Anglicized Words and Misspelled Cognates in Native Language Identification

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

[1]CLiPS, University of Antwerp, Antwerp, Belgium
[2]University of Heidelberg, Heidelberg, Germany
[3]Fondazione Bruno Kessler, Trento, Italy
ilia.markov@uantwerpen.be, nastase@cl.uni-heidelberg.de,
strappa@fbk.eu

Abstract

In this paper, we present experiments that estimate the impact of specific lexical choices of people writing in a second language (L2). In particular, we look at misspelled words that indicate lexical uncertainty on the part of the author, and separate them into three categories: misspelled cognates, "L2-ed" (in our case, anglicized) words, and all other spelling errors. We test the assumption that such errors contain clues about the native language of an essay's author through the task of native language identification. The results of the experiments show that the information brought by each of these categories is complementary. We also note that while the distribution of such features changes with the proficiency level of the writer, their contribution towards native language identification remains significant at all levels.

1 Introduction

Producing an utterance in a language, be it the native, second or n-th one, relies in large part on the vocabulary range of the speaker. When dealing with a second language L2, this range may be correctly or incorrectly expanded through commonalities or similarities of form with the vocabulary of the native language L1. Examples of this process are cognates, which are words that have the same ancestors or were derived from the same sources, that we often approximate in computational approaches as words having similar forms and similar meaning in L1 and L2, for example, SPA. *religión* and ENG. *religion*. Research in psycholinguistics and native language identification have shown that using cognates when producing L2 is common and shared across native speakers of the same L1 to the degree that a quite accurate phylogenetic language tree can be reconstructed (Rabinovich et al., 2018).

In this paper, 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. Misspelled cognates are words that are misspellings from the point of view of L2, but have a very close form in L2 and L1. L2-ed words are something like false cognates (not in the sense of false friends): words in L1 that were "adjusted" to seem and sound like legitimate L2 words. For example, a Spanish native speaker could use the incorrectly anglicized word *lentaly* instead of *slowly* (SPA. *lentamente*). From the point of view of the L2 vocabulary, L2-ed words are spelling errors, but they are special because they have a very similar L1 form. Chen et al. (2017) have shown that spelling errors, represented as character n-grams, are also very indicative of an author's L1, as they may capture language-specific sound-to-spelling mappings.

The experiments presented in this paper aim to analyze how much each of these phenomena reveal about the L2 speaker's native language. We analyze misspelled words and split them into cognates, L2-ed words or all other misspellings, and analyze their impact through the task of native language identification (NLI). The goal of NLI is to identify the native language (L1) of a person based on his/her writing in the second language (L2). The underlying hypothesis is that the L1 influences learners' second language writing as a result of the language transfer effect (Odlin, 1989). NLI is usually approached as a multi-class classification problem of assigning class labels representing L1s to essays written in L2. 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. NLI is an interesting example of a task which is hard to perform for humans: the study of human per-

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 275–284
Florence, Italy, August 2, 2019.

formance in NLI (Malmasi et al., 2015) showed that automated systems significantly outperform human annotators (73% vs. 37% accuracy, respectively).

We test the impact of the three phenomena – misspelled cognates, L2-ed words, spelling errors – on the subsets of the TOEFL (Blanchard et al., 2013) and ICLE (Granger et al., 2009) datasets that cover languages that use the Latin script. The results of the multi-class classification experiments show that the role of all these phenomena is significant. Higher results are achieved when features representing each of these are combined, indicating that they are complementary for the NLI task. Experiments on data split by proficiency levels show that the L2-ed based features have a higher impact the lower the proficiency level, while the influence of the cognates grows with the proficiency level. This is not surprising, but it reveals an interesting phenomenon – when people do not know a word in a target language, they may make a "false cognate", and while the vocabulary of a proficient speaker is larger, they still resort occasionally to this incorrect lexicon expansion. Understanding the source and effects of lexical choice in L2 speakers, and how this changes with proficiency levels, could have direct applications in second language teaching.

2 Related Work

Cognates. Cognates are words that have the same ancestor, or were derived from the same "borrowed" sources. The "cognatehood" of word pairs may be obscured by phonological and spelling changes in different languages, and by the drift in their meaning from the common source: e.g., *milk* (ENG.), *latte* (ITA.), *gala* (GER.) are all cognates despite their current different forms, while *journey* (ENG.) and *journeé* (FRA. *day*) have a common etymological ancestor but their current meaning has lost this connection (*journey* used to mean *a day's travelling*). Because of the lack of computational resources on word etymologies until relatively recently, cognates have been approximated in computational linguistics as words that have similar form and meaning. The influence of cognates as indicators of an author's native language has been explored in various ways through the task of native language identification.[1]

Nicolai et al. (2013) add cognate-based features to frequently used ones (e.g., character and word n-grams, syntax production rules, misspelling features) for the NLI shared task 2013 (Tetreault et al., 2013). Cognates were detected by identifying misspelled words whose form is closer to an L2 word w_{L2} than to w_{L2}'s translation in L1. The authors report that cognate features, in spite of being extracted just for 4 out of 11 languages, improved the accuracy by 0.7% and reduced the relative error rate by about 4%.

