Feature analysis for native language identification

Sergiu Nisioi[1,2]

[1] Center for Computational Linguistics,
University of Bucharest
[2] Oracle RightNow,
Bucharest, Romania
sergiu.nisioi@gmail.com

Abstract. In this study we investigate the role of different features for the task of native language identification. For this purpose, we compile a learner corpus based on a subset of the EF Cambridge Open Language Database - EFCAMDAT [10] developed at the University of Cambridge in collaboration with EF Education. The features we are taking into consideration include character n-grams, positional token frequencies, part of speech n-grams, function words, shell nouns and a set of annotated errors. Last but not least, we examine whether the essays of English learners that share the same mother tongue can be distinguished based on their country of origin.

1 Introduction

The concept of *interlanguage* first proposed by Selinker [27] proved to be essential in understanding the means through which adults acquire a second language. The term currently describes the entire linguistic system that emerges when second language learners - both child and adult - express meaning in the target language [26]. Interlanguage is usually regarded as a separate linguistic system that is different from the target language (TL) and the learner's mother tongue.

The main objective of native language identification (NLI) resides in the analysis and classification of texts that belong to specific groups of learners. Although the classes are usually determined by learners' mother tongues, in our study we label the documents based on the country from which the learners originate. This being the case, we pretermit the different dialects or minority languages that are spoken within a country.

Our study is focused on English texts belonging to learners originating from different geographic regions. Relying upon the existing psycholinguistic studies on the phenomenon of interlanguage, we first investigate and analyze the features that can be used for automatic text classification. In addition to the standard features used in literature [29, 32], we further suggest the use of shell nouns [25] as interlanguage markers. In our first set of experiments we train a classifier to identify the country of origin corresponding to each text. The results obtained,

further confirm previous NLI studies [16, 29, 31] and methodologies to automatically detect the native language of an individual based on his writing. In addition, in the later sets of experiments we construct a classifier to distinguish between English texts of learners sharing a similar or identical mother tongue but whose countries of birth are different. For example, learners from Spain and learners from Argentina may share the same native language (Spanish) but their linguistic backgrounds are different - cultural and social norms and possibly distinct curricula of learning English can contribute to the way learners acquire English.

We define the linguistic background of a learner as the entire set of linguistic input he was subjected to so far. The linguistic input can consist of previous languages learned (others than English, including the mother tongue), previous methodologies and curricula followed in order to acquire those languages and all the possible interactions the learner might have had with native English speakers or in native English communities. What is more, political and cultural factors can also act upon the linguistic background.

We hypothesize the existence of a thinner line of delimitation between different speakers of the same native language (such as Spanish for Colombia and Spain or German for Austria and Germany). In comparison, learners having different native languages (Korean and Japanese or Telugu and Hindi), belonging to related linguistic backgrounds are likely to go through similar developmental processes when acquiring a *foreign* language.

2 Previous work

One of the first multi-class native language identification studies [17] was conducted on the International Corpus of Learner English (ICLE) [11]. The different distribution and diversity of topics proved to be a disadvantage when evaluating cross-validated results [1]. TOEFL-11 is a learner corpus used for the 2013 NLI shared task [29], generally regarded as a better choice, the topics being similar and uniformly distributed across different learners. As Jarvis et al. [13] point out, the corpus lacks a uniform distribution of proficiency levels (low, medium, high) per native language, for example, only 1.4% of the texts coming from native German students are low proficiency texts.

A broad set of features and machine English teaching methods have been tested for the native language detection task [29, 30, 31]. Jarvis et al. [13] experiment with an L2-regularized L2-loss support vector machine [8] in combination with a mix of word n-grams, POS n-grams and lemma n-grams. In total they used around four hundred thousand features to achieve the best classification results for the NLI shared task [29]. Another approach that proved to output good results is based on a large spectrum of character n-grams. Ionescu et al. [12] combine a kernel machine with character n-grams to efficiently compute similarities when the space of features grows exponentially. Other high performance systems in the shared task [29] also used large numbers of character n-grams for classification. In cases like this, the features can cover topic related aspects

or even particular named entities (learners from Germany referring to German language, names or locations) and greatly outnumber the training/testing examples. Therefore, the results are usually difficult to interpret in terms of the psycholinguistic processes that shape the interlanguage of a learner.

