Native language identification and writing proficiency

Kristopher Kyle, Scott A. Crossley and YouJin Kim
Georgia State University

This study evaluates the impact of writing proficiency on native language identification (NLI), a topic that has important implications for the generalizability of NLI models and detection-based arguments for cross-linguistic influence (Jarvis 2010, 2012; CLI). The study uses multinomial logistic regression to classify the first language (L1) group membership of essays at two proficiency levels based on systematic lexical and phrasal choices made by members of five L1 groups. The results indicate that lower proficiency essays are significantly easier to classify than higher proficiency essays, suggesting that lower proficiency writers make lexical and phrasal choices that are more similar to other lower proficiency writers that share an L1 than higher proficiency writers that share an L1. A close analysis of the findings also indicates that the relationship between NLI accuracy and proficiency differed across L1 groups.

Keywords: native language identification: natural language processing: n-grams: learner corpus

1. Introduction[1]

Native language identification (NLI) is a statistical/machine-learning approach to the identification of the first language (L1) of a second language (L2) writer based on linguistic clues. NLI is a growing field with a number of applications such as authorship profiling and automatic writing feedback systems (Tetrault et al. 2013). Within the past few years, NLI has also recently been employed as a starting point for investigations into crosslinguistic influence (CLI) (e.g. Jarvis 2010, 2012). One

1. We would like to thank the organizers of the 2013 Native Language Identification Shared Task for providing the TOEFL11 corpus. We would also like to thank Ute Römer and Jessica Kyle for providing helpful comments on earlier versions of this paper.

International Journal of Learner Corpus Research 1:2 (2015), **187–209**. DOI 10.1075/ijlcr.1.2.01kyl
ISSN 2215–1478 / E-ISSN 2215–1486

question that has been noted (Bestgen et al. 2012; Tetreault et al. 2012) but not thoroughly addressed in NLI is the role of proficiency as a confounding variable. The role of proficiency is also an important (and contested) question in the area of CLI (Jarvis 2000). This study approaches the issue of the influence of proficiency on NLI with an eye towards building a foundation for future detection-based CLI studies (Jarvis 2010, 2012). Specifically, in this study, we explore the relationship between writing proficiency and NLI accuracy and whether this relationship is stable across different language groups.

2. Native Language Identification

The task of NLI involves three important features: corpora (typically learner corpora with L2 texts from multiple L1 groups), linguistic feature variables (e.g., lexical choices), and statistical/machine-learning algorithms. Generally speaking, an NLI model is created by identifying linguistic feature variables that distinguish L1 groups based on corpora of their L2 writing. If one wanted to distinguish L2 texts written by groups of L1 speakers of language A and language B, for example, one might determine if L2 writers from the language A group used any particular lexical items more often than those of the language B group. We might find, for example, that L2 texts written by the language group A tended to include more instances of the pronoun *we*, while L2 texts written by the language B group tended to include more instances of the pronoun *I*. We could then build a very simple NLI predictor model that would count the instances of *we* and *I* in an L2 text and predict whether the text was written by language A group or B based on the instances of *we* and *I*. This model is overly simplistic and, in practice, a larger number of linguistic variables would be used as predictor variables, which would require the use of statistical and/or machine-learning algorithms to make accurate predictions.

A number of learner-corpora have been used to build NLI predictor models in the past but, until recently, the most-used learner corpus in NLI studies has been the *International Corpus of Learner English* (ICLE) (Granger et al. 2009). ICLE was designed to include a set of comparable subcorpora divided by L1 (among other potential classifiers, such as gender and time spent in a country where English is spoken) and includes L2 texts (mostly argumentative essays) written by writers from 16 L1 groups. Although the ICLE is a robust resource for a number of applications, two main issues have been raised with regard to its usefulness in constructing NLI models. First, essay type/essay prompts are not equally distributed among language groups (see Brooke & Hirst 2012). Additionally, proficiency levels in the ICLE are not constant across language groups (as suggested by Koppel et al. 2005 and empirically demonstrated by Bestgen et al. 2012). These imbalances

in the subsets of the ICLE raise questions of generalizability for NLI predictor models constructed using this dataset. Recently, NLI studies have employed new learner corpora in an attempt to control for variables such as prompt and proficiency. Brooke & Hirst (2012), for example, employed the Lang-8 web corpus of learner texts (comprised mostly of journal entries) to avoid prompt bias. Another corpus that has been used to avoid the limitations of ICLE is the TOEFL11 corpus (Tetreault et al. 2012), which is a prompt-balanced and proficiency controlled corpus of argumentative essays written as part of the Test of English as a Foreign Language (TOEFL).

Within these various corpora, many different types of linguistic features have been employed as predictor variables in NLI models. These have included relatively transparent features such as lexical items (e.g. Jarvis et al. 2012), lexical n-grams (e.g. multi-word units; Jarvis & Paquot 2012, Jarvis et al. 2013, Kyle et al. 2013), lemma n-grams (Jarvis et al. 2013) and error patterns (e.g. Bestgen et al. 2012, Gebre et al. 2013, Wong & Dras 2009). Linguistic feature variables have also included character n-grams (e.g. Tsur & Rappoport 2007, Jarvis et al. 2013), part of speech (POS) n-grams (e.g. Gebre et al. 2013, Jarvis et al. 2013), indices of cohesion, lexical sophistication, syntactic complexity, and conceptual knowledge (Crossley & McNamara 2012), error patterns (Bestgen et al. 2012), and syntactic representations (e.g. Swanson 2013). Although a single linguistic feature variable type (e.g. lexical items) can be used to create successful NLI models (e.g. Brooke & Hirst 2012, Jarvis et al. 2012), the most accurate models include a variety of variable types. For example, Tetreault et al. (2012) used a large number of simple (e.g. lexical items) and complex (e.g., syntactic dependency relations) variables to achieve a classification accuracy of 84.6% for the 11 language groups in the TOEFL11. Additionally, Jarvis et al. (2013) used lexical 1–3 grams, lemma 1–3 grams, and POS 1–3 grams to achieve a classification accuracy of 84.7% for the 11 language groups included in the TOEFL11, which is the highest accuracy published for the TOEFL11 dataset.

