Native Language Identification: A Key N-gram Category Approach

Kristopher Kyle, Scott Crossley
Georgia State University
34 Peachtree Ave, Ste 1200
Atlanta, GA 30303
kkyle3@student.gsu.edu,
scrossley@gsu.edu

Jianmin Dai, Danielle S. McNamara
Arizona State University
PO Box 872111
Tempe, AZ 85287
Jianmin.Dai@asu.edu,
dsmcnamara1@gmail.com

Abstract

This study explores the efficacy of an approach to native language identification that utilizes grammatical, rhetorical, semantic, syntactic, and cohesive function categories comprised of key n-grams. The study found that a model based on these categories of key n-grams was able to successfully predict the L1 of essays written in English by L2 learners from 11 different L1 backgrounds with an accuracy of 59%. Preliminary findings concerning instances of crosslinguistic influence are discussed, along with evidence of language similarities based on patterns of language misclassification.

1. Introduction

Native language identification (NLI) is generally an automated task that can be used in authorship profiling (Wong & Dras, 2009) and in assisting automatic writing evaluation systems provide focused feedback (e.g., Rozovskaya & Roth, 2011). NLI is achieved by identifying patterns of language use that are common to a group of users of a particular second language (L2; e.g., English) that share a native language (L1). Useful to the discussion of these patterns is the concept of crosslinguistic influence (CLI), which references 'the consequences - both direct and indirect - that being a speaker of a particular native language (L1) has on the person's use of a later learned language (Jarvis, 2012, p.1). Beyond its theoretical applications, CLI can also be used to inform L2 classroom pedagogy (Granger, 2009; Laufer & Girsai, 2008). NLI studies, then, are informed by and can inform CLI, and have diverse applications.

The current study seeks to add to the discussions of NLI and CLI by testing the efficacy of a new approach – the use of grammatical, rhetorical, semantic, syntactic, and cohesive function categories of key n-grams.

2. Background

In this section we outline two approaches to CLI, provide a selected review of relevant literature, and address gaps in the current body of NLI research.

2.1 Approaches to CLI

Jarvis (2000, 2010, 2012) has outlined two approaches to the investigation of CLI: a comparison-based and a detection-based approach. The comparison-based approach is generally constructed based on specific observed difference between language systems (e.g., article usage in English as compared to article usage in Korean). Whether or not these L1 differences affect L2 production is then analyzed by examining example texts (e.g., inappropriate use of articles by native speakers of Korean writing in English as an L2). The detection-based argument, on the other hand, is built with the opposite trajectory. Instead of beginning with hypotheses based on differences in language systems, researchers begin by identifying patterns of language use (e.g., inappropriate article use) that occur regularly by members of an L1 that

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 242–250,
Atlanta, Georgia, June 13 2013. ©2013 Association for Computational Linguistics

use a particular L2 (intragroup homogeneity) but do not occur regularly by other L1 users of the same L2 (intergroup heterogeneity). These patterns of use are then verified through statistical and machine learning techniques that use these patterns to predict the L1 group membership of L2 texts (i.e., NLI).

Recent advances in corpus development and natural language processing allow for larger numbers of texts to be searched using a greater number of linguistic features. These features can then be used to create an NLI predictor model. A successful model not only fulfills the NLI task, but provides further evidence that the observed patterns of language use can be attributable to CLI. While Type I errors are certainly a potential issue in this argument, Jarvis (2012) explains that false positives can be mitigated by balancing or controlling for potentially confounding variables (e.g., proficiency levels and essay prompts) during the construction of the target corpus.

