Exploiting native language interference for native language identification

Reference:
Markov Ilia, Nastase Vivi, Strapparava Carlo.- Exploiting native language interference for native language identification
Natural language engineering - ISSN 1469-8110 - 28:2(2022), PII S1351324920000595
Full text (Publisher's DOI): https://doi.org/10.1017/S1351324920000595
To cite this reference: https://hdl.handle.net/10067/1748210151162165141

Exploiting Native Language Interference for Native Language Identification

Ilia Markov[1], Vivi Nastase[2], and Carlo Strapparava[3]

[1] University of Antwerp, CLiPS, Antwerp, Belgium
ilia.markov@uantwerpen.be
[2] University of Stuttgart, Stuttgart, Germany
vivi.nastase@ims.uni-stuttgart.de
[3] FBK-irst, Fondazione Bruno Kessler, Trento, Italy
strappa@fbk.eu

Abstract Native Language Identification (NLI) – the task of automatically identifying the native language (L1) of persons based on their writings in the second language (L2) – is based on the hypothesis that characteristics of L1 will surface and interfere in the production of texts in L2 to the extent that L1 is identifiable. We present an in-depth investigation of features that model a variety of linguistic phenomena potentially involved in native language interference in the context of the NLI task: the languages' structuring of information through punctuation usage, emotion expression in language, similarities of form with the L1 vocabulary through the use of anglicized words, cognates and other misspellings. The results of experiments with different combinations of features in a variety of settings, allow us to quantify the native language interference value of these linguistic phenomena, and show how robust they are in cross-corpus experiments and with respect to proficiency in L2. These experiments provide a deeper insight into the NLI task, showing how native language interference explains the gap between baseline, corpus-independent features and the state-of-the-art that relies on features/representations that cover (indiscriminately) a variety of linguistic phenomena.

1 Introduction

In computational linguistics, Native Language Identification (NLI) is the task of identifying the native language (L1) of persons based on their writings in the second language (L2). NLI is based on the assumption that characteristics of a person's native language interfere in the production of texts in a different language, to the extent that L1 is identifiable. Often, as for security applications, identifying the L1s is the main focus – finding the set of features and the learning paradigm that lead to the best L1 predictions. Other applications, such as marketing and educational materials tuned to different L1s, would benefit from information about what L1 phenomena got transferred to L2 (native language interference [40]).

Word and character n-grams are considered the best-performing features for NLI (see Section 2). They cover indiscriminately a multitude of linguistic particularities.

Using such document representations do not provide any insight into the linguistic phenomena that a speaker subconsciously transfers from his/her native language to the new language.

The focus of this article is an investigation of several language characteristics that we hypothesize play a part in native language interference: (i) the structuring of information in written language through the use of punctuation; (ii) emotionally charged words usage; (iii) incorrect expansion of L2's vocabulary using L1 material through anglicized words, cognates, and other misspellings. The analysis confirms our hypotheses, and leads to several strong conclusions: (i) the structuring of written language through punctuation is a strong bias, which leaves L1 traces even for speakers at high proficiency levels, across different corpora and topics; (ii) the usage of emotionally charged words is culture dependent, and has a strong signature in L2 production including at high proficiency levels, and across different corpora and topics. Other kind of interference such as anglicized words, cognates, and spelling errors in L2 have a positive impact on identifying the L1 of the speaker.

Section 2 presents an overview of work on Native Language Identification. Section 3 describes the datasets, features, and experimental setup used in this work. Sections 4 and 5 describe the experiments, their results and analysis, and Section 6 draws the conclusions.

2 Related Work

Native language identification is focused on identifying the native language of a speaker based on the language material they produce in a different language, and is commonly approached as a multi-class classification problem: for a given text sample, predict the native language (L1) of the writer from a (small) set of options [51]. Word and character n-grams were proved to be the best features. Jarvis et al. [21] achieved the highest classification accuracy (83.6%) in the first NLI Shared Task 2013 [51] using lexical and part-of-speech (POS) n-gram features on the 11-way TOEFL11 dataset [3] composed of English essays. The authors report that a model based on character n-gram features produced nearly the same high level of NLI accuracy. This observation was confirmed by Markov et al. [31], one of the best performing systems in the recent NLI shared task 2017 [29]. Several other studies, e.g., [19, 20], achieved state-of-the-art results using character n-gram models with high values of n (up to n = 9). The state-of-the-art results for this task are usually in the 80%–90% accuracy range, depending on the number of languages being considered, amount of data, etc.

On social media data, Kumar et al. [24, 25] summarized two shared tasks on Indian NLI, both on identifying six Indian languages from Facebook English comments. The highest results in 2017 (48.8% accuracy) were achieved using character and word n-gram features, while in 2018 the best score (37.0% accuracy) was achieved using the most discriminative words for each of the six languages. Volkova et al. [54] addressed 12-way NLI task of non-English speakers based on their English posts on Twitter. The best result was obtained using words and word 3-grams (72% F1-score for both words and word 3-grams). Goldin et al. [14] identified 23 European L1s of advanced non-native English speakers from Reddit posts. The best textual features (i.e., excluding social network features) were character 3-grams (62.06% accuracy) and words (31.26%).

Word and character n-grams cover a wide range of phenomena and facilitate identifying this L1, but they obscure which ones are actually "caused" by the speaker's L1. Our goal in this study was to provide an in-depth analysis using features that explicitly model specific linguistic phenomena that potentially causes interference, and quantify their contribution to the NLI task. We essentially investigate the results gap between a configuration that uses baseline (corpus-independent) features, and the state-of-the-art, and try to see how this gap (and how much of it) is filled by the linguistic phenomena – use of punctuation, use of emotion-expressing words, L1-based expansion of L2-vocabulary – we identified.

Punctuation In high-level Natural Language Processing (NLP) tasks (e.g., information extraction, text categorization), punctuation is often disregarded or discarded. Punctuation is considered a stylistic choice, and an important feature in stylometric analysis for authorship attribution, e.g., [8, 17, 33].

While punctuation has been included in some of the studies on NLI (e.g., in character-level models), its impact has not been studied. It is however an important, and often revealing, aspect of written language. From a linguistic point of view, punctuation has been disputed as following prosodic principles or as a clarifier of grammatical structure [1, 6]. Moore [37] finds a common ground for these two views by observing that prosody and punctuation realize the same function – revealing/emphasizing the information structure of an utterance – in the spoken and respectively written modes of language. Since grammar and prosodic structure are language specific, indicators that reveal them would be language specific as well. As with other aspects of language, grammatical/prosodic influences from the native language may surface in the new language as particular punctuation choices.

Emotions Communicating emotions in L2 is not easy, and it involves important aspects of socio-pragmatic competence. Caldwell-Harris [7] shows that usage of emotionally charged words depends on the language. By focusing on differences between L1 and L2, she shows that there is a correlation between the usage of emotions and proficiency levels and the age a language is acquired.

Emotion-based features have been used in other NLP tasks, such as sentiment analysis [48], aggressiveness detection [15], identification of deceptive texts [38], etc. With respect to second language writing, they are underexplored. Torney et al. [53] use psycholinguistic features extracted by the Linguistic Inquiry and Word Count (LIWC) tool [42] to identify the first language of an author. Emotion-based features are included as part of the feature vector, e.g., percentage of positive/negative emotion words. The LIWC feature set used in the paper also contains other types of features, e.g., personal concern categories (work, leisure), paralinguistic dimensions (assents, fillers, nonfluencies), which obscure the contribution of the actual emotion features.

Rangel and Rosso [44, 45] confirm the hypothesis that the use of emotions depends on the author's age and gender using a graph-based approach, where each node is represented by the corresponding POS tag, and the representation is enriched with semantic information, emoticons, and with emotion information, which included polarity of words (polarity of common nouns, adjectives, adverbs or verbs in a sentiment lexicon) and emotionally charged words (replacing common nouns, adjectives, adverbs or verbs with the emotion information from the Spanish Emotion Lexicon [48]).

We examine the hypothesis that there are commonalities in the use of emotions in L2 by different writers with the same L1s, suggested by linguistic and psycholinguistic studies [26, 55].

Cognates and anglicized words True cognates (or true friends) are words in different languages that are translations and have a similar orthography: e.g., Spanish *conclusión* and English *conclusion*. False cognates (or false friends) are words in different languages with similar orthography that are not translations: e.g., Spanish *embarazada* (pregnant) and English *embarrassed*.[4]

Nicolai et al. [39] tested two hypotheses: (i) cognate interference may cause an L1-speaker to use a cognate word instead of a correct English translation; (ii) an intended English word may be misspelled due to the influence of the L1 spelling. They detect cognates by identifying words that are closer to the misspelling than to the intended word. Despite being applied to only 4 out of 11 languages in the TOEFL11 dataset, cognate features led to 4% fewer errors.

Rabinovich et al. [43] showed that non-native speakers, when required to pick an English word that has a set of synonyms, are more likely to select a word that has a cognate in their L1. Based only on the frequencies of specific words in English, Rabinovich et al. [43] were able to reconstruct a phylogenetic tree of the Indo-European language family.

L2-ed words (in our case anglicized) are similar to false cognates, but not in the sense of false friends: these are words in L1 that were “adjusted” to look and sound like legitimate L2 words: the incorrectly anglicized word *facily* instead of *easily* (Spa. fácilmente). This phenomenon is similar to code switching, which focuses on the mixing of languages within one text, e.g., [50], where the change from one language to another occurs at the word level. The particularity of L2-ing is that the switching/mixing occurs within words, at the morpheme or character level.

Spelling errors Spelling errors in L2 are sometimes caused by erroneous sound-to-letter mapping. Koppel et al. [23] suggested that spelling errors are indicative features for NLI as writers might be affected by the spelling convention in their native languages. They explored eight types of spelling errors and collect the statistics of each error type as features.

Nicolai et al. [39] extract spelling error features from character-level alignments between a misspelled word and the intended word. They report that the spelling error features contribute 0.4% accuracy when combined with other commonly used features on the TOEFL11 test set.

Chen et al. [10] extract character n-gram features from misspelled words. They show that adding spelling error character n-grams to other commonly used NLI features (word, lemma, and character n-grams) improves NLI accuracy by 1.2% on the TOEFL11 test set.

[4] In this work we do not distinguish between true cognates and false friends, so when we refer to cognates in the related literature or in our own work, we mean both.

Flanagan and Hirokawa [12] classified five L1s from the lang-8 dataset (Japanese, Chinese, Korean, Taiwanese, and Spanish) using 15 automatically identified types of writing errors, achieving higher results than when using unbiased words.