3 Corpora

The corpus used in our study is based on a subset from the EF Cambridge Open Language Database - EFCAMDAT [10] developed at the University of Cambridge in collaboration with EF Education. The corpus consists from texts of various proficiency levels, submitted at Englishtown, the online school of EF Education [7].

The size of our extracted corpus has a total 18 million tokens and has essays collected from learners of 29 different countries: Argentina, Austria, Belgium, Brazil, Chile, People's Republic of China (PRC), Republic of China (Taiwan), Colombia, Costa Rica, Egypt, France, Germany, Indonesia, Italy, Japan, Kuwait, Mexico, Peru, Portugal, Russia, Saudi Arabia, South Korea, Spain, Switzerland, Thailand, Turkey, Ukraine, United Arab Emirates and Venezuela.

Out of the entire set of extracted texts, we have compiled the following corpora :

B13: learner texts from every unit in each level from one to six
- thirteen different countries: Brazil, Turkey, Italy, Mexico, People's Republic of China (PRC), France, Germany, Saudi Arabia, Colombia, Japan, Taiwan, Russia, South Korea
- for each country the sentences are merged and split into chunks of 1000 tokens
- each class is represented by the same number of chunks
- the total size of the corpus is approximately two million tokens

LB_Lang: groups of corpora to study the linguistic background hypothesis
- level one to six texts grouped from English learners that share similar mother tongues
- for each country the sentences are merged and split into chunks of 500 tokens
- each class is represented by the same number of chunks
- the term "Lang" is used to describe the native language
- (Table 1) contains the size and countries included for each language

For each level we have selected only learners who completed every unit to ensure as much as possible a uniform spread of topic and proficiency. The units found in level one are fairly basic and cover topics like greetings, family, jobs, describing people, food and drinks, etc. We could not control for a uniform distribution of levels across the documents from each country and we assume that a certain bias may have been introduced.

Among these topics, learners are required to describe facts about their place of birth which can reveal the first language of the learner.

Table 1: The corpora used to investigate the linguistic background hypothesis

Corpus	Countries	Total nr. of tokens
LB_Ar	Egypt, Kuwait, Saudi Arabia, United Arab Emirates (UAE)	174,000
LB_Ch1	People's Republic of China (PRC), Taiwan	1,000,000
LB_Ch2	Hong Kong (HK), PRC, Taiwan	132,000
LB_Ge	Austria, Germany, Switzerland	102,000
LB_RuUk	Russia, Ukraine	66,000
LB_Sp	Argentina, Colombia, Costa Rica, Mexico, Peru, Spain, Venezuela	199,500

We used the Stanford named entity recognition (NER) tool [9] trained on the CoNLL 2003 shared task data [24] to remove locations, person names, organizations and misc entities found in the texts. Moreover, we removed language names that were not identified by the NER system to avoid having biased classifications when using character n-grams. If a speaker claims to know Italian, then it's more likely for him to have a European mother tongue.

The corpus also comes with manual annotations of errors [10] together with the corrected alternatives. The most common type of errors encountered are the misuse of punctuation, capitalization and spelling errors. In total, there are 23 types of annotated errors [10], the least common are expressions of idioms, the use of possessive and the use of singular.

Each sub-corpus is extracted for distinct experiments, apart from this, it contains similar error annotations which allows us to have results consistent across various experiments. The extracted, pre-processed corpus is freely available at request from the author.

4 Features of interlanguage

Chomsky [2] traces the starting point of language learning to a simple, basic (universal) grammar to which all language users have access. Strong similarities were observed between the sequential process of acquiring a mother tongue and a second language: *pidgin*, *baby talk*, *simplified registers* [21]. There is also an important distinction between the two learning processes: children always succeed in completely acquiring their native language, but adults only very rarely succeed in completely acquiring a second language [28]. The notion of *fossilization* as defined by Selinker and Rutherford [26] designates the permanent cessation

of TL learning before the learner has attained the TL norms at all levels of linguistic structure.