2.2 Selected literature review

A limited but growing number of studies have investigated CLI using the detection-based approach, many of which are included in a volume edited by Jarvis and Crossley (2012). Researchers have explored the topic of CLI in the areas of lexical style (Jarvis et al., 2012a), lexical n-grams (Jarvis & Paquot, 2012), character n-grams (Tsur & Rappoprot, 2007), using variables related to cohesion, lexical sophistication, syntactic complexity and conceptual knowledge (Crossley & McNamara, 2012), error patterns (Bestgen, et al., 2012; Wong & Dras, 2009), and a combination of these approaches (Jarvis et al., 2012b; Koppel et al., 2005; Mayfield Tomokiyo & Jones, 2001, Wong & Dras, 2009).

Such studies have demonstrated relatively strong success rates for classifying an L2 writing sample based on the L1 of the writer. For instance, Jarvis and Paquot (2012), using 1-4-grams as predictor variables on a subset of argumentative essays included in the International Corpus of Learner English (ICLE) (Granger et al., 2009) achieved a 53.6% classification accuracy for 12 groups of L1s. Crossley and McNamara (2012) used features related to cohesion, lexical sophistication, syntactic complexity, and conceptual knowledge taken from the computational tool Coh-Metrix (Graesser et al., 2004) to classify essays written in English by Czech, Finnish, German, and Spanish participants and achieved an L1 classification accuracy of 65-67.6%. Using error types, Bestgen et al. (2012), on 223 ICLE essays written by French, German, and Spanish L1 participants, achieved a classification accuracy of 65%. In a follow-up study, Jarvis et al. (2012b) explored the relative efficacy of these three CLI methods (n-grams, Coh-Metrix indices, and error types) using the corpus found in Bestgen et al. (2012). When all three approaches were used in the classification task, the accuracy increased to 79%.

2.3 Weakness of extant research in CLI

Although the studies discussed so far have produced statistical models that can predict the L1 group of a text written in L2 English with accuracies well above chance, the degree to which these studies have demonstrated instances of CLI may be questionable as they draw on the ICLE corpus, which is arguably imbalanced (Jarvis et al., 2012a, and Mayfield Tomokiyo, & Jones, 2001 being the exceptions). While ICLE was designed with an attempt to control for a number of variables, the proficiency levels vary across language groups (as suggested by Koppel et al., 2005, and empirically confirmed by Bestgen et al., 2012) and though the argumentative texts are limited to a particular set of prompts within the corpus, these prompts are not equally distributed across language groups, raising the question of the degree to which the observed differences in texts were due to CLI, proficiency level, or essay prompt.

In addition, many of the linguistic features previously investigated did not lend themselves to providing strong links between observed differences and CLI (e.g., the word concreteness and word frequency variables investigated in Crossley & McNamara, 2012). A potentially promising method that has not been applied to detection-based CLI studies that may address these limitations is the use of rhetorical, syntactic, grammatical and cohesive categories comprised of key n-grams. Such features have recently been investigated by Crossley, Defore, Kyle, Dai, and McNamara (submitted for publication), in which they explored their usefulness for assessing the efficacy of an automatic writing evaluation (AWE) system. In this study, Crossley et al. separated a corpus of essays into introduction, body, and conclusion paragraphs, and then further separated these into high and low proficiency categories based on overall essay score. They then identified n-grams that occurred significantly more often (positive keyness values) in paragraphs of a certain type (e.g., introduction) from high scoring essays than the same type of paragraphs from low-scoring essays. Additionally, they identified n-grams that occurred significantly less often (negative keyness values) in high-scoring paragraphs of a certain type than low-scoring paragraphs of the same type. Positively and negatively key n-grams for each paragraph type were then separated into categories based on their rhetorical, syntactic, grammatical, and cohesive features. These categories were then successfully used as variables in a multiple regression to create a model that accounted for between 24%-33% of the variance in essay scores. This study demonstrates the efficacy of using grammatical, rhetorical, syntactic, and cohesive function categories of key n-grams to identify instances of linguistic variation that successfully predict essay quality. These findings hold promise for the use of similar methods to contribute to the study of CLI by identifying linguistic variation across different L1 groups writing in the same L2.