Goldin et al. [14] used the average Levenshtein distance between a misspelled word and the corrected version, as well as the insertions, deletions and substitutions operations as features. The spelling error features were among the best feature types with 27.74% accuracy on a 23-way NLI task.

3 Methodology

To test the hypotheses that punctuation, emotion and specific misspellings interfere with language production in a second language, we use two datasets of English essays written by non-native English speakers. These essays will be represented through various sets of features that are designed to capture the targeted linguistic phenomena, and then used in multi-class classification experiments.

3.1 Datasets

We conducted experiments on two datasets commonly used in NLI research: TOEFL11 [3] and ICLE*v*2 [16]. Both datasets cover English L2 data and represent learner corpora.

TOEFL11 dataset The ETS Corpus of Non-Native Written English (TOEFL11) [3] was designed specifically to support the NLI task and has become a standard frame of reference for NLI research. TOEFL11 contains 1,100 essays in English (avg. 348 tokens/essay) for each of the following 11 native languages: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, and Turkish.

The essays were written in response to eight different writing prompts/topics (P0–P7), all of which appear in all 11 L1 groups. The topics include old vs. young people comparison, education, advertisement, car usage, travel, and risk management (statistics in Table 1). Within four of the L1 groups (Arabic, Chinese, Japanese, Korean), all prompts are approximately equally represented (12.5% per prompt); in other groups, there is more variability. The Italian group shows the largest variations: P1 representing 1.1% of the essays, while P0 and P5 each cover 17.0%.

The proficiency level of the author of each essay (low, medium or high) is provided as metadata (statistics in Table 2). The distribution of learners' proficiency levels (determined by assessment specialists) is more variable across groups than the writing prompts. The distribution is especially sparse for the German speakers (1.4% participants have low-proficiency, 61.2% have high-proficiency). Overall, the low-proficiency category represents only 11.0% of the essays, while the medium-proficiency category comprised 54.3% and the high-proficiency category represented the remaining 34.7%.

For the evaluation of cognates, L2-ed (anglicized) words, and all other misspellings, we used a 4-language subset of the corpus, focusing on the Indo-European languages that use the Latin script: French, German, Italian, and Spanish. This subset, to which we refer as TOEFL4, contains 1,100 essays (avg. 353 tokens/essay) for each of the four languages.

Table 1. Distribution of topics in TOEFL11.

	Number of essays per prompt							
L1	**P0**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**
Arabic	139	133	141	138	138	136	138	137
Chinese	140	141	139	139	140	134	126	141
French	156	68	160	151	158	160	87	160
German	151	28	153	152	155	150	157	154
Hindi	86	53	161	158	161	156	163	162
Italian	187	12	141	173	173	187	138	89
Japanese	138	142	143	141	116	138	140	142
Korean	128	142	143	141	140	137	136	133
Spanish	159	157	162	160	141	134	54	133
Telugu	55	41	171	166	165	169	167	166
Turkish	170	43	169	167	169	147	90	145
Total	1,509	960	1,683	1,686	1,656	1,648	1,396	1,562

Table 2. Data statistics for the three English proficiency levels in TOEFL11.

	English proficiency					
L1	**Low**		**Medium**		**High**	
Arabic	296	26.9%	605	55.0%	199	18.1%
Chinese	98	8.9%	727	66.1%	275	25.0%
French	63	5.7%	577	52.5%	460	41.8%
German	15	1.4%	412	37.5%	673	61.2%
Hindi	29	2.6%	429	39.0%	642	58.4%
Italian	164	14.9%	623	56.6%	313	28.5%
Japanese	233	21.2%	679	61.7%	188	17.1%
Korean	169	15.4%	678	61.6%	253	23.0%
Spanish	79	7.2%	563	51.2%	458	41.6%
Telugu	94	8.5%	659	59.9%	347	31.5%
Turkish	90	8.2%	616	56.0%	394	35.8%
Total	1,330	11.0%	6,568	54.3%	4,202	34.7%

ICLE dataset The ICLE*v*2 dataset [16] consists of essays written by highly-proficient non-native college-level students of English. The 6,085 essays, with length between 500 and 1,000 words, were written in response to 1,302 prompts, manually grouped into 736 topics. The essays are of two types: literature critiques or argumentative essays. The dataset covers 16 L1s: Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish, and Tswana.

ICLE was originally collected for the purpose of corpus linguistic research, and thus has certain limitations when used for NLI research [4, 52]: (i) several topics contain essays by learners from a single L1 [4]; (ii) there are several encoding errors and annotation issues, which only occur in essays written by learners from certain L1s [52]. Because of these shortcomings, and following previous work [20,52], we use a 7-language subset of the corpus normalized for topic bias and character encoding (the so-called ICLE-NLI subset [52]). This subset contains 110 essays for the 7 first languages shown in Table 3. The average number of tokens per essay is 747 (after tokenization and removal of metadata). In the paper, we refer to this 7-way subset as ICLE unless stated otherwise.

Table 3. Number of essays per class in the 7-way ICLE dataset (the ICLE-NLI subset).

Language	No. of essays
Bulgarian	110
Chinese	110
Czech	110
French	110
Japanese	110
Russian	110
Spanish	110
Total	770

For cognates, L2-ed words and other misspellings evaluation we used a 4-language subset of the corpus that represents the same languages as those included in TOEFL4: French (347 essays), German (437), Italian (392), and Spanish (251). This subset, which we refer to as ICLE4, contains 1,427 essays with avg. 690 tokens/essay.

3.2 Features

Part-of-speech (POS) tag n-grams and function words (FWs) are considered core features in NLI research [28] and not influenced by topic bias. Because of this they do not capture very specific document characteristics, and their capacity to identify the writer's L1 is limited. At the other end of the spectrum are representations based on word and character n-grams, that perform best on the NLI task, but their performance also exploits the fact that they capture topic and corpus-specific characteristics [4].

We use POS n-grams and function words as the baseline features to represent the documents (essays) in our data. For each of the phenomena we investigate – punctuation usage, emotionally-charged words, vocabulary expansion – we design a set of features.

These feature sets, which are explained below, will be added to the baseline features in a variety of experiments designed to analyze their impact and robustness. Table 4 shows some examples.

Table 4. Example for the features used. PH = place holder for punctuation marks (PMs).

phrase	*She said , “ This is very good ! ”*
POS unigrams	PRP VBD DT VBZ RB JJ
POS & FW features	She VBD This is very JJ
POS & FW with PHs	She VBD PH PH This is very JJ PH PH
POS & FW with PMs	She VBD , “ This is very JJ ! ”
POS & FW with emotion	She VBD This is very JJ-0011101001

Ultimately, we check how much of the gap between the baseline and the state-of-the-art is filled in by modeling the phenomena we identified.

Part-of-speech (POS) tags POS features capture the morpho-syntactic patterns in a text. POS tags were obtained with TreeTagger [47], which uses the Penn Treebank tagset (36 tags). Table 4 shows an example using POS unigrams (without punctuation marks (PMs)). From these, POS n-grams (n = 1–3) are built.

Function words (FWs) Function words clarify the relationships between the content-carrying elements of a sentence, and introduce syntactic structures like verbal complements, relative clauses, and questions [49]. They are considered one of the most important stylometric features [22]. The FW feature set consists of 318 English FWs from the scikit-learn package [41]. Since FWs are tied to the structuring of information, function words and punctuation marks fill roles within the same domain, and could provide complementary features with respect to the NLI task. With respect to emotion features, function words can appear as quantifiers, intensifiers (e.g., *very good*) or modify the emotion expressed in other ways. Table 4 shows an example for POS & FW features, from which n-grams (n = 1–3) are extracted.

Punctuation marks (PMs) Punctuation serves to clarify the message conveyed by an utterance, by introducing structure in the linear expression through grouping, delineating and emphasizing units of information. This structuring being language dependent, it has the potential to interfere in L2 production, as exemplified by the following English sentence written by a German native speaker:

I think the biggest question is , how to defin an “ enjoyed life.[5]

[5] Extracted from one of the training essays in the data (NLI: 10086.txt), with the author’s punctuation choices and misspelling of *define*.

A native English speaker would not insert a comma between *is* and *how*, but this reflects punctuation usage in German:

Ich denke, die größte Frage ist, wie man ein "glückliches Leben" definiert.

We evaluate the impact of PMs by adding them to POS and POS & FW representations. We evaluate first the impact of the presence of punctuation marks using a place holder (PH) for PMs, and then the actual PMs. Examples of these two representation variations are shown in Table 4.

Emotion polarity features (emoP) We encode emotion information using the information from the NRC Word-Emotion Association Lexicon (NRC emotion lexicon) [36], where binary associations are provided for each emotion word for 8 emotions (anger, fear, anticipation, trust, surprise, sadness, joy, or disgust) and two sentiments (negative or positive) – e.g., *good* = "0011101001". This representation is used as a categorial feature (not a 10-dimensional binary vector). It performed best compared to other ways of encoding the emotion information, e.g., using a 10-dimensional binary vector or excluding the sentiment information.

Table 4 shows an example representation when emotion polarity features are added to the POS & FW representation (without PMs).

Emotion load features (emoL) Speakers of different L1s may use a higher or lower number of emotionally charged words than speakers of other L1s, reflecting cultural customs or linguistic habits of the respective cultures. We modeled this information using three types of emotion load features:

1. two binary features, *emoL* (binary) that capture whether an essay has a high or low emotional load: (a) we compute the average proportion of emotion words in all essays in each dataset: for TOEFL11 this was 0.236 and for ICLE 0.246; (b) if the proportion of emotion words in an essay was higher/lower than the average,it gets assigned a "highly-emotional"/"low-emotional" feature respectively. This representation was used to examine whether the polarity as such is informative.
2. the *proportion of emotion words* in each essay as a numeric feature (1 feature, *emoL* (1)),
3. the *proportion of each emotion/sentiment* in each essay (8 emotions and 2 sentiments, *emoL* (10)).

Misspelled cognates Mann and Yarowsky [30] Bergsma and Kondrak [2] and Nicolai et al. [39] used string similarity for cognate identification. Following Nicolai et al. [39], we detect cognates by identifying misspelled words w_m, whose closest correctly spelled L2 word w_e has a translation w_f in an L1, and w_m is closer in form to w_f than to w_e. Formally:

1. For each misspelled English word w_m identify the intended word w_e using a spell-checking tool.[6]
2. For each L1:
   (a) Look up the translation w_f of the intended word w_e in L1.[7]
   (b) Replace diacritics in w_f with the corresponding Latin equivalent (e.g., “é” $\rightarrow$ “e”).
   (c) Compute the Levenshtein distance D between w_e and w_f.
   (d) If $D(w_e, w_f) < 3$ then w_f is assumed to be a cognate of w_e.[8]
   (e) If w_f is a cognate and $D(w_m, w_f) < D(w_e, w_f)$ then consider the L1 as a clue of the native language of the author.[9]

L2-ed words To identify L2-ed (in our case anglicized) words we take a misspelled word w_m and look for forms close to it in the L1 vocabularies. The idea is that a misspelled word may be an L1 word that got anglicized, which is a clue for the L1 of the author.

