words are used (rows marked with * in Table 1), with the largest difference as much as 77.75% for PROD-RULE[both]* ($n = all$) versus 64.29% for LEXICAL; these differences with respect to LEXICAL are statistically significant. (To give an idea, the paired t-test standard error for this largest difference is 2.52%.) In terms of error reduction, this is over 30%.

There appears to be no difference according to the parser used, regardless of their differing accuracy on the WSJ. Using the selection metric for PROD-RULE without rules lexicalised with function words produces results all around those for LEXICAL; using fewer reranking features is worse as the quality of RERANKING declines as feature cut-offs are raised.

Another, somewhat surprising point is that the RERANKING results are also generally around those of LEXICAL even though like PROD-RULE they are also using cross-sections of the parse tree. We consider there might be two possible reasons for this. The first is that the feature schemas used were originally chosen for the specific purpose of augmenting the performance of the Charniak parser; perhaps others might be more appropriate here. The second is that we selected only those instantiated feature schemas that occurred in the WSJ, and then applied them to ICLE. As the WSJ is filled with predominantly grammatical text, perhaps those that were not seen on the WSJ are precisely those that might indicate ungrammaticality. In contrast, the production rules of PROD-RULE were selected only from the ICLE training data.

| Features | MaxEnt |
|---|---|
| LEXICAL ($n = 798$) | 64.29 |
| PROD-RULE[Stanford] ($n = 1000$) | 65.72 |
| PROD-RULE[Stanford]* ($n = 1000$) | 74.08 |
| PROD-RULE[Stanford]* ($n = all$) | 74.49 |
| PROD-RULE[C&J] ($n = 1000$) | 62.25 |
| PROD-RULE[C&J]* ($n = 1000$) | 71.84 |
| PROD-RULE[C&J]* ($n = all$) | 71.63 |
| PROD-RULE[both] ($n = 2000$) | 67.96 |
| PROD-RULE[both]* ($n = 2000$) | 74.69 |
| PROD-RULE[both]* ($n = all$) | 77.75 |
| RERANKING (all features) | 67.96 |
| RERANKING (>5 counts) | 66.33 |
| RERANKING (>10 counts) | 64.90 |

Table 1: Classification results based on 5-fold cross validation with parse rules as syntactic features (accuracy %)

| Features | MaxEnt |
|---|---|
| Lexical features ($n = 798$) | 75.43 |
| PROD-RULE[Stanford] ($n = 1000$) | 74.29 |
| PROD-RULE[Stanford]* ($n = 1000$) | 79.43 |
| PROD-RULE[Stanford]* ($n = all$) | 78.86 |
| PROD-RULE[C&J] ($n = 1000$) | 73.71 |
| PROD-RULE[C&J] ($n = 1000$)* | 79.43 |
| PROD-RULE[C&J] ($n = all$)* | 80.00 |
| PROD-RULE[both] ($n = 2000$) | 77.71 |
| PROD-RULE[both] ($n = 2000$)* | 78.85 |
| PROD-RULE[both] ($n = all$)* | 80.00 |
| RERANKING (all features) | 77.14 |
| RERANKING (>5 counts) | 76.57 |
| RERANKING (>10 counts) | 75.43 |

Table 2: Classification results based on hold-out validation with parse rules as syntactic features (accuracy %)

Table 2 presents the results for the individual models on the held-out test set. The results are generally higher than for cross-validation—this is not surprising, as the texts are of the same type, but all the training data is used (rather than the $1 - 1/k$ proportion for cross-validation). Overall, the pattern is still the same, with PROD-RULE best, then RERANKING and LEXICAL broadly similar; as expected, no differences are significant with this smaller dataset. The gap has narrowed, but without significance test-

| Features | MaxEnt |
|---|---|
| LEXICAL ($n = 798$) | 64.29 |
| LEXICAL + PROD-RULE[both] ($n = 2000$) | 63.06 |
| LEXICAL + PROD-RULE[both]* ($n = 2000$) | 72.45 |
| LEXICAL + PROD-RULE[both]* ($n = all$) | 70.82 |
| LEXICAL + RERANKING ($n = all$) | 68.17 |

Table 3: Classification results based on 5-fold cross validation for combined models (accuracy %)

| Features | MaxEnt |
|---|---|
| LEXICAL ($n = 798$) | 75.43 |
| LEXICAL + PROD-RULE[both] ($n = 2000$) | 80.57 |
| LEXICAL + PROD-RULE[both]* ($n = 2000$) | 81.14 |
| LEXICAL + PROD-RULE[both]* ($n = all$) | 81.71 |
| LEXICAL + RERANKING ($n = all$) | 76.00 |

Table 4: Classification results based on hold-out validation for combined models (accuracy %)

ing it is difficult to say whether this is a genuine phenomenon. The accuracy rate for LEXICAL here is in line with Wong and Dras (2009); and given the smaller dataset and larger set of languages, also broadly in line with Koppel et al. (2005).