The final component of an NLI model is the use of a statistical and/or machine-learning algorithm to predict the L1 group membership of an L2 text based on linguistic predictor variables. A number of approaches have been used, including multivariate analyses of variance (MANOVAs) and discriminant function analyses (DFA) (e.g. Jarvis & Paquot 2012; Crossley & McNamara 2012, Kyle et al. 2013), support vector machines (SVM) (e.g. Koppel et al. 2005), Naïve Bayes classifiers (e.g. Mayfield Tomokiyo & Jones 2001), multinomial logistic regression/maximum entropy (e.g. Tetreault et al. 2012) and decision trees (e.g., Brooke & Hirst 2012). Each classifier has various strengths and weaknesses, and no single classifier has emerged as clearly superior. Jarvis et al. (2013) for example used SVM and Tetreault et al. (2012) used multinomial logistic regression to achieve

similar classification results on the TOEFL11 database. DFA has also been used to achieve relatively high classification accuracies (e.g., Jarvis et al. 2012). DFA has been the main classification method used in studies that employ NLI as a starting point for CLI because it is the one of the more transparent classifiers with regard to the interpretation of results (Jarvis 2012).

Results from NLI studies have been informed through and recently have been used to inform the field of CLI. Traditionally, CLI has been explored using a comparison-based approach, which involves a combination of three types of evidence: intragroup homogeneity, intergroup heterogeneity, and cross-language congruity (Jarvis 2000). Intragroup homogeneity refers to similarities in L2 language use within a particular L1 group. In our previous example, the writers in language group A would demonstrate intragroup homogeneity because they systematically use the pronoun *we* instead of *I*. Intergroup heterogeneity refers to differences in L2 language use between L1 groups. In our previous example, language groups A and B would demonstrate intergroup heterogeneity in their use of personal pronouns in that language group A reports a higher incidence of the word *we* while language group B reports a higher incidence of the word *I*. Cross-language congruity refers to the similarities between an individuals' use of their L1 and their use of their L2. CLI studies that rely on the comparison-based argument generally use either frequently observed learner errors (e.g., article errors) and/or observed differences between a particular linguistic aspect of an L1 and an L2 (e.g., L2 learners of English whose L1 does not have an article system) as a starting point, and often investigate a single construct (e.g., article use; Diez-Bedmar & Perez-Paredes 2012).

Noting the potential advantage of using NLI as a starting point for CLI, Jarvis (2010, 2012) broadened his earlier model of CLI argumentation to include the detection-based argument. A detection-based argument for CLI is constructed using three forms of evidence (intragroup homogeneity, intergroup heterogeneity, and classification accuracy). Studies that build a detection-based argument for CLI rely on identifying patterns of language use that are shared by a group of users of an L2 (e.g. English) with the same L1 (e.g. Korean), but different from other L1 users (e.g. Chinese) of the same L2 (in this case English). The systematic patterns of language use by a particular L1 group are then used to predict the L1 group membership of an L2 text using statistical or machine learning models. The degree to which the models can accurately classify the L1 of the texts in question is a preliminary indicator of the strength of the CLI argument (Jarvis 2010). Essentially, the starting point of a detection-based argument for CLI is an NLI classification problem that is followed up with focused investigations of the linguistic predictors used in classification. Because the end-goal of such studies is to investigate

specific instances of CLI, most detection-based argument studies use linguistically straightforward predictors such as lexical items and lexical n-grams.

For example, Jarvis et al. (2012) investigated the lexical choices of Danish, Finnish, Portuguese, Spanish, and Swedish L1 users of L2 English using a corpus of written narrative descriptions of a short segment of a silent film. Using lexical items that were frequently produced by each L1 group as predictor variables in a discriminant function analysis, Jarvis et al. were able to predict the L1 group membership with an accuracy of 76.9%, demonstrating that accurate NLI results can be achieved using simple predictors. A post-hoc analysis identified 18 lexical items that created clear distinctions between L1 groups, a number of which were linked to L1 characteristics. Finnish writers, for example, were clearly distinguished from writers from other language groups by their frequent use of nouns and infrequent use of *he* and *she*. Jarvis et al. preliminarily conclude that this trend may be due to the absence of separate third person pronouns for males and females in Finnish.

In a follow-up study, Jarvis & Paquot (2012) explored the n-gram use of L2 English writers from 12 L1 backgrounds based on a subset of essays included in the *International Corpus of Learner English* (ICLE; Granger et al. 2009). Using the most frequent 1-grams, 2-grams, 3-grams, and 4-grams that were not prompt-based as predictors in a number of DFAs, Jarvis & Paquot achieved L1 classification accuracies ranging from 22% (using only 4-grams as predictors) to 53.6% (using 1-grams, 2-grams, 3-grams, and 4-grams as predictors in a stepwise DFA). The inclusion of n-grams as predictors increased the accuracy of the model due to the relative overuse of particular n-grams by particular L1 groups, which Jarvis & Paquot suggest may be due to L1 influence. The n-gram predictor *going to*, for example, was thought to be used more often by Spanish L1 writers of L2 English due to the corollary *ir a + infinitive* (go to + infinitive) construction in Spanish. In this study, Jarvis & Paquot demonstrated that although 1-grams tend to be more predictive of L1 than n-grams, the inclusion of n-grams in NLI predictor models increases the accuracy of the model.

Diverging from a lexical choice-based approach to NLI, Bestgen et al. (2012) investigated whether error patterns could be used to identify the L1 group membership of ICLE essays written in L2 English by L1 users of French, German, and Spanish. Using 48 formal, grammatical, lexical, lexico-grammatical, punctuation, word (redundant/missing words), and style errors as predictor variables in a DFA, Bestgen et al. were able to accurately classify 65.5% of the essays. Their findings indicated, for example, that essays written by Spanish L1 writers of L2 English had more spelling, article, lexical and phrase errors than essays written by L1 French or German writers of L2 English. The initial analysis suggested that error types can successfully be used to identify language groups, and can, therefore be useful in discussions of CLI.