As reference vocabularies for the L1 languages in our data sets, we use the freely available lists of expressions provided by the OmegaWiki project[10] and extract the unigrams. The statistics for each language in terms of the number of expressions and the extracted vocabularies is provided in Table 5.

Table 5. Statistics of the number of expressions and the extracted vocabularies for each of the languages.

Language	**No. of expressions**	**No. of unique words (vocabulary)**
French	32,184	21,433
German	31,450	28,378
Italian	26,764	18,561
Spanish	39,566	27,321

We apply the following algorithm:

1. For each misspelled English word w_m identify its closest word in some L1:
2. For w_f in each L1:
   (a) Replace diacritics in w_f with the corresponding Latin equivalent (e.g., “é” $\rightarrow$ “e”).

[6] We use the Enchant spellchecking library: https://www.abisource.com/projects/enchant/; 14,176 unique misspelled words were identified in TOEFL4 and 6,912 in ICLE4.

[7] We use Python’s translation tool: https://pypi.org/project/translate/

[8] Following [30] we consider a word pair (w_e, w_f) to be cognate if their Levenshtein distance [27] is less than three.

[9] If $D(w_m, w_f) < D(w_e, w_f)$ was for several L1s, we opted for the one with the lowest $D(w_m, w_f)$ value. If the lowest $D(w_m, w_f)$ value was the same for several L1s, the word was discarded.

[10] http://www.omegawiki.org/Meta:Main_Page

(b) Compute the Levenshtein distance $D(w_m, w_f)$.
(c) Identify the L1 with the smallest $D(w_m, w_f)$ value, and if $D(w_m, w_f) < 5$ then take w_m to be an L2-ed version of w_f, and consider w_m as a clue for the native language of the author. [11]

Table 6 presents the statistics of misspelled words, cognates, and L2-ed words for each language in the TOEFL4 and ICLE4 datasets, respectively. The number of L2-ed words is much larger than the number of cognates: in both datasets around 40% were assigned the corresponding L1 (5,754 out of the 14,176 unique misspelled words in TOEFL4 and 2,770 out of 6,912 in ICLE4). This could be because of the tight constraint for "cognatehood" we followed [30]. In TOEFL4, the cognate and the L2-ed word lists have 350 elements in common (310 of which have the same identified L1). 230 cognates were not identified as L2-ed words and 5,404 L2-ed words were not identified as cognates. In ICLE4, the cognate and the L2-ed word lists have 266 elements in common (231 of which have the same identified L1). 148 cognates were not identified as L2-ed words and 2,504 L2-ed words were not identified as cognates.

Table 6. Statistics (absolute number and ratio (%) to the total number of words) of misspelled words, cognates, and L2-ed words for each language in the TOEFL4 and ICLE4 datasets.

L1	TOEFL4						ICLE4					
	Misspelled	**Ratio, %**	**Cognates**	**Ratio, %**	**L2-ed**	**Ratio, %**	**Misspelled**	**Ratio, %**	**Cognates**	**Ratio, %**	**L2-ed**	**Ratio, %**
French	8,150	2.31	884	0.25	3,457	0.98	3,038	1.34	281	0.12	1,211	0.53
German	7,544	1.99	425	0.11	2,869	0.76	3,913	1.69	244	0.11	1,259	0.54
Italian	8,403	2.58	585	0.18	3,249	1.00	3,223	1.43	267	0.12	1,105	0.49
Spanish	10,224	2.82	617	0.17	3,988	1.10	5,899	2.96	613	0.31	2,323	1.16
Total	**34,321**	**2.41**	**2,511**	**0.18**	**13,563**	**0.95**	**16,072**	**1.82**	**1,405**	**0.16**	**5,898**	**0.67**
Unique	**14,176**		**580**		**5,754**		**6,912**		**414**		**2,770**	

We combine the L1s of misspelled cognates and L2-ed words with the POS & FW representation, as shown below for the two phrases: *have a happy anciainty* and *a good inocent man*.[12]

have a happy ancianity
ancianity ≈ SPA. ancianidad → **L2-ed**
→ **have a JJ SPA-L2**

a good inocent man
inocent ≈ SPA. *inocente* → **cognate**
→ **a JJ SPA-cognate NN**

We extract n-grams (n = 1–3) from this representation.

[11] If the lowest $D(w_m, w_f)$ value was the same for several L1s, the word was discarded.

[12] Extracted from the training essays in the data we work with (ICLE4: SPM04022.txt and TOEFL4: 00284.txt, respectively).

Spelling errors Spelling errors are considered a strong indicator of an author's L1, since they reflect L1 influences, such as sound-to-character mappings. Chen et al. [10] showed that adding spelling error character n-grams to other features common in the literature (word and lemma n-grams) improves NLI classification accuracy.

Following [10] we represent misspelled words through character n-grams (n = 1–3), and add them as a separate subset of the feature vector representing an essay.

3.3 Learning set-up

We used the liblinear scikit-learn [41] implementation of Support Vector Machines (SVM) with OvR (one vs. the rest) multi-class strategy. The effectiveness of SVM has been proven by numerous experiments on text classification tasks in general and on NLI in particular [11, 31, 32].

Currently, the most successful learning formalism for numerous (maybe most) NLP tasks is deep learning. Its main strength comes from the fact that it can learn not only the model – in the form of weights of its various units – but also an input representation that is adapted to the learning task and the architecture. This formalism was not appropriate for our task for the following reasons: (i) the nature of the data – identifying a speaker's L1 relies on features very particular to speakers of L1, and could not be generated based on the given data as it is rather small; (ii) the phenomena investigated – we focused on specific linguistic phenomena, which can be captured by particular features that model very specific, and unconventional, contexts (e.g., usage of punctuation, particular misspellings, anglicized words), the model cannot rely on features built based on general corpora that essentially summarize thousands of occurrences of the same words or character sequences in different contexts. We confirmed these assumptions empirically, using different types of representations, including embeddings and one-hot vectors to model occurrence as opposed to meaning (thus simulating the kind of features we used). These assumptions are also supported by the literature: [46] and [13] compare approaches using deep learning with SVM, Naive Bayes and logistic regression. String kernels provide better performance, with a slight increase in a meta-learning set-up that includes word embeddings [13]. [46] use deep learning on representations of documents built based on a word x document occurrence matrix (and not word or document embeddings as more commonly used with deep learning methods). [9] compared the performance of various deep learning models (e.g., convolutional neural networks (CNN), several sequence models: recurrent neural networks (RNN), long short-term memory networks (LSTM)) with machine learning models (SVM) for NLI, concluding that machine learning models outperform deep learning models for this task. This was also confirmed by our experiments with deep learning models (CNN, LSTM, and bidirectional encoder representations from transformers (BERT)), which showed lower results than machine learning models. This can be partly related to the large number of out-of-vocabulary words in our data (around 50% of the words are out-of-vocabulary words), for which an informative representation cannot (and maybe, should not – as their occurrence is what is interesting, not their meaning) be derived based on the available data.

3.4 Evaluation

The results were measured in terms of classification accuracy. Where not otherwise specified, the experiments were carried out under 10-fold cross-validation.

4 Experiments and Results

We report the results of the experiments conducted in different settings to measure the impact of punctuation, emotion expressions, cognates, L2-ed words, and other misspellings on the NLI task.

4.1 Punctuation

Punctuation is specific to each language, and is part of the indicators that overtly represent the manner in which each language organizes and conveys information. We have designed a suite of experiments to help clarify the role/impact of punctuation usage as indicators of the author's native language: (i) the (usual) one-step classification, (ii) a two-step classification – classify the language family/geographical group, and then the actual L1 – to check whether there are commonalities across language families, and also particularities that distinguish punctuation usage for specific languages within a family/group, (iii) we test the use of punctuation within different proficiency levels, to check whether with better L2 skills, the speakers also adopt punctuation usage closer to a native speaker's, (iv) cross-corpus and cross-topic testing, to verify the robustness of punctuation features.

Multi-class and two-step classifications Multi-class classification setting is the usual NLI task, where the L1 of the author of a document is predicted based on a specific representation of the document.

We postulate that punctuation usage from L1 is reflected in the text produced in L2. Since native languages belong to specific families, we check whether there are strong influences within the language families, as well as at individual language level through a 2-step classification: (i) a coarse classification into language families/geographical grouping of languages, (ii) fine-grained classification within each language group.

Based on the languages represented in each dataset, we group them either by language family or by geographical location[13]. The language grouping used is the following:

TOEFL11: Arabic; Asian = {Chinese, Korean, Japanese}; Romance = {French, Italian, Spanish}; German; Indian = {Hindi, Telugu}; Turkish.

[13] While a grouping based on language family is more theoretically justifiable, the close results (and for some settings better) in terms of accuracy for the 2-step classification seem to support the geographical grouping of languages as well, which can be explained by shared prosody – and in the written mode, shared information organizational patterns (also evidenced by the results presented below).

ICLE: Slavic = {Bulgarian, Czech, Russian}; Asian = {Chinese, Japanese}; Romance = {French, Spanish}.

Table 7 shows the multi-class classification results (column *1-step*) in terms of accuracy (%) for POS n-grams (n = 2, 3, and 1–3) and POS & FW n-grams (n = 1–3) without PMs, with a place holder (PH), and with PMs. As a reference point we provide the random baseline (also used in the NLI shared tasks 2013 [51] and 2017 [29]): 9.09% for 11 classes in the TOEFL11 dataset and 14.29% for 7 classes in the ICLE dataset. In this and further experiments, the number of features (No.) is provided. Statistically significant gains/drops according to McNemar's statistical significance test [34] with an α value < 0.05 compared to the same representation without PMs are marked with '*' and compared to the same representation with PH are marked with '$^{+}$'.

Table 7. 10-fold cross-validation results (accuracy, %); POS and POS & FW n-grams with and without PMs; 1- and 2-step approaches. '*' and '$^{+}$' mark statistically significant differences with the same representation without PM and with PH, respectively.