Tables 3 and 4 present results for model combinations. It can be seen that the model combinations do not produce results better than PROD-RULE alone. Combining all features (results not presented here) seems to degrade the overall performance even of the MegaM: perhaps we need to derive feature vectors more compactly than by feature concatenation.

## 6 Discussion

As illustrated in the confusion matrices (Table 5 for the PROD-RULE model, and Table 6 for the LEXICAL model), misclassifications occur largely in Spanish and Slavic languages, Bulgarian and Russian in particular. Unsurprisingly, Chinese is almost completely identified since it comes from a entirely different language family, Sino-Tibetan, as compared to the rest of the languages which are from the branches of the Indo-European family (with Japanese as the exception). Japanese and French also appear to be easily distinguished, which could probably be attributed to their word order or sentence structure which are, to some extent, quite different from English. Japanese is a 'subject-object-verb' language; and French, although having the same word order as English, heads of phrases in

| | BL | CZ | FR | RU | SP | CN | JP |
|---|---|---|---|---|---|---|---|
| **BL** | [14] | 6 | 2 | 3 | - | - | - |
| **CZ** | 1 | [20] | - | 3 | 1 | - | - |
| **FR** | - | - | [25] | - | - | - | - |
| **RU** | 1 | 4 | 3 | [17] | - | - | - |
| **SP** | 2 | 1 | 3 | 1 | [18] | - | - |
| **CN** | - | - | - | - | - | [24] | 1 |
| **JP** | - | - | - | - | 1 | 2 | [22] |

Table 5: Confusion matrix based on all non-lexicalised parse rules from both parsers on the held-out set (BL:Bulgarian, CZ:Czech, FR:French, RU:Russian, SP:Spanish, CN:Chinese, JP:Japanese)

| | BL | CZ | FR | RU | SP | CN | JP |
|---|---|---|---|---|---|---|---|
| **BL** | [14] | 3 | 2 | 4 | 2 | - | - |
| **CZ** | 6 | [16] | - | 2 | 1 | - | - |
| **FR** | 1 | - | [24] | - | - | - | - |
| **RU** | 3 | 2 | 3 | [16] | 1 | - | - |
| **SP** | 1 | 2 | 3 | 1 | [17] | - | 1 |
| **CN** | - | - | - | - | - | [24] | 1 |
| **JP** | - | - | - | - | 1 | 3 | [21] |

Table 6: Confusion matrix based on lexical features on the held-out set (BL:Bulgarian, CZ:Czech, FR:French, RU:Russian, SP:Spanish, CN:Chinese, JP:Japanese)

French typically come before modifiers as opposed to English. Overall, the PROD-RULE model results in fewer misclassifications compared to the LEXICAL model; there are mostly only incremental improvements for each language, with perhaps the exception of the reduction in confusion in the Slavic languages.

We looked at some of the data, to see what kind of syntactic substructure is useful in classifying native language. Although using feature selection with only 1000 features did not improve performance, the information gain ranking does identify particular constructions as characteristic of one of the languages, and so are useful for inspection.

A phenomenon that the literature has noted as occurring with Chinese speakers is that of the missing determiner.[5] This corresponds to a higher frequency of NP rules without determiners. These rules may be valid in other contexts, but are also used to describe ungrammatical constituents. One example is

[5]This does happen with native speakers of some other languages, such as Slavic ones, but not generally (from our knowledge of the literature) with native speakers of others, such as Romance ones.

| Rules | Counts | | | | | | |
|---|---|---|---|---|---|---|---|
| | **BL** | **CZ** | **FR** | **RU** | **SP** | **CN** | **JP** |
| NNP → <R> | 0 | 0 | 3 | 0 | 0 | 67 | 0 |
| : → - | 55 | 51 | 23 | 39 | 10 | 9 | 4 |
| PRN → -LRB- X -RRB- | 0 | 1 | 7 | 2 | 0 | 42 | 0 |
| SYM → * | 0 | 1 | 7 | 3 | 1 | 42 | 0 |
| : → : | 30 | 39 | 58 | 46 | 47 | 11 | 6 |
| X → SYM | 0 | 2 | 7 | 4 | 4 | 42 | 6 |
| NP → NNP NNP NNS | 0 | 3 | 1 | 0 | 0 | 31 | 0 |
| S → S : S . | 36 | 34 | 53 | 39 | 41 | 5 | 9 |
| PP → VBG PP | 9 | 15 | 16 | 12 | 13 | 54 | 13 |
| : → ... | 16 | 13 | 39 | 11 | 24 | 1 | 3 |