Furthermore, the speed of learning a second language is highly correlated with the mother tongue [3]. The shared grammatical similarities between NL and TL can facilitate or impede the learning process. A so called positive language transfer phenomenon can intervene, *facilitating* a more rapid discovery of mother tongue-like features in the target language. A negative language transfer can also occur when a learner wrongly applies already acquired grammatical rules from his native language to express meanings in the TL. *Language transfer* is a key phenomenon that shapes the form of interlanguage.

In order to acquire a second language, Selinker [27] hypothesized that adults make use of a latent psychological structure, i.e. an already formulated arrangement in the brain, which is activated whenever an adult or child tries to produce meaning in the TL. The psycholinguistic processes of this structure together with brief examples are further provided:

native language transfer	NL-specific syntactic structures combined with target language words
overgeneralization / simplification	learners have the tendency to extensively use already acquired TL rules; for example, the use of past tense marker "-ed" for all verbs
transfer of training	e.g. fossilization can occur more rapidly for *street* learners compared to *classroom* learners [33], the former may successfully communicate to suit their needs albeit with lexical and syntactic errors
strategies of communication	resorting to more general nouns ("kind of", "sort", "thing") when the TL word is not known; the use of anaphoric shell nouns
strategies of learning	particular to learner: use of mnemonics, associations with cognates

5 Classification approach

5.1 Classification features

For classifying documents, we experiment with different features to cover every psycholinguistic aspect of interlanguage.

POS n-grams: part of speech bigrams and trigrams
character n-grams: bigrams, trigrams, 4-grams
function words: the closed class of words for English - connectors, determiners, particles, prepositions, adverbs, etc.

shell nouns: anaphoric nouns used to encapsulate more complex pieces of information [14, 25] - "fact", "thing", "task", "goal", "act", etc.

errors: manually annotated errors available in the EFCAMDAT corpus [10]

positional token frequencies: tokens appearing on the first two and last three positions in each sentence [34]

To cover the language transfer phenomenon, we consider function words and POS n-grams to be good features for two main reasons. (1) These types of features are (as much as possible) topic independent, unlike character n-grams or positional tokens. (2) Native language syntactic chunks have a tendency to transfer and influence the interlanguage. Function words reveal syntactic constructs, they are used unconsciously to tie sentences and create meaning. Hence, they were successfully used in a wide variety of text classification tasks from authorship attribution and analysis of style [5, 15], gender identification [19] to translation studies [34]. Münte et al. [22] argue that different brain functions are used to process the closed class and the open class of words.

POS n-grams are a type of shallow syntactic chunks that can be used as an indicator of the learner's coherence [4]. Apart from this, in combination with function words, POS n-grams can be used to reconstruct the phylogenetic tree of language similarities from learner texts [23] or to increase the accuracy of native language identification [13, 31].

Neither of these features are completely topic-independent, for example, literary and argumentative essays often employ different types of syntactic constructs that influence the way documents are classified, a fact observed [1] on the ICLE corpus as well.

Character n-grams have been successfully used for the task of NLI before, either in combination with words [13] or as standalone features in a kernel machine [12]. Widely used [20, 29], character n-grams have the advantage of covering language transfer and overgeneralization by encompassing both syntactic and morphologic features. The main drawback relies in the content it covers. Beside cultural particularities, learners often utter named entities like person names, organizations or locations which *betray* their actual native language. The features space usually grows to greatly outnumber the training/testing examples, making a feature selection process difficult, if not impossible.

We also investigate the importance of anaphoric shell nouns for the NLI task. Schmid [25] provides a list of shell nouns classified by six semantic classes: circumstantial, linguistic, modal, eventive, factual and mental. These features are quasi-topic-independent and we use them in combination with function words to increase the classification accuracy.

Positional frequencies [34] are obtained by counting the number of occurrences of tokens on the first, second and the last three positions. They can be an indicator that learners have certain ways of starting and ending a sentence (strategies of communication) which might be mother-tongue related.