As previously noted, however, proficiency may be a confounding factor in NLI studies. In a post-hoc analysis, Bestgen et al. (2012) assigned each ICLE essay used in their NLI analysis a proficiency score according to the Common European Framework (CEF). They found significant differences between the language groups represented in their dataset with regard to proficiency, which called into question whether the previously observed linguistic trends reported in other studies using ICLE (e.g., L1 Spanish writers' high frequency of spelling errors) were attributable to CLI or simply to proficiency differences. Tetreault et al. (2012) also reported on NLI accuracy differences based on proficiency levels in the TOEFL11 corpus. They reported highest classification accuracies for medium-proficiency essays, although their corpus also had a greater number of training essays for medium-proficiency learners. Overall, these studies demonstrated that proficiency may be an important confounding variable in NLI studies, though more work is clearly needed in this area.

In summary, even though much has been learned about NLI over the past decade, there are still gaps in research. One important area that needs more attention is the relationship between NLI and writing proficiency. This is an important issue for the generalizability of NLI accuracy across contexts, and is especially important for detection-based approaches to CLI that use NLI as a starting point. Thus, the current study explores the relationship between writing proficiency and NLI. Although this topic has been explored in the field of CLI from a comparison-argument based approach with regard to specific language constructs such as articles (e.g., Master 1987, 1997; Diez-Bedmar & Perez-Paredes 2012), it has not, to our knowledge, been explored systematically in the field of NLI (though see Bestgen et al. 2012, Tetreault et al. 2012). This study is guided by the following research questions:

1. Is there a relationship between the strength of NLI models and a writer's proficiency level?
2. If a relationship between the strength of NLI models and writing proficiency exists, is it consistent across L1 groups?

3. Method

3.1 Corpus

The current study uses a subset of the TOEFL11 corpus (Blanchard et al. 2013). As Blanchard et al. (2013) describe, the TOEFL11 corpus is a collection of argumentative independent essays from actual administrations of the TOEFL between

2006–2007. The essays included in the TOEFL11 corpus were written in English by individuals from 11 L1 backgrounds in response to one of eight prompts that ask test takers to give their opinion on an aspect of academic life, travel, economics, or community dynamics (see the Appendix for a complete list of prompts represented in the TOEFL11). During the independent writing task, test takers are given thirty minutes to write an essay about a given topic. Each essay in the corpus is coded for three characteristics: L1, essay prompt, and writing proficiency.

In the TOEFL11 corpus, learner writing proficiency is based on the holistic score given to the essay by two ETS-trained raters according to the TOEFL independent essay rubric. The TOEFL independent essay rubric ranges from a score of 0–5 (a copy of this rubric can be obtained at https://www.ets.org/), and is based on how well a test-taker addresses the topic and task, how well a test-taker organizes and develops an essay, and the language ability demonstrated by the test-taker in the essay (e.g., the sophistication of the test-taker's syntax, word choice, and idiomaticity). If scores given by the two raters agree or are adjacent, the essay scores are averaged. If the scores are not exact or adjacent matches, a third rater scores the essay and the two closest scores are averaged. For the TOEFL11 corpus, the original essay scores were reduced to three categories; low proficiency (scores between 1.0–2.0), medium proficiency (scores between 2.5–3.5), and high proficiency (scores between 4.0–5.0) (Blanchard et al. 2013). While these writing proficiency classifications may not be representative of overall language proficiency (Hulstjin 2007) they are likely representative of writing proficiency (e.g., Chapelle et al. 2008) and thus provide a statistically reliable basis for comparison among writing proficiency levels (although see Deluca et al. 2013 for a counter-argument).

One potential problem with the TOEFL11 corpus is that while it is relatively comparable across languages, it is not well balanced across writing proficiency levels (see Table 1). In order to investigate the relationship between writing proficiency and NLI with regard to lexical choices it is necessary to create a corpus that is as balanced as possible across language groups, prompts, and writing proficiency. This criterion effectively eliminated the low proficiency group due to the relatively low representation of low-proficiency essays in the TOEFL11 corpus. In addition, because holistic scores are highly correlated with essay length (e.g., Chodorow & Burstein 2004), a comparable number of medium or high proficiency essays would contain a much higher number of words than low proficiency groups, further complicating comparisons in lexical production between the groups. For these reasons, the decision was made to only compare lexical choices across medium and high proficiency groups.

Table 1. Distribution of writing proficiency levels in the TOEFL11 corpus

Language	Low	Medium	High
Arabic	296	605	199
Chinese	98	727	275
French	63	577	460
German	15	412	673
Hindi	29	429	642
Italian	164	623	313
Japanese	233	679	188
Korean	169	678	253
Spanish	79	563	458
Telugu	94	659	347
Turkish	90	616	394
Total	**1330**	**6568**	**4202**

Note. Adapted from Blanchard et al. (2013)

Five of the 11 language groups were chosen for analysis, based on the minimum number of essays available across the two writing proficiency groups, language family membership, and our own familiarity with the features of the languages. Although it was not possible to strictly control for prompt because administrations of the prompts differed geographically, each prompt is represented in each language and writing proficiency group (see Table 2 and Table 3 for an overview of the distribution of texts in the corpus). Languages selected for inclusion were Chinese, German, Hindi, Korean, and Spanish. Of these languages, the fewest essays represented in either writing proficiency level were high proficiency Korean essays (n=229[2]). Thus, using the Korean sub-corpus as a limiting factor, we randomly selected 229 essays from each of the five language groups at each of the two writing proficiency levels.

Table 2. Distribution of essay prompts in the medium proficiency corpus

Prompt	Chinese	German	Hindi	Korean	Spanish	Prompt Total	% of Corpus
1	31	39	29	37	30	166	14.5%
2	25	28	38	27	27	143	12.5%
3	27	39	20	21	14	121	10.6%
4	28	30	16	21	23	118	10.3%

2. This differs slightly from the information provided in Table 1 because a 1,100-essay subset of the TOEFL11 corpus had not been made available at the time of data analysis.