Table 7: Top 10 rules for the Stanford parser according to Information Gain on the held-out set

```
(ROOT
  (S
    (NP
      (NP (DT The) (NN development))
      (PP (IN of)
        (NP (NN country) (NN park))))
    (VP (MD can)
      (ADVP (RB directly))
      (VP (VB elp)
        (S
          (VP (TO to)
            (VP (VB alleviate)
              (NP (NNS overcrowdedness)
                (CC and)
                (NN overpopulation))
              (PP (IN in)
                (NP (JJ urban)
                    (NN area))))))))
    (. .)))
```

Figure 1: Parse from Chinese-speaking authors, illustrating missing determiner

```
(ROOT
  (S
    (PP (VBG According)
      (PP (TO to)
        (NP (NNP <R>))))
    (, ,)
    (NP
      (NP (NN burning))
      (PP (IN of)
        (NP (JJ plastic)
            (NN waste))))
    (VP (VBZ generates)
      (NP (JJ toxic)
          (NNS by-products)))
    (. .)))
```

Figure 2: Parse from Chinese-speaking authors, illustrating *according to*

NP → NN NN. In Figure 1 we give the parse (from the Stanford parser) of the sentence *The development of **country park** can directly elp to alleviate overcrowdedness and overpopulation in urban area*. The phrase *country park* should either have a determiner or be plural (in which case the appropriate rule would be NP → NN NNS). There is a similar phenomenon with *in urban area*, although this is an instance of the rule NP → JJ NN.

Another production rule that occurs typically—in fact, almost exclusively—in the texts of native Chinese speakers is PP → VBG PP (by the Stanford parser), which almost always corresponds to the phrase *according to*. In Figure 2 we give the parse of a short sentence (*According to <R>, burning of*

```
(S1
  (S
    (ADVP (RB Overall))
    (, ,)
    (NP (NNP cyber))
    (VP (VBD cafeis)
      (NP (DT a) (JJ good) (NN place))
      (PP (IN as)
          (NP (JJ recreational)
          (NNP centre)))
      (PP (IN with)
          (NP
            (NP
              (DT a) (NN bundle))
              (PP (IN of)
                (NP (JJ up-to-dated)
                  (NN information))))))
  (. .)))
```

Figure 3: Parse illustrating parser correction

*plastic waste generates toxic by-products*—<R>is an in-text citation that was removed in the preparation of ICLE) that illustrates this particular construction. It appears that speakers of Chinese frequently use this phrase as a translation of *gēn jù*. So in this case, what is identified is not the sort of error that is of interest to contrastive analysis, but just a particular construction that is characteristic of a certain native speaker's language, one that is perfectly grammatical but which is used relatively infrequently by others and has a slightly unusual analysis by the parser.

We had expected to see more rules that displayed obvious ungrammaticality, such as VP → DT IN. However, both parsers appear to be good at 'ignoring' errors, and producing relatively grammatical structures (albeit ones with different frequencies for different native languages). Figure 3 gives the C&J parse for *Overall, cyber **cafeis** a good place as recreational centre with a bundle of **up-to-dated** information*. The correction of *up-to-dated* rather than *up-to-date* is straightforward, but the simple typographical error of running together *cafe* and *is* leads to more complex problems for the parser. Nevertheless, the parser produces a solid grammatical tree, specifically assigning the category VBD to the compound *cafeis*. This appears to be because both the Stanford and C&J parsers have implicit linguistic constraints such as assumptions about heads; these are imposed even when the text does not provide evidence for them.

We also present in Table 7 the top 10 rules chosen under the IG feature selection for the Stanford parser on the held-out set. A number of these, and those ranked lower, are concerned with punctuation: these seem unlikely to be related to native language, but perhaps rather to how students of a particular language background are taught. Others are more typical of the sorts of example we illustrated above: PP → VBG PP, for example, is typically connected to the *according to* construction discussed in connection with Figure 2, and it can be seen that the dominant frequency count there is for native Chinese speakers (column 6 of the counts).

## 7 Conclusion

In this paper we have shown that, using cross-sections of parse trees, we can improve above an already good baseline in the task of native language identification. While we do not make any strong claims for the Contrastive Analysis Hypothesis, the usefulness of syntax in the context of this problem does provide some support.