5.2 Classifier

In our experiments, we use an L2-regularized L2-loss support vector classification machine with further parameter selection for C [8]. We adopt the log-entropy weighting scheme to construct feature vectors from documents. This weighting method also increased the classification accuracies in previous studies [6, 13].

The log-entropy weighting is frequently encountered in latent semantic indexing [18], its purpose is to reduce the importance of high frequency features, and increase the weight for the ones that are good discriminants between documents. We compute the entropy for a feature i by the following formula:

$$g_i = 1 + \sum_{j=1}^{\mathcal{N}} \frac{p_{ij} \log 1 + p_{ij}}{\log \mathcal{N}} \tag{1}$$

where $\mathcal{N}$ is the number of documents in the corpus and p_{ij} is defined by the normalized frequency of term i in document j.

To normalize the p_{ij} values, we divide by the global frequency in the corpus:

$$\mathrm{gf}_i = \sum_{j=1}^{\mathcal{N}} \mathrm{tf}_{ij}$$

in consequence, the value of p_{ij} becomes: $p_{ij} = \frac{\mathrm{tf}_{ij}}{\mathrm{gf}_i}$.

The final weight of a feature is computed by multiplying the entropy with the log weight:

$$logent_{ij} = g_i \log(\mathrm{tf}_{ij} + 1) \tag{2}$$

6 Results and interpretation

We have conducted multiple 10-fold cross-validation experiments corresponding to each combination of feature and corpus. The classifier was optimized with a search for the best parameter C which, in the majority of cases, attains low optimal values.

We have distinguished between topic sensitive/dependent features like character n-grams or positional token frequencies and topic independent features like function words, annotated errors or POS n-grams. (Table 3) contains the complete results for each set of experiments.

6.1 B13 corpus for native language identification

The B13 set of experiments represents the standard NLI task in which we evaluated a 13-class classifier to detect the native language/country of different English chunks.

On this sub-corpus, we obtained the best overall classification accuracy with character 4-grams (99.89%), closely followed by character trigrams.

Table 3: Average accuracy for each combination of feature and corpus. The highlighted values on each column represent the best scores for topic sensitive and independent features.

		Average accuracy for data set						
	Feature	**B13**	**LB_Ar**	**LB_Ch1**	**LB_Ch2**	**LB_Ge**	**LB_RuUk**	**LB_Sp**
independent	POS bigrams	75.43	67.04	88.10	70.18	76.58	90.22	67.25
	POS trigrams	87.20	75.07	92.35	83.01	**82.43**	**97.74**	72.25
	function words (FW)	86.06	**97.42**	99.5	94.71	60.48	83.45	**95.25**
	errors (E)	47.52	39.54	98.45	66.03	79.02	88.72	26.0
	FW and shell nouns	88.08	96.84	99.4	95.47	60.97	82.70	94.5
	FW + E + shell nouns	**93.75**	97.42	**99.85**	**96.22**	74.63	93.23	94.25
dependent	positional frequency	97.42	86.24	98.95	89.43	85.85	95.48	78.75
	char bigrams	94.47	93.98	98.75	88.3	83.41	93.98	86.25
	char trigrams	99.79	97.7	**99.9**	**97.35**	**90.24**	**97.74**	94.75
	char 4-grams	**99.89**	**99.14**	99.75	96.6	88.78	97.74	**96.25**

Among the features that are topic dependent, positional token frequencies achieved a reasonable accuracy of 97.42% which indicates a correlation between the nativeness (mother tongues) of individuals and the way they start or end sentences, i.e. some *strategies of communication* may be determined by the native language.

The best topic independent features were a combination of function words, errors and shell nouns which achieved an average cross-validation accuracy of (93.75%).

Table 4: Confusion matrix containing the rounded percentages of correctly classified B13 documents. A combination of function words, annotated errors and shell nouns were used as classification features.