Table 2. (*continued*)

Prompt	Chinese	German	Hindi	Korean	Spanish	Prompt Total	% of Corpus
5	32	29	29	37	37	164	14.3%
6	22	5	9	29	32	97	8.5%
7	30	25	45	31	27	158	13.8%
8	34	34	45	26	39	178	15.5%
Texts per language	229	229	229	229	229	1145	100%
Number of words	73,731	72,775	78,832	68,772	72,455	366,565	

Table 3. Distribution of essay prompts in the high proficiency corpus

Prompt	Chinese	German	Hindi	Korean	Spanish	Prompt Total	% of Corpus
1	31	31	35	40	22	159	14.3%
2	31	31	34	29	43	168	14.7%
3	34	34	31	37	28	164	11.7%
4	29	29	33	15	18	124	10.8%
5	32	32	27	47	28	166	14.2%
6	37	37	7	7	36	124	11.4%
7	22	22	32	33	24	133	12.1%
8	13	13	30	21	30	107	10.8%
Texts per language	229	229	229	229	229	1145	100%
Number of words	87,155	84,651	87,567	86,980	84,490	430,843	

3.2 Predictor selection

As our research interests lie in the eventual application of NLI to CLI, we chose to use linguistically transparent predictor variables (following Jarvis et al. 2012 and Jarvis & Paquot 2012). Our predictor variables comprised n-grams from 1–5 words in length that were identified through series of keyness analyses conducted on training set corpora (see Section 3.3 for a description of how each corpus was divided into training and test sets). Keyness analyses identify items that occur statistically significantly more (which receive positive keyness values) or less (which receive negative keyness values) frequently in one corpus than another. For our

keyness analyses, we compared the n-gram frequencies in a particular subcorpus (e.g., medium-proficiency Chinese) with the aggregated n-gram frequencies of the other subcorpora at the same proficiency level (e.g., medium-proficiency German, Hindi, Korean and Spanish). The remainder of this section describes the predictor set selection conducted on medium proficiency essays written by the Chinese L1 group, a process that was repeated for all other language groups within the medium proficiency corpus (MPC) and the high proficiency corpus (HPC).

To create the medium Chinese predictor set, we first created a list of key n-grams using the Key Words feature in Wordsmith Tools 6 (Scott 2013). In order to ensure that the key n-gram list was not skewed by the prolific use of a particular n-gram by a particular test taker, we set the minimum threshold for inclusion as ten percent occurrence in the corpus (i.e. a particular n-gram had to occur in at least ten percent of the Chinese essays to be included in the key n-gram list). In addition, we set the significance threshold to $p < .0001$, and used the default log-likelihood method to calculate the keyness values. After the n-gram list was completed, all prompt-based words and their lemmas were removed in an attempt to enhance the generalizability of the findings. Prompt-based words were operationally defined as words included in the writing task prompt that had a frequency of 715 or less in the Brown corpus (Kucera & Francis 1967). This cut-off point, which was based on a qualitative analysis of the prompt-word lists, excluded function words and other commonly used words that are often not included in function-word stop lists. Finally, for the medium Chinese group, key n-grams were sorted into two lists based on whether the keyness values were positive (i.e. n-grams that occurred more often in Chinese essays than in essays written by other groups) or negative (i.e. n-grams that occurred less often in Chinese essays than essays written by other groups). For each, the lists were ranked by absolute keyness values (higher absolute keyness values indicate a stronger statistical relationship between a particular n-gram and a language group), and then checked for redundancy. For example, if the n-gram with the highest positive keyness value was 'more', and the n-gram 'have more' also occurred (but with a lower positive keyness value), the latter was removed from the list. This is important because automated counts of instances of the n-gram 'more' in an essay will also include all of the instances of 'have more'. As 'more' in this situation potentially has more predictive power (given the higher keyness value), it would remain in the predictor set. Each remaining n-gram was used as an individual predictor variable. Additionally, two aggregated indices were created for the Chinese group using the refined n-grams. The first aggregated index included all of the positive keyness n-grams identified in the Chinese group, while the second included all of the negative keyness n-grams.

After variables had been compiled for all of the languages in the MPC and HPC respectively, any redundant individual predictor variables were combined.

Instances of the variables were then automatically counted for each essay in the corpus using the *Custom List Analyzer* (CLA[3]), a freely available program developed for this project that counts user-defined lists of words, n-grams, and wildcard entries for a batch of texts. CLA provides normed (instances of variable in essay/number of words in essay) scores for each variable.

3.3 Statistical analysis

A number of previous studies have used DFA due to the relative ease of interpreting the results for follow-up investigations of CLI (e.g. Jarvis & Paquot 2012). Despite the advantages of DFA, two important assumptions of the statistical classifier are normality of data and homogeneity of variance (these assumptions also hold for MANOVAs). A preliminary analysis of the data indicated that almost all of the variables were not normally distributed (a problem that should be common in all analyses that rely on unigrams and other n-grams that are not extremely frequent). Thus, a non-parametric Kruskal Wallis test was conducted for each subcorpus using the languages from one writing proficiency group as independent variables and the predictor indices/n-grams as dependent variables. Variables that had significant differences were preliminarily selected as predictor variables for a multinomial simple logistic regression (SLR) model, which is analogous to maximum entropy methods used in a number of studies (e.g. Brooke & Hirst 2012, Tetreault 2012). SLR is an alternative statistical method to DFA that can be used to classify categorical variables (such as L1 groups) using continuous variables (such as the frequency of n-grams), which does not assume normal distributions of the data. One drawback of SLR is that it requires a relatively large sample size, though this is not usually a pertinent problem in corpus-based studies. Prior to use in an SLR, the preliminary variables for each subcorpus were checked for multicollinearity using a correlation matrix. Any two variables above a threshold of $r > .899$ were flagged for further analysis (Tabachnick & Fidell 2001). If two variables showed multicollinearity, the effect sizes produced by the Kruskal Wallis test were used to select which variables flagged in the correlation matrix would be retained, and which would be eliminated (i.e., the variable with the largest effect size was kept).