2.4 Goals of the current study

The current study, while drawing on previous research (notably Jarvis & Paquot, 2012 and Crossley et al., submitted for publication), contributes to the detection-based CLI discussion by: a) examining a prompt and proficiency-controlled corpus and, b) using n-gram indices related to grammatical, rhetorical, semantic, syntactic, and cohesive functions to assess difference in L2 essays based on the L1 of the writers. This study is guided by the following research questions:

1. Can a model consisting of functional categorical n-grams predict the native language of an L2 writer of English?

2. Does the resulting model inform theories of CLI?

3. Method

In this section, we describe the corpus used for our training and test set, the methods used for key n-gram identification, and the grouping of these n-grams into grammatical, rhetorical, semantic, syntactic, and cohesive categories.

3.1 Corpus

For this project we used an 11,000 essay subset of the 12,100 essay TOEFL11 corpus (Blanchard, Tetreault, Higgins, Cahill, & Chodorow, 2013). The TOEFL11 corpus is comprised of independent task essays written during administrations of the Test of English as a Foreign Language (TOEFL) between 2006-2007 (Blanchard et al., 2013). The corpus is balanced across 11 native language (L1) groups, includes responses to eight different independent-task prompts, and includes essays written by low, medium, and high proficiency writers. The languages represented include Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish. Following the procedures of the NLI shared task (Tetreault, Blanchard, & Cahill, 2013), 1,100 of the original 11,000 essays were set aside as the test set, leaving a training corpus of 9,900 essays.

3.2 Identifying key n-grams

In this study, we considered n-grams from 1-10 words in length. N-grams were considered to be *key* if they occur in a corpus significantly more or less frequently than in a reference corpus. We identified key n-grams using the KeyWords function of Wordsmith Tools 6 (Scott, 2013) and the default log likelihood method of identifying key n-grams (McEnery & Hardie, 2012). To ensure that the keyness of a particular n-gram was representative of use across a particular L1 group and not due to prolific use by a small number of individuals, we set the minimum threshold for inclusion at a range of 10 percent (n-grams had to occur in at least 10 percent of the texts written by a particular L1 group). Using these parameters, we conducted keyness tests for each language group. To create the key n-gram list for the Arabic group, for example, we compared the frequency of n-grams in the Arabic group to the frequency of n-grams in all of the other language groups combined. This process was completed for each language group until a key n-gram list existed for each.

Because one of the goals of our study was to generalize instances of CLI to essays written on prompts other than those included in the TOEFL11 Corpus, it was important to remove all prompt-based words from our key n-gram lists. Removing *all* words occurring in the prompts from the n-grams list would remove a number of high frequency words that may not be prompt-based (e.g., *the*, *to*), so prompt-based words were operationally defined as content words and their lemmas included in the prompt that had a Kucera and Francis (1967) written frequency value of 715 or less. N-grams were removed from potential predictor sets if they contained any of these prompt-based words. The remaining key n-grams for each language group were then sorted by absolute keyness in each group and filtered for redundancy. For example, prior to this stage, the Chinese key n-gram list included both *more* and *have more*. Because *more* had a higher absolute keyness value than *have more*, *have more* was removed from the Chinese key n-gram list.

Table 1 provides a summary of the length of key n-grams identified in each stage of the selection process. Although n-grams from 1-10 words in length were initially considered, no n-grams longer than 5-grams were identified as being key. Additionally, all 5-grams, such as the key Chinese n-gram ‘group led by a tour’, and the Telugu n-gram ‘agree with the statement that’ contained prompt-based words and were removed from further consideration. After the final n-gram refining step, the longest n-gram was a single 4-gram, the Turkish n-gram ‘on the other hand’.

N-gram Length	Original	No Prompt Words	After Final Sort
5	5	0	0
4	19	3	1
3	110	54	8
2	699	512	147
1	1100	877	770
Total	1933	1446	926

Table 1: Length of key n-grams.