	TOEFL11 dataset			**ICLE dataset**		
Features	**1-step acc.**	**2-step acc.**	**No.**	**1-step acc.**	**2-step acc.**	**No.**
Random baseline	9.09	9.09	–	14.29	14.29	–
POS 2-grams w/o PMs	34.65	34.78	1,056	54.42	52.47	951
POS 2-grams w/ PH	36.96*	38.35*	1,078	60.78*	57.27*	935
POS 2-grams w/ PMs	43.88*$^{+}$	42.92*$^{+}$	2,389	65.32*$^{+}$	63.90*$^{+}$	1,766
POS 3-grams w/o PMs	38.98	38.93	16,390	61.17	53.77	10,767
POS 3-grams w/ PH	44.40*	44.35*	16,769	66.49*	61.82*	10,746
POS 3-grams w/ PMs	47.95*$^{+}$	48.36*$^{+}$	29,525	69.09*$^{+}$	63.77*	17,992
POS 1–3 w/o PMs	40.16	40.39	17,483	62.86	55.19	11,755
POS 1–3 w/ PH	45.62*	45.14*	17,885	69.87*	60.91*	11,719
POS 1–3 w/ PMs	49.74*$^{+}$	49.65*$^{+}$	31,985	72.08*	66.62*$^{+}$	19,824
POS & FW 1–3 w/o PMs	64.06	63.07	411,599	74.42	69.87	138,170
POS & FW 1–3 w/ PH	65.57*	65.75*	342,179	79.09*	73.51*	119,306
POS & FW 1–3 w/ PMs	67.03*$^{+}$	66.51*	380,132	80.26*	76.62*	133,818

The inclusion of place holders (PH) significantly improves the results for all the considered settings. When PHs are replaced by the actual PM, the improvements are significant in the vast majority of settings.

The purpose of the 2-step classification set-up was to determine whether there are commonalities in punctuation usage across languages within the same family/geographical group. This would reflect the grammatical/prosody/information structuring in different language families or groups.

The improvement for the 2-step approach demonstrates that there are shared patterns of punctuation usage across the grouped languages and within individual languages.

The analysis of the 10 top features according to their weights for each dataset revealed that PMs are present among the 10 top features for all of the classes. The most frequent punctuation marks in these highly ranked features (bigrams and trigrams) were commas and full stops. An ablation study conducted to reveal the most indicative PM-enriched features showed that the performance does not come from one pattern, but L1-specific combinations.

It is interesting to note that the combination of POS & FW 1–3-gram features significantly outperforms the POS representation. This result is in line with previous studies on NLI, e.g., [28], and provides additional evidence for the effectiveness of FW features in this task.

Proficiency-level classification As students increase their language proficiency, it would be expected that their usage of punctuation will get closer to a native's, and the influence of their native language to get weaker. To test whether this is indeed the case, we have built a balanced dataset (from the point of view of proficiency levels) as a subset of the TOEFL11 dataset. The distribution of English proficiency levels in the TOEFL11 dataset is quite imbalanced, as shown in Table 2. To produce a balanced subset, we extract the same number of essays within each proficiency level (equal to the minimum number of essays for each level for each L1, Table 8).

Table 8. Balanced distribution of English proficiency levels in the TOEFL11 dataset.

	English Proficiency		
L1	**Low**	**Medium**	**High**
Arabic	199	199	199
Chinese	98	98	98
French	63	63	63
German	15	15	15
Hindi	29	29	29
Italian	164	164	164
Japanese	188	188	188
Korean	169	169	169
Spanish	79	79	79
Telugu	94	94	94
Turkish	90	90	90
Total	1,188	1,188	1,188

We use both the imbalanced and balanced subsets to perform multi-class classification based on the proficiency level using POS and POS & FW n-grams without PM features, with place holders (PHs), and with the actual PMs, to determine the impact the punctuation has within each proficiency level.

We investigate whether higher proficiency levels lead to punctuation usage closer to an L2 native speaker. Should that be the case, the performance in native language identification should decrease with higher proficiency levels, particularly when adding punctuation marks to the document representations.

The results for each proficiency level on the imbalanced and balanced subsets of the TOEFL11 dataset are shown in Table 9.

Table 9. 10-fold cross-validation results (accuracy, %) for the imbalanced and balanced settings for each proficiency level.

Features	Imbalanced setting			Balanced setting		
	1-step acc.	2-step acc.	No.	1-step acc.	2-step acc.	No.
Low proficiency						
POS 1–3-grams w/o PMs	41.10	39.77	9,751	37.56	38.64	9,509
POS 1–3-grams w/ PH	42.10	43.91*	10,226	42.39*	41.75*	9,959
POS 1–3-grams w/ PMs	45.03*$^{+}$	46.39*	14,228	44.64*$^{+}$	42.51*	13,791
POS & FW 1–3-grams w/o PMs	52.40	51.35	91,340	50.09	48.48	86,167
POS & FW 1–3-grams w/ PH	52.76	52.56	83,295	50.08	49.66	78,672
POS & FW 1–3-grams w/ PMs	53.75	53.76	89,873	51.18	52.44*$^{+}$	84,869
Medium proficiency						
POS 1–3-grams w/o PMs	43.07	42.83	15,334	36.46	38.05	10,072
POS 1–3-grams w/ PH	48.39*	47.59*	15,824	38.31	38.55	10,467
POS 1–3-grams w/ PMs	51.91*$^{+}$	52.57*$^{+}$	26,711	42.80*$^{+}$	42.17*$^{+}$	14,907
POS & FW 1–3-grams w/o PMs	66.52	64.07	288,658	52.40	53.03	104,367
POS & FW 1–3-grams w/ PH	68.42*	66.61*	245,100	53.24	53.11	93,170
POS & FW 1–3-grams w/ PMs	69.17*$^{+}$	68.38*$^{+}$	270,216	54.06	53.20	100,341
High proficiency						
POS 1–3-grams w/o PMs	34.65	34.89	14,454	32.49	32.66	10,639
POS 1–3-grams w/ PH	38.19*	40.31*	14,686	35.34*	35.10	10,810
POS 1–3-grams w/ PMs	42.17*$^{+}$	43.67*$^{+}$	24,644	37.68*$^{+}$	37.63*	16,039
POS & FW 1–3-grams w/o PMs	54.25	54.14	242,880	43.96	43.69	114,928
POS & FW 1–3-grams w/ PH	57.75*	57.38*	206,111	45.75	46.13	101,018
POS & FW 1–3-grams w/ PMs	58.94*$^{+}$	57.95*	227,043	47.01*	45.12	109,227

It is interesting to note that while the L1 classification results based on POS and POS & FW n-grams go down for high proficiency levels, the impact of adding the punctuation marks is still high. According to Hirvela et al. [18], L2 English learners are confident about their use of punctuation. However, the high improvement for high-proficiency learners in both imbalanced and balanced settings suggests that learners keep their native language (L1) punctuation style even when achieving high English proficiency.

Cross-topic and cross-corpus classification Brooke and Hirst [5] have criticized the datasets used for NLI because they represent different topics, and thus the performance of the n-gram-based classifiers is questionable as they capture topics rather than native language phenomena. To investigate the impact of punctuation features which are rather abstract, we perform cross-topic and cross-corpus classification.

Cross-topic The essays in the TOEFL11 dataset were written in response to eight different topics or prompts (P0–P7), and all eight prompts are represented in all 11 L1 groups. We split the dataset in two ways:

1. make folds based on the topics – a topic will be present in only one fold (8 topics → 8-fold cross-validation).
2. use 5,838 essays written on the first four prompts (P0–P3) for training and 6,262 essays written on the P4–P7 prompts for testing. To compare the result of this experiment with a mixed-topic scenario with approximately the same number of essays for training and testing, we split the TOEFL11 dataset using half of the essays on each prompt for training (6,050 essays) and testing (6,050 essays): e.g. from the 140 essays of Chinese learners on P0, 70 are used for training and 70 for testing.

Cross-corpus We extract subsets of our two datasets that represent the same languages. The TOEFL11 and the ICLE datasets have 7 common languages: Chinese, French, German, Italian, Japanese, Spanish, and Turkish. We extract the subsets corresponding to these languages from the two corpora. We use each in turn for training and testing, respectively. For this experiment, we did not balance the ICLE7 dataset and used all the essays for each of the selected languages. The data statistics are shown in Table 10.

Table 10. Number (No.) of essays per class in the ICLE dataset used for the cross-corpus experiment.

Language	No. of essays
Chinese	982
French	347
German	437
Italian	392
Japanese	366
Spanish	251
Turkish	280
Total	3,055

The purpose of the cross-topic and cross-corpus experiments is to show that the influence of punctuation from the native language transcends topics and corpora. The results for cross-topic classification are presented in Tables 11–12. Separating the training and test data based on topics leads to a drop in performance of approx. 5 percentage points in both the cross-validation and train/test split conditions. But for both settings, adding the punctuation-based features leads to very similar increases in performance whether the topics are separated or mixed. This indicates that the punctuation-based features are robust and portable across topics.

The cross-corpus experiments explore further the robustness of the punctuation-based features. An overfitting model would lead to lower scores when tested on a corpus different to the training corpus. We include here experiments done using word n-grams (n = 1–3), tf weighted just like our other features (as described in Section 3). The results

Table 11. 8-fold cross-validation and one fold/topic setting results.

	TOEFL11		**TOEFL11**	
	(8FCV)		**(topic = fold)**	
Features	**Acc.**	**No.**	**Acc.**	**No.**
POS 1–3-grams w/o PMs	40.13	17,483	34.44	17,040
POS 1–3-grams w/ PMs	49.68*	31,985	43.61*	30,782
POS & FW 1–3-grams w/o PMs	63.84	411,599	56.58	382,147
POS & FW 1–3-grams w/ PMs	66.59*	380,132	59.87*	353,842

Table 12. Mixed- and cross-topic settings results.

	TOEFL11		**TOEFL11**	
	(mixed-topic)		**(cross-topic)**	
Features	**Acc.**	**No.**	**Acc.**	**No.**
POS 1–3-grams w/o PMs	38.35	15,290	32.05	14,993
POS 1–3-grams w/ PMs	46.31*	26,117	40.96*	25,676
POS & FW 1–3-grams w/o PMs	59.85	282,178	49.60	273,579
POS & FW 1–3-grams w/ PMs	61.87*	263,682	55.48*	256,345

(training on TOEFL7 and testing on ICLE7, and vice versa) are shown in Table 13. 10FCV stands for 10-fold cross-validation on the training data (accuracy, %).

Table 13. Cross-corpus classification results for POS, FW, and word n-grams with and without PMs.