The data was also divided into training and test sets via a random sample that was stratified by language group. Essentially, the first data set (the training set) was used to create the predictor model, and the second data set (the test set) was used to test the accuracy of the model (Crossley & McNamara 2009). Although different training/test splits have been used ranging from 50/50 (Crossley & McNamara 2009) to 10/1 (e.g. Tetreault et al. 2013), or LOOCV (e.g., Jarvis & Paquot 2012)

3. http://www.kristopherkyle.com/tools.html (accessed May 8th 2015).

the current study uses a 67/33 split as suggested by Witten & Frank (2010), resulting in a training set for each corpus that included 770 texts and a test set for each corpus that included 335 essays. To prevent over-fitting the model, we also constrained the number of predictor variables in each SLR model to achieve a 10:1 ratio of cases to predictors. We thus chose the 77 variables with the highest effect size as predictor variables in each analysis based on the Kruskal-Wallis difference tests we conducted. A SLR was then conducted on the MPC and HPC training sets based on the predictor variables produced for each. The predictor model sets identified in the SLR were then used on the test sets to determine whether the model sets could generalize to a new population.

The difference (or lack thereof) in overall classification results between the two SLRs addresses whether writing proficiency level affects NLI (research question 1). The relative differences (or lack thereof) in classification accuracies for each language addresses whether the effect of writing proficiency on NLI is stable across language groups (research question 2).

4. Results

4.1 Medium proficiency

The SLR achieved a classification accuracy of 70.7% on the medium proficiency test set, which is significantly higher ($df = 16$, $n = 375$, $\chi^2 = 611.432$, $p < .001$) than the baseline accuracy of 20%. The reported Kappa = .633, indicates substantial agreement between actual and predicted L1 (Landis & Koch 1977). Table 4 includes the confusion matrix for the medium corpus. Rows indicate how essays from a particular language group were classified by the SLR. Columns indicate how many essays were classified as belonging to a particular L1 group by the SLR.

Table 5 includes the precision, recall and F-measure values for the medium proficiency SLR model. In machine learning applications, precision refers to the ratio of correct predictions to the total number of predictions made. Table 4, for example, indicates that 78 essays in the medium proficiency test set were predicted to be written by Chinese L1 writers. Of these, only 50 were actually written by Chinese writers. Our precision value for Chinese, then, is 50/78 = 0.641, which is reflected in Table 5. In other words, 64.1% of texts that were classified as 'Chinese' were correctly classified. Recall, on the other hand, is what one might traditionally refer to as accuracy. Recall is calculated by dividing the number of correct predictions by the number of instances that exist. Returning to our Chinese example, 50 out of 75 Chinese texts were correctly classified. The recall value for Chinese then

is 50/75 = .667. The F-measure is the harmonic mean of the recall and precision measures, which is calculated using the following formula:

$$F = 2\left(\frac{recall \cdot precision}{recall + precision}\right)$$

For more information regarding the evaluation of classifier models, see Witten & Frank (2010). Overall, the findings preliminarily indicate that the SLR model was able to classify the L1 groups based on their lexical choices. Table 6 includes the predictor n-grams used by the simple logistic regression to classify each language. Three letter, capitalized sequences are abbreviations for the positive and negative lists identified in the keyness analysis. The first letter indicates the corpus (in Table 6 these are all *M* for the medium corpus), the second letter indicates the first letter of the language the list represents (e.g., *C* stands for *Chinese*, *G* stands for *German*), and the third letter indicates whether the list indicates positive keyness (denoted by a *P*) or negative keyness (denoted by an *N*).

Table 4. Medium corpus confusion matrix

	Chinese	German	Hindi	Korean	Spanish	Total	Percent Correct
Chinese	**50**	9	2	10	4	75	66.67%
German	6	**56**	4	2	7	75	74.67%
Hindi	5	7	**55**	4	4	75	73.33%
Korean	9	2	6	**51**	7	75	68.00%
Spanish	8	5	6	3	**53**	75	70.67%
Total	78	79	73	70	75	375	70.67%

Table 5. Precision, recall, and f-measure for medium proficiency SLR model

Language	Precision	Recall	F-Measure
Chinese	0.641	0.667	0.654
German	0.709	0.747	0.727
Hindi	0.753	0.733	0.743
Korean	0.729	0.680	0.703
Spanish	0.707	0.707	0.707
Average	0.708	0.707	0.707

Table 6. Predictors used to classify each language in the medium proficiency model

Chinese		German		Hindi		Korean		Spanish	
Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
MCP	MCN	MGP	MSP	MHP	MHN	MKP	MCN	MCN	MSN
MHN	MHP	able to	choose	able to	because	be able	MCP	MGN	able to
a	MKP	an	hence	according to me	beeing	even though	MHN	MHN	can not
can not	able to	beeing	however	any	choose	first	MKN	MSP	first
choose	an	but on the	I	but	first	however	able to	be able	has to
first	any	has to	may	conclude	I think	in	an	but	have to
he	be able	often	particular	going	nt	nt	has to	going	however
hold	beeing	on	such as	have to	opinion	often	hence	his	may
however	but	opinion	then	hence	probably	second	his	its	nt
may	has to	or	various	hold	school	such as	hold	of	often
school	have to	people	we	I	special	the one hand	I	or	on
still	his	point	you	its	still	think	on	people	opinion
such as	its	possible		may	such as	various	on the other	person	school
you	often	probably		particular	the	we	people	probably	still
your	on the other	special		person	the one hand	you	point	this	than
	person	still		then	think It	you are	probably		then
	probably	than		towards	to be		this		various
	to be	that		would	why		towards		which
	we	the		your			why		would
	which	the one hand							
		this							
		to be							
		why							
		would							

Note: Three-letter sequences indicate keyness n-gram lists (e.g., MCP). The initial letter indicates writing proficiency level, the second indicates language group (C = China, G = German, etc.), and the third letter indicates the keyness polarity (P = positive, N = negative).

4.2 High proficiency

The SLR achieved a classification accuracy of 57.6% on the high proficiency test set, which is significantly higher (df = 16, n = 375, χ^2=360.818, p< .001) than the baseline accuracy of 20%. The reported Kappa = .470, indicates moderate agreement between actual and predicted L1 (Landis & Koch 1977). Table 7 includes the high proficiency test set confusion matrix. Table 8 includes the precision, recall and F-measure values for the high proficiency test set. Overall, the findings preliminarily indicate that the SLR model was able to classify the L1 groups based on their lexical choices. Table 9 includes the n-gram variables used by the logistic regression to identify each L1 group.