	Brazil	Turkey	Italy	Mexico	PRC	France	Germany	S. Arabia	Colombia	Japan	Taiwan	Russia	S. Korea
Brazil	99	0	0	0	0	0	1	0	0	0	0	0	0
Turkey	1	96	0	2	1	0	0	0	0	0	0	0	0
Italy	0	0	83	1	0	7	1	1	3	2	2	0	1
Mexico	0	0	0	99	1	0	0	0	0	0	0	0	0
PRC	0	0	0	0	100	0	0	0	0	0	0	0	0
France	0	0	13	0	0	70	6	1	2	3	0	3	1
Germany	0	0	2	0	0	3	83	1	0	2	0	4	5
S. Arabia	1	0	0	0	0	0	0	96	0	0	1	3	0
Colombia	0	0	1	0	0	3	1	1	91	2	1	0	0
Japan	1	0	6	0	0	6	5	0	0	63	3	1	15
Taiwan	0	0	1	0	0	0	0	0	0	0	99	0	0
Russia	0	0	2	0	1	0	1	1	1	0	0	94	1
S. Korea	0	0	1	1	0	1	9	1	1	11	1	3	72

We regard the number of misclassified documents as a measure of similarity between two classes, (Table 4) contains the resulted confusion matrix for the B13 corpus. If native language relatedness can explain the 13% of the French documents misclassified as Italian, it cannot explain why native Mexican-Spanish

and Colombian-Spanish do not trigger any confusion at all. Learners from distinct families of languages (Japanese and Korean, French and German) coming from related geographic areas/linguistic backgrounds evidence more similarities through the percentage of classification confusion. Similar confusions were also observed at the NLI 2013 shared task in which pairs of non-related languages - Japanese-Korean and Telugu-Hindi [13] exhibit confusion because learners belong to related linguistic or cultural backgrounds.

6.2 Linguistic background analysis

We have experimented with texts coming from learners that share the same mother tongue including variations or dialects (Russian and Ukrainian). The LB_Lang columns reflect the classification accuracy for these sub corpora.

For the different varieties of Arabic spoken in Egypt, Kuwait, United Arab Emirates (UAE) and Saudi Arabia, we show that the classifier is able to distinguish between the English texts of these learners. Function words achieved the best accuracy for topic independent features (97.42%) whereas errors or shell nouns did not improve the classification results. Among the topic dependent features, positional token frequencies obtained the lowest accuracy (lower than function words), hence, in (Table 5) we render the resulted confusion matrix.

Table 5: Confusion matrix containing the rounded percentages of correctly classified L_Ar documents using positional token frequencies.

	Egypt	Kuwait	UAE	S. Arabia
Egypt	91	2	6	1
Kuwait	5	82	11	2
UAE	9	6	77	8
S. Arabia	1	0	3	95

Table 6: Confusion matrix containing the rounded percentages of correctly classified L_Ch2 documents using function words, errors and shell nouns.

	HK	Taiwan	PRC
HK	90	6	4
Taiwan	1	99	0
PRC	0	0	100

The confusion matrix in (Table 5) shows that a significant amount of learner English from Egypt was misclassified as United Arab Emirates and vice-versa. Documents from Kuwait are also frequently confused as being from UAE (11%) in contrast, Saudi Arabian English can be differentiated from the remaining texts - 95% correctly classified. Positional token frequencies cover similar types of starting and ending a sentence, a good classification result could indicate that the differences do not necessarily emerge due to specific language variations getting transfered onto English, but rather because of different strategies of teaching/learning English in these countries.

The LB_Ch1 corpus contains one million tokens equally extracted from learners from People's Republic of China (PRC) and Taiwan. Standard Chinese (the Mandarin dialect) is spoken in both countries, with the mention that in People's Republic of China at least 13 more major dialects exists specific for different

provinces. In (Table 3) we can observe that classifying English from Taiwan and English from PRC can be done with high accuracy values - using only function words, we get a 99.5% accuracy. Almost every type of feature (including errors) can act as excellent discriminants. For this particular instance we cannot be sure whether diverse dialects within PRC transfer to English yielding texts that are structurally different from the ones coming from Taiwan, or whether distinct learning methods are being used within the two countries.