Training on TOEFL7, testing on ICLE7			
Features	**10FCV acc.**	**Test set acc.**	**No.**
POS 1–3-grams w/o PMs	50.22	46.12	15,733
POS 1–3-grams w/ PMs	60.79*	52.47*	27,731
POS & FW 1–3-grams w/o PMs	72.09	65.79	315,822
POS & FW 1–3-grams w/ PMs	75.71*	68.48*	291,740
Word 1–3-grams w/o PMs	80.91	73.81	1,904,839
Word 1–3-grams w/ PMs	83.52*	74.47	1,806,102
Training on ICLE7, testing on TOEFL7			
Features	**10FCV acc.**	**Test set acc.**	**No.**
POS 1–3-grams w/o PMs	77.24	35.39	15,500
POS 1–3-grams w/ PMs	85.66*	41.75*	28,432
POS & FW 1–3-grams w/o PMs	87.29	43.68	271,318
POS & FW 1–3-grams w/ PMs	90.86*	47.21*	256,521
Word 1–3-grams w/o PMs	92.47	43.66	1,706,554
Word 1–3-grams w/ PMs	94.11*	47.19*	1,644,978

Despite the loss in performance suffered by the model based on POS and word n-gram features, the PM features are robust and lead to the same increase in performance on testing as they did on training. While the loss in performance when training on TOEFL / testing on ICLE is relatively small (5–7 percentage points for accuracy), training on ICLE and testing on TOEFL leads to much more dramatic drops (45 percentage points for accuracy). For the models based on word n-grams, their high results are harder to improve by the addition of PM features, but they contribute nonetheless, and when added to the model trained on ICLE their impact on the TOEFL data is higher than on ICLE.

4.2 Emotions

We explore the hypothesis that emotion is one of the dimensions of language that surfaces from the native language into a second language. To check the role of emotions in native language identification (NLI), we model emotion information through polarity and emotion load features, and use document representations using these features to classify the native language of the author. After confirming the potential of emotion words for NLI, we follow the same experimental procedure as for punctuation: (i) one-step classification, (ii) two-step classification – classify the language family/geographical group, and then the actual L1, (iii) proficiency level analysis, (iv) cross-corpus and cross-topic testing, to verify the robustness of punctuation features.

As emotion words, we consider the 14,182 words listed in the NRC Word-Emotion Association Lexicon (NRC emotion lexicon) [36] and their associations with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were gathered through Amazon Mechanical Turk[14]. Table 14 presents the emotion words statistics for our data.

Table 14. Emotion words statistics (absolute number and frequency) sorted from the highest to the lowest.

TOEFL11				ICLE			
L1	**No.**	**L1**	**%**	**L1**	**No.**	**L1**	**%**
Hindi	96,184	Korean	24.93	Czech	20,162	Chinese	26.81
Teluglu	88,979	Hindi	24.62	Russian	20,142	Bulgarian	25.06
German	88,268	Chinese	24.32	Bulgarian	18,939	Japanese	24.74
Chinese	87,486	Teluglu	24.19	Spanish	17,187	Russian	24.72
Turkish	83,945	Japanese	24.15	Chinese	16,794	French	23.88
Korean	82,878	Turkish	23.90	French	16,750	Czech	23.81
French	82,454	French	23.30	Japanese	16,234	Spanish	23.33
Spanish	81,497	German	23.21				
Italian	75,339	Italian	23.16				
Japanese	73,740	Spanish	22.40				
Arabic	69,156	Arabic	21.91				

[14] https://www.mturk.com

As a first step we assess the impact of emotion words on the NLI task. Because the bag-of-words (BoW) representation covers a variety of phenomena, including emotion, we experiment using the BoW representation, and subsequently the BoW without the words that have an emotional dimension. To verify that the effect in classification is not just due to the removal of features, we randomly select a set of words of the same size as the emotion words, and remove those from the BoW to test the effect. Table 15 presents the 10-fold cross-validation results (accuracy, %) on the TOEFL11 and ICLE datasets, when using emotion words and random words as features, as well as the results when excluding emotion words or random words from the BoW approach. Random words accuracy was calculated as average over five experiments with five different sets of random words.

Table 15. Performance of emotion words.

	TOEFL11		**ICLE**	
Features	**Acc., %**	**No.**	**Acc., %**	**No.**
BoW	68.65	61,339	80.65	20,032
Random words	36.15	8,187	70.21	6,465
Emotion words	46.75	8,187	72.86	6,465
BoW w/o random words	66.68	53,152	76.83	13,567
BoW w/o emotion words	63.11	53,152	75.19	13,567

The results in Table 15 show that emotion words have higher impact on classification accuracy than random words when evaluated in isolation. Moreover, the accuracy drop is higher when excluding emotion words from the BoW approach than when excluding random words, confirming that emotion is a useful dimension for L1 classification, and not just an effect of having additional features.

Multi-class and two-step classification Following our experiments with punctuation, we provide the results when adding emotion-based features to POS and POS & FW n-gram (n = 1–3) feature set. The 10-fold cross-validation results in terms of accuracy (%) on the TOEFL11 and ICLE datasets are shown in Tables 16 and 17, respectively.

The experimental results show that emotion features, in particular the *emoP* features, significantly contribute to the results for all the considered settings, indicating that different cultures (as defined by the authors' L1) have different emotion word usage. It is interesting to note that despite being very general, the three types of *emoL* features – 13 features that characterize the emotional load of a document – also improve the results in the majority of settings, including when combined with the *emoP* features. This supports the hypothesis that some cultures use a larger or smaller emotional vocabulary. More fine grained emotional load features could improve the results further.

The confusion matrices for the POS & FW 1–3-grams combined with all emotion-based features (*emoP* and three types of *emoL* features) on the TOEFL11 and ICLE datasets are shown in Fig. 1.

Table 16. 10-fold cross-validation accuracy for POS 1–3-grams combined with emotion-based features. '*' marks statistically significant differences with the baseline.

Features	TOEFL11		ICLE	
	Acc., %	No.	Acc., %	No.
POS 1–3-grams (baseline)	40.16	17,483	62.86	11,755
POS 1–3-grams + emoL (binary)	40.60	17,485	62.86	11,757
POS 1–3-grams + emoL (1)	40.55	17,484	62.73	11,756
POS 1–3-grams + emoL (10)	40.31	17,439	62.73	11,765
POS 1–3-grams + emoL (binary) + emoL (1) + emoL (10)	40.60*	17,496	62.60	11,768
POS 1–3-grams + emoP	50.36*	216,090	67.66*	90,920
POS 1–3-grams + emotion-based features	50.33*	216,013	67.79*	90,933

Table 17. 10-fold cross-validation accuracy for POS & FW 1–3-grams combined with emotion-based features. '*' marks statistically significant differences with the baseline.

Features	TOEFL11		ICLE	
	Acc., %	No.	Acc., %	No.
POS 1–3-grams	40.16	17,483	62.86	11,755
POS & FW 1–3-grams (baseline)	64.06	411,599	74.42	138,170
POS & FW 1–3-grams + emoL (binary)	64.10	411,601	74.42	138,172
POS & FW 1–3-grams + emoL (1)	64.21	411,600	74.42	138,171
POS & FW 1–3-grams + emoL (10)	64.22	411,609	74.42	138,180
POS & FW 1–3-grams + emoL (binary) + emoL (1) + emoL (10)	64.32*	411,612	74.42	138,183
POS & FW 1–3-grams + emoP	67.73*	880,595	77.92*	268,605
POS & FW 1–3-grams + emotion-based features	68.05*	880,608	78.31*	268,618

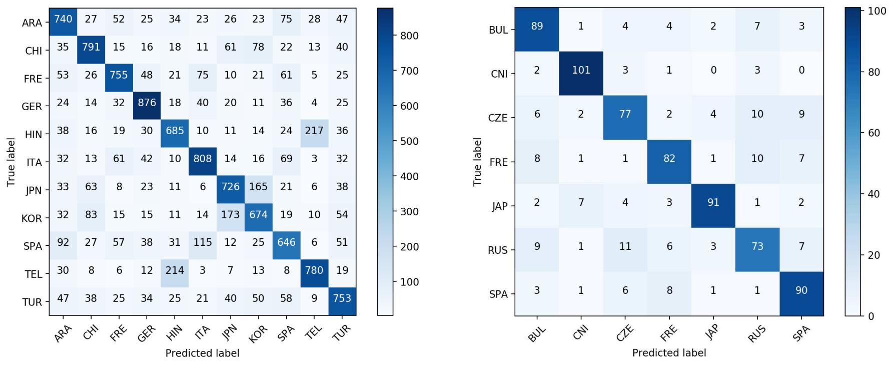

Figure 1. Confusion matrices for the POS & FW 1–3-grams in combination with all emotion-based features (*emoP* and three types of *emoL* features) on the TOEFL11 (left) and ICLE (right) datasets.

It is notable that there is a high confusion between Hindi and Telugu, the two L1s that use the largest absolute number of emotion words when producing texts in English as L2 (Table 14). Hindi and Telugu were the most problematic L1s to identify, with the highest degree of confusion, in both editions of the NLI shared task [29, 51]. The fact that these languages share a geographical space and as such have highly interacting populations could account for shared traits of their native speakers that translate into similar linguistic markers produced in L2. The lowest results on the ICLE dataset were obtained for Czech and Russian, again the languages with the largest number of emotion word usage. Czech and Russian are both European Slavic languages, representing cultures that have historically interacted for long periods of time. As in the case of Hindi and Telugu, this could explain shared traits that surface in L2 and make them harder to distinguish. Understanding which phenomena (in this case, similar emotional profiles) makes languages harder to distinguish could provide insights and allow for the development of more focused and detailed representations to develop more robust NLI systems to identify these languages.

The relative confusion between Telugu/Hindi and Czech/Russian raises the question whether emotion usage indeed depends on language family/geographical group. The high improvement provided by the emotion-based features for the two-step classification (Table 18) suggests that there are commonalities across language families, and also particularities that distinguish emotion usage for specific languages within a family/group.

Table 18. 2-step 10-fold cross-validation accuracy. '*' marks statistically significant differences.

Features	**TOEFL11**		**ICLE**	
	Acc., %	**No.**	**Acc., %**	**No.**
POS 1–3-grams	40.39	17,483	55.19	11,755
POS 1–3-grams + emotion-based features	49.78*	216,103	64.94*	90,933
POS & FW 1–3-grams	63.07	411,599	69.87	138,170
POS & FW 1–3-grams + emotion-based features	66.65*	880,608	73.25	268,618

The results for the first step using the POS & FW 1–3-grams in combination with emotion-based features, and the classification results within each group in the two-step approach are shown in Table 19. The confusion matrices for the first step are presented in Fig. 2.