Table 7. High training set confusion matrix

	Chinese	German	Hindi	Korean	Spanish	Total	Percent Correct
Chinese	**36**	5	6	21	7	75	48.00%
German	4	**50**	8	5	8	75	66.67%
Hindi	5	10	**52**	5	3	75	69.33%
Korean	17	7	7	**38**	6	75	50.67%
Spanish	6	10	10	9	**40**	75	53.33%
Total	68	82	83	78	64	375	57.60%

Table 8. Precision, recall, and f-measure for high proficiency SLR model

Language	Precision	Recall	F-Measure
Chinese	0.529	0.480	0.503
German	0.610	0.667	0.637
Hindi	0.627	0.693	0.658
Korean	0.487	0.507	0.497
Spanish	0.625	0.533	0.576
Average	0.576	0.576	0.574

4.3 Statistical comparison of model accuracy

To determine whether the difference in classification accuracy between the predictor models for the medium and high corpora was statistically significant and meaningful, a Mann-Whitney U test was conducted. Each text for each writing proficiency level was coded as being correctly predicted (1) or incorrectly

Table 9. Predictors used to classify each language in the high proficiency model

Chinese		German		Hindi		Korean		Spanish	
Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
HCP	HCN	HGP	HGN	HCN	HCP	HKP	HCN	HGN	HCP
a person	HGN	HSN	a person	HGN	HGP	a person	HGP	HSP	HSN
any	HHP	and to	any	HHP	HHN	became	HKN	an	a person
choose	an individual	at	beneficial	HSN	HKN	beneficial	HSN	but	any
especially	because	easier	choose	any	because	cannot	a	easier	beneficial
experience	easier	especially	even though	an individual	even though	even though	an	even though	cannot
far as I	etc	has to	experience	but	far as I	however	at	etc	especially
feel that	fuel	have to	far as I	come	first	I	come	experience	first
first	have to	I	fuel	etc	however	I could	easier	has	fuel
hence	I could	I could	has	feel that	I think	I feel	experience	its	has to
however	increase In	I think	particular	fuel	in my	increase in	feel that	more	his
I think	or	in my	the place	hence	maybe	more	has	of	however
of course	particular	maybe	these days	I	necessary	necessary	hence	or	I
often	person	necessary	visit	I feel	of course	often	maybe	particular	I feel
the place	this	of course	was	its	the	person	question	person	I think
was	towards	often	we	jack	think that	the	this	that	jack
	transport	question	you	now		these days	transport	think that	often
	very	that		particular		various	very	this	these days
	visit	the		these days		visit	will always	transport	various
		this		towards		was	would	very	which
		transport		we				visit	
		which		which				will always	
		will always		you/your					

Note: Three-letter capitalized sequences indicate keyness n-gram lists. The initial letter indicates writing proficiency level (H = high), the second indicates language group (C = China, G = German, etc.), and the third letter indicates the keyness polarity (P = positive, N = negative).

predicted (0). The results of the Mann-Whitney U test indicate that the classification accuracy of the model built on the medium-proficiency corpus is statistically significantly more accurate than the classification accuracy of the model built on the high-proficiency corpus ($z = -3.728$, $p < .001$). Furthermore, the effect size ($r = .136$) is meaningful but small, according to Cohen (1988). This finding indicates a relationship between writing proficiency and the lexical choices made by particular L1 groups writing in English, and suggests that the lexical choices made by writers of particular L1 background writing in English become less uniform as writing proficiency increases. Table 10 includes a summary of the differences in classification accuracy for the overall models and by language. The differences in classification accuracies for each language group across writing proficiency levels suggest that the relationship between writing proficiency and NLI is not uniform across L1 groups.

Table 10. Classification across writing proficiency groups

Language	Medium	High	Difference	Effect Size (r)
Chinese	66.67%	48.00%	*18.67%	0.188
German	74.67%	66.67%	8.00%	0.086
Hindi	73.33%	69.33%	4.00%	0.044
Korean	68.00%	50.67%	*17.33%	0.176
Spanish	70.67%	53.33%	*17.34%	0.178
Total	70.67%	57.60%	**13.07%	0.136

Note: * indicates p < .05, ** indicates p < .001

5. Discussion

The SLR predictor model created for the medium proficiency corpus using key n-grams from the training set as predictor variables was able to predict the L1 group membership of the essays included in the medium proficiency test set with an accuracy of 70.7%. This was significantly more accurate than the high proficiency test-set performance of the SLR model built using key n-grams from the high proficiency training set (57.6% classification accuracy). The comparison of these statistical models, which were based on differences in language use between five L1 groups writing in L2 English across two writing proficiency levels, indicates that NLI models can more accurately classify essays that are considered to be "medium proficiency" than those that are considered to be "high proficiency" based on the lexical choices made.

Tables 6 and 9 above comprise the variables used by the SLR to predict whether an L2 English essay was written by a particular L1 group. Some of these lexical choices include systematic spelling choices, such as the tendency for medium-proficiency German writers to spell the word *being* as *beeing*. Others, such as the medium-proficiency Korean writers to use the pronoun *we* more than *I* can preliminarily be linked to L1 tendencies. Kim (2009), for example, suggested that Korean writers tend to use *wuli*, the Korean equivalent of English *we* more often than other personal pronouns to show in-group membership (e.g., Cutting 2001). The aggregated indices in the medium-proficiency corpus loaded as expected with few exceptions in that positively key aggregated lists for a particular language group loaded as a positive predictor for that language group and the same trend followed for the negative key lists. In the high-proficiency corpus, however, substantial overlap in aggregated indices occurred. Both the positively key aggregated Hindi and Spanish lists, for example, were positive predictors for Hindi. These results may be due to the general similarity of L2 writing at high proficiency levels, and may also be due to the way in which we calculated keyness. We identified key n-grams by comparing the writing of one language group to the aggregate writing of the other language groups, which may have resulted in overlap of key n-grams between some of the language groups. Overall it was found that the lexical choices made by writers demonstrate more intergroup heterogeneity and intragroup homogeneity at the medium proficiency level than at the high proficiency level. A closer inspection, however, reveals differences among L1 groups in the relative classification accuracy between the medium and high proficiency groups. The difference in classification accuracy between the medium and high proficiency Chinese, Korean, and Spanish L1 groups were all statistically significant with meaningful but small effect sizes. The difference in classification accuracy between the medium and high proficiency Hindi and German L1 groups followed the same trend as the others (i.e., the medium proficiency essays were classified with higher accuracy than the high proficiency essays), but the difference in classification accuracy did not reach statistical significance. For both the medium and high proficiency test sets, both German and Hindi L1 were classified most accurately, suggesting that the lexical choices made by German and Hindi L1 writers of English may demonstrate higher levels of intragroup homogeneity at both the medium and high proficiency levels than Chinese, Korean, and Spanish L1 writers of L2 English.