3.3 Grouping of key n-grams into indices

The last stage in our variable selection process was to group the key n-grams in each language group into categories. First, two indices for each language group were created. The first included all n-grams with positive keyness values that remained after the filtering process described above. The second included all of the n-grams with negative keyness values after filtering. Next, positive and negative n-grams were sorted into grammatical, rhetorical, semantic, syntactic, and cohesive function categories by two trained linguists with experience in the area of second language writing. The purpose of sorting n-grams in this manner was to identify patterns of relative over/underuse by each language group. See Table 2 for a list of all of the indices created during this process.

3.4 Evaluation of model

In CLI studies and other studies that attempt to predict the group membership of a text, discriminant function analysis (DFA) is often used (Jarvis & Paquot, 2012; Crossley & McNamara, 2012). Although other methods can be used, such as support vector machine decision trees (e.g., Koppel et al., 2005) or Naïve Bayes (e.g., Mayfield Tomokiyo & Jones, 2001), DFA has the advantage of being the most transparent of these with regard to interpreting results (Jarvis, 2012). DFA was therefore chosen as the method of analysis for this study, using L1 as the dependent variable and n-gram indices as independent variables.

The first step in the analysis was to check the independent variables for multicollinearity using a Pearson correlation matrix. Any two variables above a threshold of p>.899 were flagged for further analysis. A MANOVA was then conducted using the languages from one proficiency group as independent variables and the predictor indices/n-grams as dependent variables. The effect sizes produced by the MANOVA were used to select which variables flagged in the correlation matrix would be retained, and which would be eliminated. Within each highly correlated pair, the variable with the largest effect size was kept. Finally, a DFA was conducted on the training set. The predictor model sets identified in the DFA were then

		L1 Index Coverage			
Variable	Category	-	+	Total	Examples
ALL		11	11	22	see below
Adjectives	Syntactic	0	1	1	little, kind, real
Adverbs	Syntactic	0	2	4	always, easily just, still
Articles	Cohesion	8	8	16	a, an, the
Auxilliary Verbs	Syntactic	2	0	2	has, have, will
Certainty	Semantic	0	1	1	necessary, sure, true
Cognition	Semantic	0	1	1	experience, thought
Comparatives	Rhetorical	0	1	1	easier, much more
Conjunctions	Cohesion	6	5	11	and, because, or
Connectives	Cohesion	1	2	3	and to, and that, also
Determiners	Cohesion	1	0	1	that, this
Evaluation	Semantic	0	1	1	good, fun, like to
Examples	Semantic	0	1	1	particular, etc
Explanation	Semantic	0	4	4	explain, in order to, that is
Go	Semantic	0	1	1	are going, go, going to
Irrealis	Grammatical	0	1	1	what, will
Modality	Rhetorical	9	9	18	we can, could, can be
Negation	Syntactic	3	8	11	but not, no
Nouns	Syntactic	3	7	10	country, person, places
Options	Rhetorical	0	1	1	consider, different, instead
People	Semantic	1	4	5	people, society, friends
Place	Semantic	0	1	1	city, place, places
Possession	Semantic	1	1	2	his, having, your
Possibility	Rhetorical	0	3	3	probably, maybe, possible
Pre-infinitive	Syntactic	0	1	1	how to, time to, way to
Prepositions	Grammatical	10	9	19	from, about, with a
Problems	Semantic	1	1	2	problem, problems
Pronouns	Cohesion	10	11	21	he, his, your
Quantity	Semantic	11	11	22	every, more than, some
Questions	Syntactic	7	6	13	where, who, why, question
Science/ Technology	Semantic	0	2	2	computer, internet
Signifying	Rhetorical	0	1	1	see, mean
Specificity	Rhetorical	0	3	3	certain, especially, special
Stance	Rhetorical	2	6	8	feel that, in my, opinion
Temporality	Semantic	6	7	13	during, more and more, often
To Be	Syntactic	6	8	14	are, been, it is
Transitions	Cohesion	4	9	13	but, however, therefore
Vagueness	Semantic	0	1	1	general, someone, something
Verbs	Syntactic	5	8	13	choose, make, play
Work/Study	Semantic	2	7	9	money, study, parents
Total		110	167	277	

Table 2: Negative and positive key n-gram variables.