The LB_Ch2 corpus is smaller including additional documents from English learners from Hong Kong. (Table 6) renders the confusion matrix which shows that only 10% of the learner English from Hong Kong is confused as being from Taiwan or PRC. The overall accuracy for topic dependent features is of 97.35% with character trigrams while function words, errors and shell nouns used together obtain a 96.22% accuracy.

Table 7: Confusion matrix containing the rounded percentages of correctly classified LB_Ge documents using POS trigrams.

	Switzerland	Austria	Germany
Switzerland	76	12	12
Austria	7	86	7
Germany	13	1	86

Table 8: First nine feature-selected character n-grams sorted by their corresponding F-score in the LB_Ge corpus.

trigram	**F-score**	**examples**
“hi ”	2.25	hi
“pu ”	1.93	punctuation error
“pe ”	0.54	type, hope
“oon”	0.31	soon, afternoon
“ope”	0.29	hope, open, opera
“tab”	0.25	table, vegetables,
“wn ”	0.24	down, brown, town,
“ af”	0.23	after, afternoon
“hit”	0.22	white

Lower accuracy values were obtained for countries in which German varieties are commonly spoken (Austria, Germany and Switzerland) - POS trigrams achieved 82.43% which is the best accuracy for the topic independent features. In (Table 7) we can observe the confusion matrix obtained with function words combined with errors: a significant number of documents are uniformly misclassified to each of the other countries.

Character trigrams - topic dependent features - attain the best overall accuracy value of 90.24%. Naturally, we are interested to observe which character n-grams increase the accuracy of the classifier. As a result, we investigate the n-grams with the highest F-score given the feature selection method proposed by Yi-Wei and Chih-Jen [35]. After extracting the most relevant trigrams for classification, we search for their occurrences in texts to find the most frequent examples. As (Table 8) indicates, the most discriminant trigrams cover topic independent features such as punctuation errors, function words (“soon”, “after”) and shell nouns (“type”). Yet, these features also cover content related words for example: “white”, “brown”, “afternoon”, “opera”, “table”, “brown”, etc. Under

these circumstances, character n-grams do not necessarily reveal only interlanguage markers, but also hidden content that is not uniform for different groups of learners.

Even though Russian and Ukrainian are considered dialects or separate languages, we investigated whether the classifier can distinguish between English essays written by natives of these countries. The penultimate column in (Table 4) surprisingly indicates that both topic dependent and independent features achieve similar classification accuracies (97.74%). As in the case of PRC versus Taiwan, learners could also be influenced by different varieties of languages spoken across Russia, a fact which can determine separate linguistic backgrounds.

Last but not least, we carried a 7-class classification of texts coming from different regions of the Spanish-speaking world (LB_Sp): Spain, Mexico, Costa Rica, Peru, Colombia, Venezuela and Argentina. Learner English from these countries can be classified with a 95.25% accuracy using only the list of function words while shell nouns or errors slightly decrease the value. (Table 9) contains the confusion matrix of the classification results using only the function words which obtained an overall accuracy of 95.25%. Character 4-grams can increase this accuracy with only 1%.

Table 9: Confusion matrix containing the rounded percentages of correctly classified LB_Sp documents using function words.

	Colombia	Mexico	Peru	Argentina	Venezuela	Costa Rica	Spain
Colombia	98	2	0	0	0	0	0
Mexico	0	100	0	0	0	0	0
Peru	0	0	100	0	0	0	0
Argentina	0	0	0	96	2	2	0
Venezuela	0	0	0	4	93	4	0
Costa Rica	0	0	0	9	11	79	2
Spain	0	0	0	0	0	0	100

English texts from Argentina, Colombia, Mexico, Peru and Spain can be distinguished almost perfectly from the rest while Costa Rica and Venezuela share the largest amounts of classification confusion using function words (topic independent features).

These results indicate that students from each country go through similar stages of learning English and possibly any foreign language. For example, the learners from Mexico may be influenced by linguistic and political factors (USA being a neighboring country) so that they achieve good proficiency levels at earlier stages of learning English, compared to students from other countries which experience less interaction with native English communities. Our investigation does not account the different grades students had for the Cambridge examination which, we assume, might also be a factor of influence. Furthermore, the distributions of different levels across the corpus can also be a factor of influence and more work is prepared in this direction.