On the ICLE dataset, the confusion is mostly between the Romance and Slavic languages, which to a certain degree coexist in the same geographical area, while the confusion with Asian languages is much weaker.

Proficiency-level classification The ability to choose the proper words to express oneself increases with the proficiency level. This could be reflected in different ways from the point of view of the emotions expressed: (i) L2 speakers will choose words with emotion content that is closer to their L1 or (ii) their emotional word usage will become closer to a native speaker's, and thus further from their L1. We experiment with

Table 19. 10-fold cross-validation accuracy for the first step and within each group in the two-step approach.

Step/group	**TOEFL11** **Acc., %**	**ICLE** **Acc., %**
Step 1	82.03	86.36
Asian (ASI)	74.09	94.44
Romance (ROM)	78.57	85.71
Indian (IND)	76.02	77.74
Overall	66.65	73.25

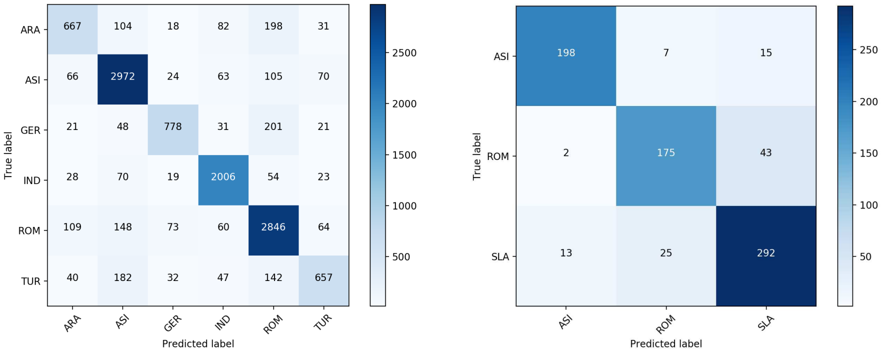

Figure 2. Confusion matrices for the first step in the two-step approach using the POS+FW 1–3-grams in combination with emotion-based features on the TOEFL11 (left) and ICLE (right) datasets.

L1 classification separating the data based on the three different proficiency levels in the TOEFL11 dataset. The results are included in Table 20. The "%" row shows the rate of increase in performance when emotion features are included in the representation. These features have a clear positive impact for all proficiency levels, and a higher impact for the medium and high levels. This could be explained by the fact that with higher proficiency, L2 speakers are better able to choose the desired lexical material in L2, and can express themselves in a manner closer to what they intend.

Table 20. 10-fold cross-validation accuracy for each proficiency level with and without emotion-based features (imbalanced setting). '*' marks statistically significant differences.

	Low		**Medium**		**High**	
	Acc., %	No.	Acc., %	No.	Acc., %	No.
POS 1–3-grams	41.10	9,751	43.07	15,334	34.65	14,454
POS 1–3-grams + emotion-based features	44.56*	51,108	52.64*	152,059	42.55*	136,783
% improvement:	8.42		22.22		22.80	
POS & FW 1–3-grams	52.40	91,340	66.52	288,658	54.25	242,880
POS & FW 1–3-grams + emotion-based features	54.58	155,725	69.12*	585,083	57.23*	491,342
% improvement:	4.16		3.91		5.49	
No. of emotion words:	62,223		475,665		372,025	
Ratio:	0.228		0.235		0.242	

Cross-topic and cross-corpus classification Different topics may also elicit different levels of emotion word usage. To explore this issue, we conducted experiments for the topics in the TOEFL11 dataset (Table 21).[15] The improvement brought by the emotion-based features does seem to depend on the topic. The highest improvements were achieved for P5 (car usage) and P7 (young vs. old people comparison). When combined with the POS & FW representation, emotion-based features are less helpful (not statistically significant improvements) for the topics discussing traveling (P1), ideas vs. facts (P3), and education (P4). Overall, adding emotion-based features to POS and POS & FW representations leads to accuracy improvement for all the topics.

4.3 L2-ed words, misspelled cognates, and other misspellings

We analyze in parallel three of the phenomena responsible for the *incorrect* expansion of L2's vocabulary using L1 material: misspelled cognates, L2-ed words, and all other spelling errors. We analyze how much each of these phenomena reveal about the L2 speaker's native language. We use the subsets of the TOEFL11 and ICLE datasets that cover languages that use the Latin script (cf. Section 3).

[15] We do not perform such an experiment on the ICLE dataset because of its high number of topics, and low number of essays per topic.

Table 21. 10-fold cross-validation accuracy for each topic in the TOEFL11 dataset. '*' marks statistically significant differences.

	P0	P1	P2	P3	P4	P5	P6	P7
POS 1–3-grams	33.74	39.26	38.54	39.89	42.40	38.29	42.08	38.15
POS 1–3-grams + emotion-based features	41.08*	47.44*	44.85*	45.96*	49.06*	49.61*	49.01*	47.25*
POS & FW 1–3-grams	50.54	56.54	53.28	55.62	60.34	56.84	57.79	55.31
POS & FW 1–3-grams + emotion-based features	53.11*	57.66	56.91*	57.04	62.28	62.40*	61.46*	58.97*
No. of emotion words:	99,606	75,308	116,795	118,427	122,741	129,837	107,924	139,288
Ratio:	0.213	0.239	0.222	0.226	0.238	0.239	0.243	0.274

Multi-class classification We first examine the features obtained from misspelled words – cognates, L2-ed, spelling error (SE) character n-grams – and verify whether they are informative for NLI. First, we use only the aggregated information about identified L1s as features, then we use them in combination with the spelling error character n-grams (n = 1–3). We compare the obtained results with the majority baselines of 25.00% and 30.62% accuracy for TOEFL4 and ICLE4, respectively. We then use as a baseline the POS & FW features, to which we add the cognates, L2-ed words, and spelling error character n-grams. The POS tags of the cognates and L2-ed words are replaced by the identified L1, and n-grams from this representation are built (cf. Subsection 3.2). SE character n-grams are represented through separate feature vectors.

The results for this experiment are shown in Table 22.

Table 22. 10-fold cross-validation accuracy for cognates, L2-ed words, their combination, and when combined with spelling error (SE) character n-grams on the TOEFL4 and ICLE4 datasets, and for POS & FW 1–3-grams combined with the cognate and L2-ed features and in combination with SE character n-grams.

	TOEFL4		ICLE4	
Features	**Acc.%**	**No.**	**Acc.%**	**No.**
Majority baseline	25.00		30.62	
Cognates	37.34*	4	38.55*	4
L2-ed	36.05*	4	44.85*	4
Cognates & L2-ed	39.84*	8	46.18*	8
Cognates & L2-ed & SE	54.55*	7,347	56.33*	6,391
POS & FW 1–3-grams	74.45	231,737	80.58	189,622
POS & FW 1–3-grams & cognates	75.50*	236,716	80.72	192,572
POS & FW 1–3-grams & L2-ed	75.80*	247,814	81.56	198,469
POS & FW 1–3-grams & cognates & L2-ed	76.20*	253,175	81.77	201,623
POS & FW 1–3-grams & SE	78.23*	238,929	82.75*	195,869
POS & FW 1–3-grams & cognates & L2-ed & SE	78.80*	260,367	82.61*	207,870

The improvement in terms of accuracy over the majority baselines by more than 10 percentage points achieved when using the proposed features in isolation confirms that these features are highly relevant for NLI. Combining these features further boosts the results, showing that their L1 signal is strengthened with each additional source of information. The combination of L2-ed words and misspelled cognates provide statistically significant improvement in the majority of cases. Spelling error character n-grams further increase the results. Replacing the POS tags of the misspelled words with the corresponding L1s, and using word n-grams of such features (n = 1–3) provides improvement on both datasets.

A complicating factor in this classification is the fact that the four languages represented in the dataset have shared etymological ancestors and thus shared cognates. Furthermore, three of these languages are Romance languages, and thus are even closer, and may confound the Levenshtein distance computation.

Proficiency-level classification We evaluate the impact of cognates and L2-ed words within each proficiency level. It is expected that the impact (as well as the frequency) of L2-ed words will decrease with an increase in proficiency.

The statistics for the number of essays per language within each proficiency level is shown in Table 23. The statistics for the misspelled words, cognates, and L2-ed words (as a percentage of the total number of tokens) for each language within each proficiency level is provided in Fig. 3. As all these phenomena are gathered from misspelled words, it is not surprising that their overall frequency decreases with the proficiency level. The number of L2-ed words is still higher than the number of cognates throughout all proficiency levels and L1s. Analysis of the identified L2-ed words reveal that many of them do have a common etymological ancestor as a word from L2, but they are written in such a way that their Levenshtein distance from the L2 version is greater than their distance from the L1 version. Using information about shared etymologies could help make the separation between words with shared etymologies and “corrupted" L1 words clearer.

Table 23. Data statistics for the three English proficiency levels in TOEFL4.

L1	**Low**		**Medium**		**High**	
	No.	**%**	**No.**	**%**	**No.**	**%**
French	63	19.6	577	26.5	460	24.2
German	15	4.7	412	18.9	673	35.3
Italian	164	51.1	623	28.6	313	16.4
Spanish	79	24.6	563	25.9	458	24.1
Total	321	7.3	2,175	49.4	1,904	43.3

The results for each proficiency level (Table 24) indicate that, in the majority of cases, the influence of L2-ed words gets weaker from low to high proficiency, while the influence of the cognates grows with the proficiency level, despite the fact that even for higher levels of proficiency the number of L2-ed words is higher than the number of cognates. This shows that even high-proficiency language users are prone to extend their vocabulary in L2 incorrectly, but following cognate principles, when no fitting

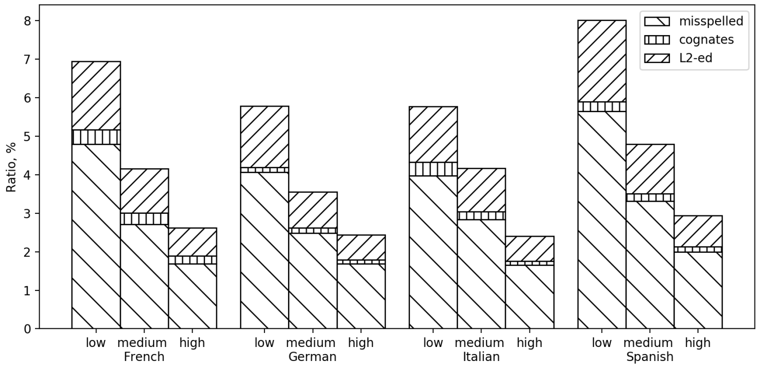

Figure 3. Ratio (%) of the misspelled words, cognates, and L2-ed words to the total number of words for each language within each proficiency level.