It is certainly conceivable that the lexical choices made by L1 Hindi writers of English represented in these corpora are affected by the variety of English that is spoken in India. This explanation is certainly a tempting one, especially because English is an official language of India. Future research is warranted to address this question empirically, a task that is beyond the scope of this paper. The reasons for the distinctiveness of the lexical choices made by L1 German writers is less clear,

though the relatively close linguistic relationship between German and English may play a role. Again, this needs to be examined empirically before any firm conclusions can be made.

Although the TOEFL11 corpus is an improvement from the ICLE with regard to controlling for proficiency and prompt, future studies should attempt to control for prompt more tightly than we were able to in this study. Although all essay prompts included in the study were represented in each language group investigated, the distribution was not perfectly equal. In addition, very little metadata is available for the essays in the TOEFL11 corpus, resulting in potentially confounding variables that we were not able to control for, such as level and type of education, socio-economic status, and gender. Future research should attempt to control for these variables and/or determine the effect they have on the accuracy of NLI models. Future studies are also warranted to look at a wider range of languages than we did to determine whether the results of this study are further generalizable.

With regard to proficiency measures, it is also pertinent to explore different measures of proficiency (such as the Common European Framework of Reference for Languages). In this study, we operationalized the construct of proficiency very narrowly as a writing proficiency score given based on a timed independent essay. Perhaps a good starting point would be a standardized test score, such as a complete TOEFL or IELTS score. Additionally, it is important to explore linguistic variables that go beyond lexical cues such as grammatical structures following recent work on Criterial Features (e.g., Hawkins & Buttery 2010). Such research would help clarify the questions we have preliminarily answered through our analyses.

6. Conclusion

This study explored the relationship between writing proficiency and NLI using a corpus-based, statistical approach that reported a significant relationship between writing proficiency and the lexical choices made by L1 groups in L2 English essays. The SLR model built using key n-grams from the medium proficiency corpus achieved a test set classification accuracy that was significantly higher than what would be expected by chance. The SLR model built using key n-grams from the high proficiency corpus also achieved a test-set classification accuracy that was significantly higher than would be expected by chance, and was also significantly lower than the classification accuracy for the medium test set. This indicates that as writing proficiency increases, the intragroup homogeneity decreases with regard to lexical choices. Although the effect was low ($r = .136$), this finding has important implications for future NLI and detection-based CLI studies, namely that writing proficiency is an important confounding variable that should be controlled.

This study also demonstrated that the effects of writing proficiency on NLI accuracy are not uniform across language groups. A statistically significant relationship existed between writing proficiency level and the accuracy of the NLI model for the Chinese, Korean, and Spanish L1 groups. Although the general trend held true (NLI accuracy decreased as writing proficiency increased) for the German and Hindi L1 groups, the differences in NLI accuracy between the two groups were not statistically significant. The findings also indicate that follow-up research should focus on the role that language family (in this case the Indo-European language family) plays in helping to predict the native language of L2 writers.

References

Bestgen, Y., Granger, S. & Thewissen, J. 2012. "Error patterns and automatic l1 identification". In S. Jarvis & S. A. Crossley (Eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-Based Approach*. Bristol: Multilingual Matters, 127–153.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A. & Chodorow, M. 2013. *TOEFL11: A Corpus of Non-native English*. Princeton: Educational Testing Service.

Brooke, J. & Hirst, G. 2012. "Robust, lexicalized native language identification". In *Proceedings of the 24th International Conference on Computational Linguistics* (COLING-2012). Mumbai: The COLING 2012 Organizing Committee, 391–407.

Chapelle, C. A., Enright, M. K. & Jamieson, J. M. 2008. *Building a validity argument for the test of English as a foreign language*. New York: Routledge.

Chodorow, M. & Burstein, J. 2004. *Beyond Essay Length: Evaluating e-rater Performance on TOEFL® Essays (Report No. 73)*. Princeton: Educational Testing Service.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum.

Crossley, S. A. & McNamara, D. S. 2009. "Computational assessment of lexical differences in L1 and L2 writing", *Journal of Second Language Writing* 18(2), 119–135. DOI: 10.1016/j.jslw.2009.02.002

Crossley, S. A. & McNamara, D. S. 2012. "Detecting the first language of second language writers using automated indices of cohesion, lexical sophistication, syntactic complexity and conceptual knowledge". In S. Jarvis & S.A. Crossley (Eds.), *Approaching Language Transfer Through Text Classification: Explorations in the Detection-based Approach*. Bristol: Multilingual Matters, 106–126.

Cutting, J. 2001. "The speech acts of the in-group", *Journal of Pragmatics* 33(8), 1207–1233. DOI: 10.1016/S0378-2166(00)00056-4

DeLuca, C., Cheng, L., Fox, J., Doe, C. & Li, M. 2013. "Putting testing researchers to the test: An exploratory study on the TOEFL iBT", *System* 41(3), 663–676. DOI: 10.1016/j.system.2013.07.010

Diez-Bedmar, M. B. & Perez-Paredes, P. 2012. "A cross-sectional analysis of the use of the English article system in Spanish learner writing". In Y. Tono, Y. Kawaguchi & M. Minegishi (Eds.), *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*. Tokyo: John Benjamins, 139–157. DOI: 10.1075/tufs.4.13die

Gebre, B. G., Zampieri, M., Wittenburg, P. & Heskes, T. 2013. "Improving native language identification with TF-IDF weighting". In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta: Association for Computational Linguistics, 111–118.

Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. (Eds.). 2009. *International Corpus of Learner English*. Version 2 (Handbook + CD-ROM). Louvain-la-Neuve: Presses universitaires de Louvain.