	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR	Precision	Recall	F-measure
ARA	66	0	5	3	1	3	2	4	8	1	7	53.2%	66.0%	58.9%
CHI	3	63	5	3	2	0	6	9	0	3	6	57.8%	63.0%	60.3%
FRE	3	4	64	7	3	6	2	1	6	0	4	64.6%	64.0%	64.3%
GER	2	5	5	64	3	5	2	4	6	0	4	62.7%	64.0%	63.4%
HIN	4	5	0	7	54	1	0	1	6	17	5	56.8%	54.0%	55.4%
ITA	4	1	9	10	1	64	2	1	6	0	2	68.8%	64.0%	66.3%
JPN	6	7	1	1	0	1	64	9	2	1	8	61.5%	64.0%	62.7%
KOR	5	9	2	1	2	0	19	56	2	0	4	57.7%	56.0%	56.9%
SPA	14	6	6	3	4	9	2	3	43	2	8	47.3%	43.0%	45.0%
TEL	5	3	0	1	22	1	1	1	4	60	2	70.6%	60.0%	64.9%
TUR	12	6	2	2	3	3	4	8	8	1	51	50.5%	51.0%	50.7%

Table 3: Test set confusion matrix.

used on the essays in the test set to determine whether the model sets could generalize to a new population.

4. Results

The training set DFA predicted L1 group membership of TOEFL independent essays with an accuracy of 60% using 184 indices (df= 100, n= 9900, χ^2= 32997.259, p< .001), which is significantly higher than the baseline chance of 9%. The reported Kappa = .560, indicates a moderate relationship between actual and predicted L1.

The predictive accuracy of the model was verified on the test set, in which L1 group membership was predicted with an accuracy of 59% (df= 100, n= 1100, χ^2= 3550.791, p< .001). The reported Kappa = .549, indicates a moderate agreement between the actual and predicted L1. Table 3 includes the test set confusion matrix.

5. Discussion

The results of this study suggest the usefulness of key n-grams grouped into categories based on their grammatical, rhetorical, semantic, syntactic, and cohesive features for NLI. The results demonstrate that such indices can correctly classify 59% of essays written in English as belonging to 1 of 11 L1 populations.

In addition, with regard to n-gram length, we found that although n-grams 1-10 words in length were initially considered, no n-grams longer than 5-grams were identified as key, and the longest n-gram that remained after removing prompt-based and redundancy was a single 4-gram. This suggests that 4-grams (or possibly even 3-grams) may be a useful threshold for future investigations.

5.1 Preliminary CLI findings

As Jarvis (2012) notes, CLI studies that use the detection-based argument to CLI are exploratory in nature, while studies that use the comparison-based argument are confirmatory in nature. The present study is, thus, exploratory in nature, and without substantial further investigation, we cannot definitively posit whether observed differences and similarities in English use can be attributed to the influence of the L1 itself or to cultural or educational norms.

Nonetheless, a few preliminary observations are worthy of discussion. First, we identified a number of patterns of language use that may be attributable to CLI. Although a full discussion of these is beyond the scope of this paper, Table 4 includes examples of potential CLI features in reference to the German writers represented in the corpus. The table demonstrates the particular n-grams that German writers are likely to use more or less often than writers of the other 10 languages. German writers, for example, are more likely to use the phrasal modals *able to*, *have to*, *has to*, and singular modals *might* and *would* more often than writers of the other language groups, but are less likely to use the modals *can* and *may*. These findings are preliminary, and further research that links these English n-grams with patterns of use in German is needed.