Table 24. 10-fold cross-validation accuracy for cognates, L2-ed words, their combination, and when combined with spelling error (SE) character n-grams for each proficiency level, and for POS & FW 1–3-grams combined with the cognate and L2-ed features and in combination with SE character n-grams.

	Low		**Medium**		**High**	
Features	**Acc., %**	**No.**	**Acc., %**	**No.**	**Acc., %**	**No.**
Majority baseline	51.09		28.64		35.35	
Cognates	56.49*	4	39.81*	4	40.23*	4
L2-ed	58.12*	4	38.39*	4	36.24	4
Cognates & L2-ed	59.24*	8	42.57*	8	40.18*	8
Cognates & L2-ed & SE	60.79*	3,241	55.26*	6,031	45.95*	5,366
POS & FW 1–3-grams	62.92	34,970	74.33	148,878	67.71	152,105
POS & FW 1–3-grams & cognates	62.38	35,609	75.57*	152,158	68.08	154,318
POS & FW 1–3-grams & L2-ed	65.16	37,214	76.17*	159,508	68.03	160,025
POS & FW 1–3-grams & cognates & L2-ed	64.54	37,922	77.09*	163,057	68.55	162,419
POS & FW 1–3-grams & SE	66.09	38,114	78.14*	154,774	70.07*	157,346
POS & FW 1–3-grams & cognates & L2-ed & SE	69.13*	41,066	79.25*	168,953	71.28*	167,660

lexical item is readily available to them. Higher results are usually achieved when these features are combined, regardless of the proficiency level.

High improvement achieved for medium proficiency can be related to a larger number of essays for this level.[16]

4.4 Correct cognates

In the experiments presented above, we exploited only misspelled words to extract L1-indicative features. While we do not expect to find L2-ed words among the correctly spelled words, there will be correct cognates. In order to detect properly spelled cognates, we used etymological information obtained from the Etymological WordNet [35]. We identify "perfect" cognates if the lemma occurs in the Etymological WordNet's L1 vocabulary, while "not perfect" cognates are identified as words (lemmas) that share an etymological ancestor and their Levenshtein distance < 3 (diacritics removed). The Levenshtein distance was used since the ancestor can have multiple descendants.

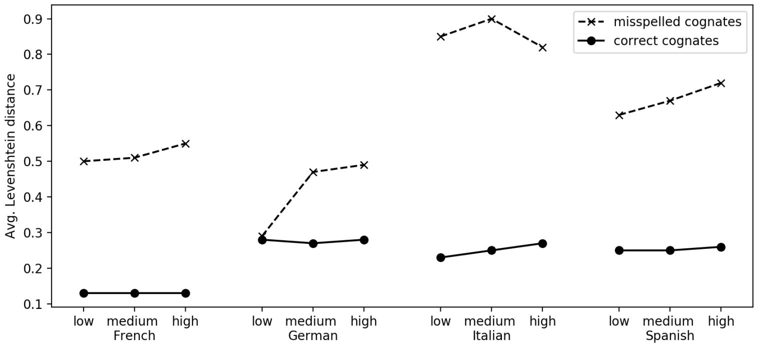

Figure 4. Average Levenshtein distances for correct and misspelled cognates for each language within each proficiency level.

When the L1s of the identified correct cognates are used as separate features, they perform by around 3 pp above the majority baseline, but do not enhance the results when combined with misspelled cognates and L2-ed words. This could be related to the fact that correct cognates are either closest to their L1 form, or are part of a more basic vocabulary that all learners have to master. We design features that capture the distance between cognates in L2 and some L1 – for correct cognates we use the average of the Levenshtein distances for each L1 as a numeric feature. These features outperform the majority baseline by around 4% on TOEFL4 and 6% on ICLE4. When combined with L2-ed words, misspelled cognates, or POS & FW 1–3 gram representations, the

[16] We do not balance the dataset by proficiency levels for this experiment, because the dataset will become too small.

improvement on ICLE4 (1%–5% improvement depending on the setting) is higher than on TOEFL4 (1%–3% improvement depending on the setting), which could be due to the topics or the high proficiency level of the ICLE essays.

Analysis of the average Levenshtein distances in our datasets and within each proficiency level for correct and misspelled cognates reveal that the average Levenshtein distance is lower for correct cognates (Fig. 4), which indicates that learners tend to correctly use cognates when they are closer to the form they are familiar with in their L1. This distance increases with the proficiency level, which can be due to the fact that learners with high proficiency use more complex vocabulary, with cognates that have a form that is more distant from the one in L1.

Another factor to consider are false friends. Since words are judged outside of their context and based only on their form, false friends are not distinguished from proper cognates. The word *became* may appear correct, unless the larger context is taken into account: *I became a letter.* Such a usage would reveal the writer to be a native German speaker, where *bekommen* means *to receive*. Detecting false friends, however, is a difficult problem.

Gathering all such information would provide additional insight on how the L1 vocabularies influence lexical choice in L2, and we plan to address some of these issues in future work.

5 Revisiting the Performance

The main goal of our investigation is to identify and analyze very specific linguistic phenomena with respect to their interference in second language production. In particular, we focused on punctuation usage (as an expression of structuring the information conveyed), emotion (as an expression of cultural biases with respect to the usage of emotion terms in written communication) and cognates and sound-driven types of spelling errors (as an expression of the influence of the forms and sounds from the native language on second language production). These phenomena, and others, are indiscriminately covered by word and character n-gram text representations, which are very successful in the native language identification task. After the separate analysis of the contribution of the three linguistic phenomena we targeted, it is natural to ask how much of the gap in performance between a baseline essay representation – POS tags and function words – and the word and character n-gram representation we can fill by adding to the baseline the features that capture these phenomena. Table 25 and 26 show the results of these experiments, under multi-class and cross-corpus settings[17].

There are several observations. First, with respect to the best performing representation using character n-grams, n varies from corpus to corpus, and also between different experimental set-ups (within-corpus or cross-corpus). The large size of n also hints at some overfitting, and this is confirmed by the large differences (avg. 19 pp for TOEFL and 54 pp for ICLE) between the within-corpus (10 fold cross validation – 10FCV) and cross-corpus (test) set-ups presented in Table 26. Second, while the phenomena

[17] For cross-corpus experiment we used the 7-way TOEFL and ICLE datasets, as described in Section 4.1.

Table 25. Multi-class classification. Our approach vs. baselines. ‘*’ and ‘$^{+}$’ mark statistically significant differences with BoW and the best representation, respectively.

	TOEFL11		ICLE	
Features	**Acc., %**	**No.**	**Acc., %**	**No.**
BoW	68.65	61,339	80.65	20,032
Character 3-grams	65.68	21,994	80.00	12,575
Character 4-grams	71.32	98,962	86.23	47,439
Character 5-grams	74.69	283,959	87.79	126,883
Character 6-grams	76.75	653,011	**87.92**	276,837
Character 7-grams	77.12	1,289,672	87.66	500,305
Character 8-grams	**77.13**	2,250,146	86.88	779,287
Character 9-grams	76.63	3,531,284	85.06	1,086,862
POS & FW & PMs & emotions & cognates & L2-ed & SE	72.83*$^{+}$	840,734	82.86^{+}	264,198

Table 26. Cross-corpus classification. Our approach vs. baselines. ‘*’ and ‘$^{+}$’ mark statistically significant differences with BoW and the best representation, respectively.

Training on TOEFL7, testing on ICLE7			
Features	**10FCV**	**Test set**	**No.**
BoW	75.66	58.82	42,693
Char. 3-grams	73.69	45.11	18,148
Char. 4-grams	78.70	57.22	76,609
Char. 5-grams	81.10	63.08	214,822
Char. 6-grams	83.12	65.24	492,425
Char. 7-grams	83.25	**68.05**	964,096
Char. 8-grams	**83.30**	67.89	1,654,907
Char. 9-grams	82.90	65.92	2,548,330
POS & FW & PMs & emotions & cognates & L2-ed & SE	80.13*$^{+}$	63.60*$^{+}$	626,826
Training on ICLE7, testing on TOEFL7			
Features	**10FCV**	**Test set**	**No.**
BoW	91.85	39.45	40,252
Char. 3-grams	92.34	32.12	18,776
Char. 4-grams	93.91	35.77	78,487
Char. 5-grams	**94.44**	40.96	222,884
Char. 6-grams	94.41	42.60	513,134
Char. 7-grams	94.21	44.38	996,599
Char. 8-grams	93.92	43.53	1,676,691
Char. 9-grams	93.78	41.95	2,510,088
POS & FW & PMs & emotions & cognates & L2-ed & SE	92.01^{+}	**46.14***$^{+}$	573,452

explicitly represented do not cover the entire performance gap (evidenced by the negative difference with respect to the best representation), they are extremely robust, and compensate for the overfitting of the BoW representation (evidenced by the consistent difference between the combined representation and the baseline). We happily note that in the cross-corpus ICLE set-up (training on ICLE7, testing on TOEFL7), the baseline model enriched with the combined representation of our targeted phenomena has a higher performance than the best character n-gram representation.

These results indicate that there is much to be learned by explicitly analysing linguistic phenomena that have the potential to interfere in second language production. The advantages are not only in understanding how to adjust materials for a more close-to-optimal second language acquisition, but also in obtaining a higher and more robust performance in native language identification.

6 Conclusions

In this article we have presented the results of our investigation of various aspects of language that seep into L2 production: the languages' structuring of information through the use of punctuation, emotion expression in language, and incorrect expansion of L2's vocabulary through commonalities or similarities of form with the vocabulary of the native language L1.

The experiments conducted to investigate the impact of punctuation on native language identification showed that punctuation marks provide useful information for the NLI task. Their impact is positive for both coarse (family-language level) and fine-grained classification, indicating that there are patterns of punctuation usage that are common across language families, but also patterns specific to individual languages. Punctuation interference does not decrease with the level of proficiency and they are abstract and robust NLI features, as shown by the results of cross-topic and cross-corpus experiments.

We investigated the hypothesis that the use of emotions is indicative of an author's native language. We used two types of emotion-based features – one that captures the types of sentiments expressed, the other captures the frequency of emotion words in essays. The fact that adding these features to POS and function word n-grams leads to improvements in predicting a text's author's native language leads us to conclude that emotion characteristics from a native language are "imported" into the production of L2.