7 Conclusions

In this paper we provide an analysis of the linguistic features that are suitable for the task of native language identification. We research our claims on a subset of the EFCAMDAT corpus [10] from which named entities and references to language names were removed. In addition to the standard classification features used in literature such as character n-grams, part of speech n-grams, function words or annotated errors, we further prove that anaphoric shell nouns and positional token frequencies represent interlanguage markers that contribute to the overall classification accuracies. Our results also suggest that topic sensitive features tend to obtain the best results across different corpora. However, we recommend additional care when employing these features since texts may contain hidden topics that can determine misleading classifications.

Our data includes error annotated documents from different countries in which the same language is spoken by a majority. Apart from this, the corpus is compiled from medium-low proficiency English texts that exhibit a significant amount of errors and interlanguage features, therefore, facilitating the classification tasks.

The novelty of our study does not only rely on the experimental analysis of interlanguage features but also on the investigation of the inner dissimilarities within a group of learners that share the same mother tongue. To explain the differences that appear between learners with distinct native countries and similar native languages, we conjecture the existence of a linguistic background which can be determined by the previous languages learned and possibly cultural and political factors. The linguistic background interacts with the process of learning, complementary to the learner's native language.

On one hand, language relatedness can explain the classification confusions that emerge between similar languages e.g. French and Italian. On the other hand, this phenomenon cannot explain why Spanish from Mexico and Spanish from Colombia do not trigger confusion or why learners from distinct families of languages (Japanese and Korean, French and German, Telugu and Hindi [13]) coming from neighboring geographic areas evidence more similarities through the percentage of misclassified documents.

We are inclined to believe these similarities fade as the learner proficiency increases, but the corpus required to investigate this hypothesis is not available yet and its development is part of our current and future work. Our results trace the existence of a linguistic background. Nevertheless, a more thorough investigation would be necessary to fully analyze and understand the roots of this phenomenon.

Acknowledgments: I would like to address special thanks to Anca Bucur for her helpful suggestions and support in improving this paper. Needless to say, any remaining errors are mine alone.

Bibliography

[1] Brooke, J., Hirst, G.: Native language detection with 'cheap' learner corpora. In: Conference of Learner Corpus Research (LCR2011). Presses universitaires de Louvain, Louvain-la-Neuve, Belgium (2011)

[2] Chomsky, N.A.: Linguistics and philosophy. In: Hook, S. (ed.) Language and Philosophy. New York University Press (1969)

[3] Corder, S.P.: Language distance and the magnitude of the language learning task. Studies in Second Language Acquisition 2, 27–36 (9 1979)

[4] Dinu, A.: On classifying coherent/incoherent romanian short texts. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco (may 2008)

[5] Dinu, L.P., Niculae, V., Şulea, O.M.: Pastiche detection based on stopword rankings: exposing impersonators of a romanian writer. In: Proceedings of the Workshop on Computational Approaches to Deception Detection. pp. 72–77. EACL 2012, Association for Computational Linguistics, Stroudsburg, PA, USA (2012)

[6] Dumais, S.: Improving the retrieval of information from external sources. Behavior Research Methods, Instruments and Computers 23(2), 229–236 (1991)

[7] Englishtown: Education first. `http://www.englishtown.com/` (2012)

[8] Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: Liblinear: A library for large linear classification. J. Mach. Learn. Res. 9, 1871–1874 (Jun 2008)

[9] Finkel, J.R., Grenager, T., Manning, C.D.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Knight, K., Ng, H.T., Oflazer, K. (eds.) ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA. The Association for Computer Linguistics (2005)

[10] Geertzen, J., Alexopoulou, T., Korhonen, A.: Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). Proceedings of the 31st Second Language Research Forum (SLRF), Cascadilla Press, MA (2013)

[11] Granger, S., Dagneaux, E., Meunier, F.: The International Corpus of Learner English: Handbook and CD-ROM, version 2. Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium (2009)