The best features arising from the classification have been horizontal cross-sections of trees, rather than the more general discriminative parse reranking features that might have been expected to perform at least as well. This relatively poorer performance by the reranking features may be due to a number of factors, all of which could be investigated in future work. One is the use of feature schema instances that did not appear in the largely grammatical WSJ; another is the extension of feature schemas; and a third is the use of a parser that does not enforce linguistic constraints such as the Berkeley parser (Petrov et al., 2006).

Examining some of the substructures showed some errors that were expected; other constructions that were grammatical, but were just characteristic translations of constructions that were common in the native language; and a large number where grammatical errors were glossed over by the parser's linguistic constraints, suggesting another purpose for further work with the Berkeley parser. Overall, the use of these led to an error reduction in over 30%

in the cross-validation evaluation with significance testing.

## Acknowledgments

The authors would like to acknowledge the support of ARC Linkage Grant LP0776267 and ARC Discovery Grant DP1095443, and thank the reviewers for useful feedback. Much gratitude is due to Mark Johnson for his guidance on the extraction of reranking features.

## References

Ethem Alpaydin. 2004. *Introduction to Machine Learning*. MIT Press, Cambridge, MA, USA.

Harald Baayen, Hans van Halteren, and Fiona Tweedie. 1996. Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing*, 11(3):121–131.

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

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 173–180, Ann Arbor, Michigan.

Stephen Clark and James R. Curran. 2007. Wide-Coverage Efficient Statistical Parsing with CCG and Log-Linear Models. *Computational Linguistics*, 33(4):493–552.

Michael Collins. 2000. Discriminative reranking for natural language processing. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML'00)*, Stanford, CA.

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

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*.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305.

Jennifer Foster, Joachim Wagner, and Josef van Genabith. 2008. Adapting a WSJ-trained parser to grammatically noisy text. In *Proceedings of ACL-08: HLT, Short Papers*, pages 221–224, Columbus, Ohio.

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, Louvian-la-Neuve.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Mark Johnson and Ahmet Engin Ural. 2010. Reranking the Berkeley and Brown Parsers. In *Proceedings of Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-10)*, pages 665–668, Los Angeles, CA, USA, June.

Mark Johnson, Stuart Geman, Stephen Canon, Zhiyi Chi, and Stefan Riezler. 1999. Estimators for stochastic unification-based grammars. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, MD.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan.

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.

Brian MacWhinney and Elizabeth Bates. 1989. *The Crosslinguistic Study of Sentence Processing*. Cambridge University Press, New York, NY, USA.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU: Automatic evaluation of

sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic.

Steven Myers. 2007. Introduction to phishing. In Markus Jakobsson and Steven Myers, editors, *Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Dominick Ng, Matthew Honnibal, and James R. Curran. 2010. Reranking a Wide-Coverage CCG Parser. In *Proceedings of Australasian Language Technology Association Workshop (ALTA'10)*, pages 90–98, Melbourne, Australia.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL'06)*, pages 433–440, Sydney, Australia, July.

Manfred Pienemann. 1998. *Language Processing and Second Language Development: Processability Theory*. John Benjamins, Amsterdam, The Netherlands.

Payam Refaeilzadeh, Lei Tang, and Huan Liu. 2009. Cross-validation. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, pages 532–538. Springer, 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.

Guihua Sun, Xiaohua Liu, Gao Cong, Ming Zhou, Zhongyang Xiong, John Lee, and Chin-Yew Lin. 2007. Detecting erroneous sentences using automatically mined sequential patterns. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 81–88, Prague, Czech Republic.

Michael Swan and Bernard Smith, editors. 2001. *Learner English: A teacher's guide to interference and other problems*. Cambridge University Press, 2nd edition.

Joel Tetreault, Jennifer Foster, and Martin Chodorow. 2010. Using parse features for preposition selection and error detection. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 353–358. Association for Computational Linguistics.

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.

Joachim Wagner, Jennifer Foster, and Josef van Genabith. 2009. Judging grammaticality: Experiments in sentence classification. *CALICO Journal*, 26(3):474–490.

Richard Wardhaugh. 1970. The Contrastive Analysis Hypothesis. *TESOL Quarterly*, 4(2):123–130.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61, Sydney, Australia, December.

Sze-Meng Jojo Wong and Mark Dras. 2010. Parser features for sentence grammaticality classification. In *Proceedings of the Australasian Language Technology Association Workshop 2010*, pages 67–75, Melbourne, Australia, December.

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.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*, pages 412–420.

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.