Additionally, our findings provide some evidence for close relationships between languages. For example, when checking for multicollinearity, we found that the All Negative Japanese and All Negative Korean categories were very strongly correlated (r =.946, p< .001). Upon further examination, 8 of the 19 n-grams (42%) in the All Negative Japanese category occurred in the corresponding Korean category. The overlapping n-grams were the n-grams *all*, *any*, *but*, *different*, *or*, *person*, *this*, and *your*, which may indicate similarities between these language systems in that speakers from both language avoid the use of these words.

Variable	Positive	Negative
Adverbs	just, only, there, necessary	
Comparatives	easier, much more	
Conjunctions	or, but, as well	
Modals	able to, have to, has to, might, would	can, may
Nouns	development, job, topic, something	person, place
Prepositions	at, on	about, by
Pronouns	everybody, this, you, your	she, its, I his, us, he, we, they, our
Quantity (and example)	another, amount of, both, less, lot, whole	any, many, some, such
Specific	certain, especially, special	
Stance	in my, of course, opinion, point	
Temporality	often, still	day, now, second, then, time to, second
To Be	be able, it is, to be	was
Transitions	furthermore, one hand, other hand	
Verbs	look, to get, work	go, going, study

Table 4: German predictor variables.

Patterns of essay categorization also provide preliminary insights into language similarities. Based on the test set confusion matrix (see Table 3), a few conflicting patterns emerged. Among the Indo-European languages represented, the Romance (French, Italian, and Spanish) and Germanic (German) languages were regularly miscategorized as one another. Italian essays, for example, were predicted to be French, German, and Spanish 9%, 10%, and 6% of the time, respectively, but were predicted to be other languages only 0%-4% of the time. This seems to confirm generally accepted language taxonomies, though Spanish was predicted to be Arabic (14%) and Turkish (8%) more often than Italian (6%) or French (6%) (as compared to 3% for German, and no more than 4% for other languages).

While similarities between language families seem to support extant language taxonomies (see Blanchard et al., 2013) and lend credence to claims of CLI, other observations may cast doubt on these. Hindi (an Indo-Iranian member of the Indo-European family) essays were predicted to be Telugu (Dravidian) essays 17% of the time, and Telugu essays were predicted to be Hindi essays 22% of the time. This may indicate instances of cultural proximity or educational similarities as opposed to linguistic transfer (and/or borrowing) because these languages are both spoken within India. Further investigations of these issues are clearly needed.

5.2 Limitations

While we have confidence in our findings, there are limitations to the analysis that need to be discussed. The TOEFL11 corpus was designed to be comparable across languages. While it largely accomplishes this goal, it is not well balanced across proficiency levels (which may reflect the relative proficiency levels of TOEFL test-takers). Although medium and high proficiency levels are well (though not equally) represented, the low proficiency group represents only 11% of the number of texts and an estimated 7.2% of total words (based on mean lengths of essays from each proficiency level given in Blanchard et al., 2013). The medium proficiency group represented 54.4% of the texts and an estimated 52.8% of words in the corpus, and the high proficiency group comprised the remaining 34.7% of the texts and an estimated 40% of the words. This indicates that caution should be used when generalizing any CLI findings from this study to low proficiency language users. Furthermore, any CLI findings will be biased towards medium proficiency language users.

Another limitation that may have affected the accuracy of the model was the way in which potential predictor variables were refined. For each language, the absolute keyness values were used when refining the lists of potential n-gram predictors (as discussed in Section 3.2). After the data had been processed, we discovered that this process removed some n-grams that should have remained. In a very few instances redundant n-grams (e.g., *have; have more)* had a positive keyness value for one n-gram (*have)* and a negative keyness value for the other (*have more*). Because all n-grams were later grouped into categories based on positive and negative keyness values, both *have* and *have more* should have been retained (as they would not have occurred in any of the same categories). In future studies, positive and negative n-grams will be kept separate during the elimination of redundant n-grams.