Hawkins, J. A. & Buttery, P. 2010. "Criterial features in learner corpora: Theory and illustrations", English Profile Journal 1(1), 1-23. DOI: 10.1017/S2041536210000036

Hulstijn, J.H. 2007. "The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency", *The Modern Language Journal* 91(4), 662–666. DOI: 10.1111/j.1540-4781.2007.00627_5.x

Jarvis, S. 2000. "Methodological rigor in the study of transfer: Identifying L1 influence in the interlanguage lexicon", *Language Learning* 50(2), 245–309. DOI: 10.1111/0023-8333.00118

Jarvis, S. 2010. "Comparison-based and detection-based approaches to transfer research". In L. Roberts, M. Howard, M. Ó. Laoire & D. Singleton (Eds.), EUROSLA Yearbook 10. Amsterdam: Benjamins, 169-192. DOI: 10.1075/eurosla.10.10jar

Jarvis, S. 2012. "The detection-based approach: An overview". In S. Jarvis & S.A. Crossley (Eds.), *Approaching Language Transfer Through Text Classification: Explorations in the Detection-based Approach*. Bristol: Multilingual Matters, 1–33.

Jarvis, S., Castañeda-Jiménez, G. & Nielsen, R. 2012. "Detecting L2 writers' L1s on the basis of their lexical styles". In S. Jarvis & S.A. Crossley (Eds.), *Approaching Language Transfer Through Text Classification: Explorations in the Detection-based Approach*. Bristol: Multilingual Matters, 34–71.

Jarvis, S., Bestgen, Y. & Pepper, S. 2013. "Maximizing classification accuracy in native language identification". In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta: Association for Computational Linguistics, 111–118.

Jarvis, S. & Paquot, M. 2012. "Exploring the role of n-grams in L1 identification". In S. Jarvis & S.A. Crossley (Eds.), *Approaching Language Transfer Through Text Classification: Explorations in the Detection-based Approach*. Bristol: Multilingual Matters, 71–105.

Kim, C. 2009. "Personal pronouns in English and Korean texts: A corpus-based study in terms of textual interaction", *Journal of Pragmatics* 41(10), 2086–2099. DOI: 10.1016/j.pragma.2009.03.004

Koppel, M., Schler, J. & Zigdon, K. 2005. "Determining an author's native language by mining for text errors". In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in data Mining*. Chicago: Association for Computing Machinery, 624–628. DOI: 10.1145/1081870.1081947

Kucera, H. & Francis, W. N. 1967. *Computational Analysis of Present-day American English*. Providence: Brown University Press.

Kyle, K., Crossley, S.A., Dai, J. & McNamara, D.S. 2013. "Native language identification: A key ngrams approach". In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta: Association for Computational Linguistics, 242–250.

Landis, J.R. & Koch, G.G. 1977. "The measurement of observer agreement for categorical data", *Biometrics* 33(1), 159–174. DOI: 10.2307/2529310

Master, P. 1987. *A Cross-linguistic Interlanguage Analysis of the Acquisition of the English Article System*. Unpublished Ph.D. Dissertation. University of California, Los Angeles.

Master, P. 1997. "The English article system: Acquisition, function, and pedagogy", *System* 25(2), 215–232. DOI: 10.1016/S0346-251X(97)00010-9

Mayfield Tomokiyo, L. & 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* (NAACL '01), unpaginated electronic document. Cambridge, MA: The Association for Computational Linguistics.

Scott, M. 2013. *WordSmith Tools 5.0*. Liverpool: Lexical Analysis Software.

Swanson, B. 2013. "Exploring syntactic representations for native language identification". In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta: Association for Computational Linguistics, 111–118.

Tabachnick, B. G. & Fidell, L. S. 2001. *Using Multivariate Statistics*(4th ed). Needham Heights: Allyn & Bacon.

Tetreault, J., Blanchard, D. & Cahill, A. 2013. "A report on the first native language identification shared task". In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta: Association for Computational Linguistics, 48–57.

Tetreault, J., Blanchard, D., Cahill, A. & Chodorow, M. 2012. "Native tongues, lost and found: resources and empirical evaluations in native language identification". In *Proceedings of the 24th International Conference on Computational Linguistics* (COLING-2012). Mumbai: The COLING 2012 Organizing Committee, 2585–2602.

Tsur, O. & 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*. Cambridge: The Association for Computational Linguistics, 9–16. DOI: 10.3115/1629795.1629797

Witten, I. A. & Frank, E. 2010. *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). San Francisco: Elsevier.

Wong, S.-M.J. & Dras, M. 2009. "Contrastive analysis and native language identification". In *Proceedings of the Australasion Language Technology Association*. Cambridge: The Association for Computational Linguistics, 53–61.

Appendix. Prompts included in the TOEFL 11 corpus

Prompt 1	Do you agree or disagree with the following statement? It is better to have broad knowledge of many academic subjects than to specialize in one specific subject. Use specific reasons and examples to support your answer.
Prompt 2	Do you agree or disagree with the following statement? Young people enjoy life more than older people do. Use specific reasons and examples to support your answer.

Prompt 3	Do you agree or disagree with the following statement? Young people nowadays do not give enough time to helping their communities. Use specific reasons and examples to support your answer.
Prompt 4	Do you agree or disagree with the following statement? Most advertisements make products seem much better than they really are. Use specific reasons and examples to support your answer.
Prompt 5	Do you agree or disagree with the following statement? In twenty years, there will be fewer cars in use than there are today. Use reasons and examples to support your answer.
Prompt 6	Do you agree or disagree with the following statement? The best way to travel is in a group led by a tour guide. Use reasons and examples to support your answer.
Prompt 7	Do you agree or disagree with the following statement? It is more important for students to understand ideas and concepts than it is for them to learn facts. Use reasons and examples to support your answer.
Prompt 8	Do you agree or disagree with the following statement? Successful people try new things and take risks rather than only doing what they already know how to do well. Use reasons and examples to support your answer.

Corresponding author's address

Kristopher Kyle
Department of Applied Linguistics & ESL
Georgia State University
P.O. Box 4099
Atlanta, GA 30302–4099
United States of America

kristopherkyle1@gmail.com