[12] Ionescu, T.R., Popescu, M., Cahill, A.: Can characters reveal your native language? a language-independent approach to native language identification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1363–1373. Association for Computational Linguistics (2014)

[13] Jarvis, S., Bestgen, Y., Pepper, S.: Maximizing classification accuracy in native language identification. In: Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 111–118. Association for Computational Linguistics, Atlanta, Georgia (June 2013)
[14] Kolhatkar, V., Zinsmeister, H., Hirst, G.: Interpreting anaphoric shell nouns using antecedents of cataphoric shell nouns as training data. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 300–310. Association for Computational Linguistics (2013)
[15] Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. J. Am. Soc. Inf. Sci. Technol. 60(1), 9–26 (Jan 2009)
[16] Koppel, M., Schler, J., Zigdon, K.: Automatically determining an anonymous author's native language. Intelligence and Security Informatics pp. 41–76 (2005)
[17] Koppel, M., Schler, J., Zigdon, K.: Determining an author's native language by mining a text for errors. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. pp. 624–628. ACM, Chicago, IL (2005)
[18] Landauer, T., McNamara, D., Dennis, S., Kintsch, W.: Handbook of Latent Semantic Analysis. Taylor and Francis (2013)
[19] Lim, C., Lee, K., Kim, G.: Multiple sets of features for automatic genre classification of web documents. Information Processing and Management 41(5), 1263–1276 (2005)
[20] Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C.: Text classification using string kernels. J. Mach. Learn. Res. 2, 419–444 (Mar 2002)
[21] Long, M.H.: Maturational constraints on language development. Studies in Second Language Acquisition 12, 251–285 (9 1990)
[22] Münte, T.F., Wieringa, B.M., Weyerts, H., Szentkuti, A., Matzke, M., Johannes, S.: Differences in brain potentials to open and closed class words: class and frequency effects. Neuropsychologia 39(1), 91 – 102 (2001)
[23] Nagata, R., Whittaker, E.W.D.: Reconstructing an indo-european family tree from non-native english texts. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers. pp. 1137–1147 (2013)
[24] Sang, E.F.T.K., Meulder, F.D.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003. pp. 142–147 (2003)
[25] Schmid, H.U.: English Abstract Nouns As Conceptual Shells: From Corpus to Cognition. Topics in English Linguistics 34, De Gruyter Mouton, Berlin (2000)
[26] Selinker, L., Rutherford, W.: Rediscovering Interlanguage. Applied Linguistics and Language Study, Routledge (2014)
[27] Selinker, L.: Interlanguage. International Review of Applied Linguistics in Language Teaching 10(1–4), 209–232 (1972)

[28] Tarone, E.: Interlanguage. Blackwell Publishing Ltd (2012)
[29] Tetreault, J., Blanchard, D., Cahill, A.: 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. Association for Computational Linguistics, Atlanta, GA, USA (June 2013)
[30] Tetreault, J., Blanchard, D., Cahill, A., Chodorow, M.: Native tongues, lost and found: Resources and empirical evaluations in native language identification. In: Proceedings of COLING 2012. pp. 2585–2602. The COLING 2012 Organizing Committee, Mumbai, India (December 2012)
[31] Tsur, O., Rappoport, A.: 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. Association for Computational Linguistics, Prague, Czech Republic (June 2007)
[32] Tsvetkov, Y., Twitto, N., Schneider, N., Ordan, N., Faruqui, M., Chahuneau, V., Wintner, S., Dyer, C.: Identifying the l1 of non-native writers: the cmu-haifa system. In: Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 279–287. Association for Computational Linguistics, Atlanta, Georgia (June 2013)
[33] Valette, R.M.: Proficiency and the prevention of fossilization an editorial. The Modern Language Journal 75(3), 325–328 (1991)
[34] Volansky, V., Ordan, N., Wintner, S.: On the features of translationese. Literary and Linguistic Computing (2013)
[35] Yi-Wei, C., Chih-Jen, L.: Combining svms with various feature selection strategies. Feature Extraction 207, 315–324 (2006)