Another limitation that was discovered after the data analysis was that a data input error caused All Negative Chinese n-gram category to be combined with two n-grams included in the Positive Chinese School and Home category. A similar error retained two positive German adverb categories (with one overlapping n-gram, *just*). The models described in this study retained these variables, as they were not highly correlated with each other or any other variable (based on the r > .899 threshold), so any CLI findings based solely on these variables should be considered with caution.

5.3 Future research

Although it is clear that categorical n-grams can be used as successful NLI predictor variables, it is unclear whether this approach is more or less effective than the use of raw counts of frequent words or n-grams (e.g., Jarvis et al., 2012a; Jarvis & Paquot, 2012). Future research should explore the relative effectiveness of these methods using the TOEFL11 corpus to determine whether the time involved to create key n-gram lists and then sort those lists into categories is warranted.

Finally, another remaining question is whether the key n-grams identified in this study are due to linguistic factors or, alternatively, other influences such as culture and educational materials.

Acknowledgements

We thank ETS for compiling and providing the TOEFL11 corpus, and we also thank the organizers of the NLI Shared Task 2013.

References

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

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

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 and S. A. Crossley (Eds.), *Approaching Language Transfer through Text Classification: Explorations in the Detection-Based Approach*. (pp. 106-126). Bristol, UK: Multilingual Matters.

Crossley, S. A., Defore, C., Kyle, K., Dai, J., & McNamara, D. S. (under review). *Paragraph specific n-gram approaches to automatically assessing essay quality.* Sixth International Conference on Educational Data Mining, Memphis, TN.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193-202.

Granger, S., Dagneaux, E.., Meunier, F., Paquot, M. (Eds.) (2009). *International corpus of learner english. version 2.* Belgium: Presses universitaires de Louvain.

Granger, S. (2009) The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In: Aijmer, K., *Corpora and Language Teaching*, Benjamins: Amsterdam and Philadelphia, 2009, p. 13-32.

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

Jarvis, S. (2010). Comparison-based and detection-based approaches to transfer research. In L. Roberts,

M. Howard, M. Ó Laoire, & D. Singleton (Eds.), *EUROSLA Yearbook 10* (pp. 169-192). Amsterdam: Benjamins.

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* (pp. 1-33). Bristol, UK: Multilingual Matters.

Jarvis, S., & Crossley, S. A. (2012). *Approaching language transfer through text classification: Explorations in the detection-based approach*. Bristol, UK: Multilingual Matters.

Jarvis, S., Bestgen, Y., Crossley, S. A., Granger, S., Paquot, M., Thewissen, J., & McNamara, D. S. (2012). The comparative and combined contributions of n-grams, Coh-Metrix indices, and error types in the L1 classification of learner texts. In S. Jarvis & S.A. Crossley (Eds.), *Approaching language transfer through text classification: Explorations in the detection-based approach* (pp. 154-177). Bristol, UK: Multilingual Matters.

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* (pp. 71-105). Bristol, UK: Multilingual Matters.

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* (pp. 34-70). Bristol, UK: Multilingual Matters.

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* (pp. 624-628). Chicago: Association for Computing Machinery.

Kucera, H. and Francis, W. N. (1967). *Computational Analysis of Present-Day American English* Providence, RI: Brown University Press.

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics, 29*(4), 694-716.

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.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: method, theory and practice*. Cambridge, New York: Cambridge University Press, 2012.

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

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

Tetreault, J., Blanchard, D., & Cahill, A. (2013). Summary report on the first shared task on native language identification. In *Proceedings of the Eighth Workshop on Building Educational Applications Using NLP*, unpaginated electronic document. Atlanta, GA: Association for Computational Linguistics.

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.* (pp. 9-16). Cambridge, MA: The Association for Computational Linguistics.

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