We analyzed misspellings for particular clues about an essay author's native language: we identified misspelled cognates and L2-ed (here, anglicized) words and analyzed the information they provided separately and combined with other misspellings. Our experiments showed that all three phenomena provide useful information for identifying the native language of the author. An analysis of these phenomena at different levels showed that although the frequency of misspellings in general – and of L2-ed words – decreases with an increase in proficiency, as expected, their contribution to the NLI task remains strong for all levels. When combined, the results increase in most tested scenarios, showing that the L1 signal is boosted by considering all these phenomena together. We find it particularly interesting that L2-ed words are still frequent at the high proficiency level, showing that the impulse of using cognates is so strong that people make them when they are not available.

The analysis of the contribution of the targeted phenomena to the native language identification task shows that explicitly analysing linguistic phenomena that have the potential to interfere in second language production is beneficial from multiple perspectives, including understanding how to adjust materials for better second language acquisition, and also building more robust native language identification systems.

References

1. Baron, N.: Commas and canaries: The role of punctuation in speech and writing. Language Sciences 23(1), 15–67 (2001)
2. Bergsma, S., Kondrak, G.: Alignment-based discriminative string similarity. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. pp. 656–663. ACL, Prague, Czech Republic (2007)
3. Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., Chodorow, M.: TOEFL11: A corpus of non-native English. ETS Research Report Series 2013(2), i–15 (2013)
4. Brooke, J., Hirst, G.: Native language detection with 'cheap' learner corpora. In: Proceedings of the Conference of Learner Corpus Research. pp. 37–47. Presses universitaires de Louvain, Louvain-la-Neuve, Belgium (2011)
5. Brooke, J., Hirst, G.: Robust, lexicalized native language identification. In: Proceedings of the 24th International Conference on Computational Linguistics. pp. 391–408. The COLING 2012 Organizing Committee, Mumbai, India (2012)
6. Bruthiaux, P.: Knowing when to stop: Investigating the nature of punctuation. Language and Communication 13(1), 27–43 (1993)
7. Caldwell-Harris, C.: Emotionality differences between a native and foreign language: Theoretical implications. Frontiers in Psychology 5(1055) (2014)
8. Chaski, C.: Empirical evaluations of language-based author identification techniques. Forensic Linguistics 8(1), 1–65 (2001)
9. Chen, L.: Native Language Identification on Learner Corpora. Master's thesis, University of Trento, Department of Information Engineering and Science, Trento, Italy (2016)
10. Chen, L., Strapparava, C., Nastase, V.: Improving native language identification by using spelling errors. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 542–546. ACL, Vancouver, Canada (2017)
11. Cimino, A., Dell'Orletta, F.: Stacked sentence-document classifier approach for improving native language identification. In: Proceedings of the 12th Workshop on Building Educational Applications Using NLP. pp. 430–437. ACL, Copenhagen, Denmark (2017)
12. Flanagan, B., Hirokawa, S.: An automatic method to extract online foreign language learner writing error characteristics. International Journal of Distance Education Technologies 16(4), 15–30 (2018)
13. Franco-Salvador, M., Kondrak, G., Rosso, P.: Bridging the native language and language variety identification tasks. Procedia Computer Science 112, 1554 – 1561 (2017), http://www.sciencedirect.com/science/article/pii/S1877050917314126
14. Goldin, G., Rabinovich, E., Wintner, S.: Native language identification with user generated content. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3591–3601. ACL, Brussels, Belgium (2018)
15. Gómez-Adorno, H., Bel-Enguix, G., Sierra, Gerardo, S.O., Quezada, D.: A machine learning approach for detecting aggressive tweets in Spanish. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages. vol. 2150, pp. 97–101. CEUR-WS.org, Seville, Spain (2018)

16. Granger, S., Dagneaux, E., Meunier, F., Paquot, M.: International Corpus of Learner English v2 (ICLE). Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium (2009)
17. Grieve, J.: Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing 22(3), 251–270 (2007)
18. Hirvela, A., Nussbaum, A., Pierson, H.: ESL students' attitudes toward punctuation. System 40(1), 11–23 (2012)
19. Ionescu, R.T., Popescu, M.: Can string kernels pass the test of time in native language identification? In: Proceedings of the 12th Workshop on Building Educational Applications Using NLP. pp. 224–234. ACL, Copenhagen, Denmark (2017)
20. Ionescu, R.T., 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. pp. 1363–1373. ACL, Doha, Qatar (2014)
21. 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. ACL, Atlanta, GA, USA (2013)
22. Kestemont, M.: Function words in authorship attribution. From black magic to theory? In: Proceedings of the 3rd Workshop on Computational Linguistics for Literature. pp. 59–66. ACL, Gothenburg, Sweden (2014)
23. 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, New York, NY, USA (2005)
24. Kumar, A., Ganesh, B., Ajay S, Soman, P.: Overview of the second shared task on Indian native language identification (INLI). In: Working notes of FIRE 2018 - Forum for Information Retrieval Evaluation. vol. 2266, pp. 39–50. CEUR Workshop Proceedings, Gandhinagar, India (2018)
25. Kumar, A., Ganesh, B., Singh, S., Soman, P., Rosso, P.: Overview of the INLI PAN at FIRE-2017 track on Indian native language identification. In: Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation. vol. 2036, pp. 99–105. CEUR Workshop Proceedings, Bangalore, India (2017)
26. Leersnyder, J.D., Mesquita, B., Kim, H.S.: Where do my emotions belong? a study of immigrants' emotional acculturation. Personality and Social Psychology Bulletin 37(4), 451–463 (2011)
27. Levenshtein, V.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10(8), 707–710 (1966)
28. Malmasi, S., Dras, M.: Multilingual native language identification. Natural Language Engineering 23(2), 163–215 (2015)
29. Malmasi, S., Evanini, K., Cahill, A., Tetreault, J., Pugh, R., Hamill, C., Napolitano, D., Qian, Y.: A report on the 2017 native language identification shared task. In: Proceedings of the 12th Workshop on Building Educational Applications Using NLP. pp. 62–75. ACL, Copenhagen, Denmark (2017)
30. Mann, G., Yarowsky, D.: Multipath translation lexicon induction via bridge languages. In: Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics. pp. 151–158. ACL, Pittsburgh, PA, USA (2001)
31. Markov, I., Chen, L., Strapparava, C., Sidorov, G.: CIC-FBK approach to native language identification. In: Proceedings of the 12th Workshop on Building Educational Applications Using NLP. pp. 374–381. ACL, Copenhagen, Denmark (2017)
32. Markov, I., Sidorov, G.: CIC-IPN@INLI2018: Indian native language identification. In: Working notes of FIRE 2018 - Forum for Information Retrieval Evaluation. vol. 2266, pp. 82–88. CEUR Workshop Proceedings, Gandhinagar, India (2018)

33. Markov, I., Stamatatos, E., Sidorov, G.: Improving cross-topic authorship attribution: The role of pre-processing. In: Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing. vol. 10762, pp. 289–302. Springer, Budapest, Hungary (2018)
34. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12(2), 153–157 (1947)
35. de Melo, G., Weikum, G.: Towards universal multilingual knowledge bases. In: Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the 5th Global WordNet Conference. pp. 149–156. Narosa Publishing House, Mumbai, India (2010)
36. Mohammad, S., Turney, P.: Crowdsourcing a word-emotion association lexicon. Computational Intelligence 29, 436–465 (2013)
37. Moore, N.: What's the point? The role of punctuation in realising information structure in written English. Functional Linguistics 3(1), 6 (2016)
38. Newman, M., Pennebaker, J., Berry, D., Richards, J.: Lying words: Predicting deception from linguistic styles. Personality and Social Psychology Bulletin 29(5) (2003)
39. Nicolai, G., Hauer, B., Salameh, M., Yao, L., Kondrak, G.: Cognate and misspelling features for natural language identification. In: Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 140–145. ACL, Atlanta, GA, USA (2013)
40. Odlin, T.: Language Transfer: cross-linguistic influence in language learning. Cambridge University Press, Cambridge, UK (1989)
41. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
42. Pennebaker, J., Booth, R., Francis, M.: Linguistic Inquiry and Word Count: LIWC2007. Austin, TX: LIWC.net (2007)
43. Rabinovich, E., Tsvetkov, Y., Wintner, S.: Native language cognate effects on second language lexical choice. Transactions of the Association for Computational Linguistics 6, 329–342 (2018)
44. Rangel, F., Rosso, P.: On the identification of emotions and authors' gender in Facebook comments on the basis of their writing style. In: Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and perspectives from AI. vol. 1096, pp. 34–46. CEUR-WS.org, Torino, Italy (2013)
45. Rangel, F., Rosso, P.: On the impact of emotions on author profiling. Information Processing & Management 52(1), 74–92 (2016)
46. Rangel, F., Rosso, P., Brooke, J., Uitdenbogerd, A.: Cross-corpus native language identification via statistical embedding. In: Proceedings of the Second Workshop on Stylistic Variation. pp. 39–43. ACL, New Orleans, LA, USA (Jun 2018), https://www.aclweb.org/anthology/W18-1605
47. Schmid, H.: Improvements In Part-of-Speech Tagging With an Application to German, pp. 13–25. Springer (1999)
48. Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A., Gordon, J.: Empirical study of machine learning based approach for opinion mining in tweets. In: Proceedings of the Mexican International Conference on Artificial Intelligence. vol. 7629, pp. 1–14. Springer, San Luis Potosí. Mexico (2013)
49. Smith, T., Witten, I.: Language inference from function words. Tech. Rep. 93/3, Department of Computer Science, University of Waikato (1993), computer Science Working Papers

50. Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., Al-Ghamdi, F., Hirschberg, J., Chang, A., Fung, P.: Overview for the first shared task on language identification in code-switched data. In: Proceedings of the First Workshop on Computational Approaches to Code Switching. pp. 62–72. ACL, Doha, Qatar (2014)
51. 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. pp. 48–57. ACL, Atlanta, GA, USA (2013)
52. Tetreault, J., Blanchard, D., Cahill, A., Chodorow, M.: Native tongues, lost and found: Resources and empirical evaluations in native language identification. In: Proceedings of the 24th International Conference on Computational Linguistics. pp. 2585–2602. The COLING 2012 Organizing Committee, Mumbai, India (2012)
53. Torney, R., Vamplew, P., Yearwood, J.: Using psycholinguistic features for profiling first language of authors. Journal of the Association for Information Science and Technology 63(6) (2012)
54. Volkova, S., Ranshous, S., Phillips, L.: Predicting foreign language usage from English-only social media posts. In: Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 608–614. ACL, New Orleans, LA, USA (2018)
55. Wierzbicka, A.: Emotions across languages and cultures: Diversity and universals. Cambridge University Press (1999)