Rabinovich et al. (2018) investigate the cognate effect on lexical choice in L2 of advanced non-native speakers. They construct a *focus set* of more than 1,000 words, that have synonyms (provided by WordNet) with different etymologies (provided by the Etymological WordNet), thus potentially leading to different patterns of usage for speakers with different L1s. The influence of cognates on lexical choice is measured through frequency of usage with respect to this list of words. Aggregated evidence for all texts belonging to the same L1 can be used to build a relatively accurate phylogenetic language tree for the Indo-European language family (31 languages).

Nastase and Strapparava (2017) did not look specifically at cognates, but used etymological information to build etymological ancestor profiles for sets of English essays written by different L1 speakers. This representation quantified the influence of different etymological ancestors when producing texts in L2, and showed that these influences are different depending on L1.

From the previous studies it is hard to see the quantitative impact of cognates on the NLI task: in the study by Nicolai et al. (2013) cognates were used in combination with a large number of features (including words and word 2-grams), while in (Nastase and Strapparava, 2017; Rabinovich et al., 2018) the authors were mostly concerned with reconstructing language family tree and not with the role of cognates in the task of NLI.

Spelling errors. Spelling errors were used in one of the first studies on NLI (Koppel et al., 2005). The authors focused on syntax errors and eight types of spelling errors, e.g., missing let-

[1] Distinguishing between actual cognates and false friends is not being done, so when we refer to cognates in the related literature or in our own work, we mean both.

ters, repeated letters, double letters appearing only once, among others. The relative frequency of each error type to the length of the essay was used as the corresponding feature value. When combining these with commonly used features, i.e., function words, the authors obtained 80.2% accuracy on a 5-way subset of the ICLE dataset.

Nicolai et al. (2013) focused on the misspelled part of a word and used pairs of correct and misspelled parts as character n-gram features. Misspelling features contributed 0.4% accuracy to their NLI shared task system when used in combination with other commonly used NLI features.

Chen et al. (2017) also explored spelling errors, testing the hypothesis that spelling errors capture L1-biased sound-to-spelling mappings. Spelling errors were represented as character n-grams, and added to other commonly used features (word, lemma, and character n-grams). Including these typo-based features leads to an increase in NLI accuracy of 1.2% on the TOEFL11 test set.

Flanagan and Hirokawa (2018) 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.

These studies clearly show that spelling errors are influenced by an author's L1. The source of such errors was not of interest though, and they may hide interesting linguistic phenomena, like cognates and L2-ed words.

L2-ed words. The combination of languages within one text has been studied before, under the name of code switching or code mixing, e.g., (Solorio et al., 2014). This switching/mixing though happens at the word level, and lexical items in the text belong fully to one language. In the phenomenon we study here, the switching/mixing happens below the word level, where the word in a language L1 is inflected or adjusted to "fit" language L2.

3 Methodology

To investigate the impact of L2-ed words and cognates, we use the native language identification task: we perform multi-class classification of essays written in L2 (English in our case) by people with different native languages (L1s) – with L1 as the class labels – using a representation of these essays through features that capture these phenomena. We use two datasets – TOEFL and ICLE – previously used for NLI, and extract the subsets that cover languages that use a Latin script.

3.1 Datasets

We use two datasets commonly used in NLI research:

TOEFL (Blanchard et al., 2013): the ETS Corpus of Non-Native Written English (TOEFL11) contains 1,100 essays in English for 11 native languages. We used a 4-language subset of the corpus, focusing on the languages that use the Latin script: French, German, Italian, and Spanish. This subset, to which we refer as TOEFL4, contains 1,100 essays (with an average of 353 tokens per essay) for each of the four languages.

ICLE (Granger et al., 2009): consists of essays written by highly-proficient non-native college-level students of English. We used a 4-language subset of the corpus that represents the same languages as included in TOEFL4: French (347 essays), German (437), Italian (392), and Spanish (251). Overall, this subset, to which we refer as ICLE4, contains 1,427 essays with avg. 690 tokens/essay.

The four languages represented in the TOEFL4 and ICLE4 datasets have shared etymological ancestors and therefore shared cognates, which is a complicating factor in the classification.

3.2 Experiment setup

We used the (pre-)tokenized version of the TOEFL4 dataset and tokenized ICLE4 with the Natural Language Toolkit (NLTK) tokenizer[2], removing metadata in pre-processing. Each essay was represented through the sets of features described below, using term frequency (tf) weighting scheme and the liblinear scikit-learn (Pedregosa et al., 2011) implementation of Support Vector Machines (SVM) with OvR (one vs. the rest) multi-class strategy. We report classification accuracy on 10-fold cross-validation experiments.

3.3 Features

Following previous studies on NLI, e.g., (Markov et al., 2018a,b), we evaluate the impact of L2-ed words and cognates in combination with the part-

[2]http://www.nltk.org

of-speech (POS) tag and function word (FW) representations. POS tags and function words (FWs) are considered core features in NLI research (Malmasi and Dras, 2015), not susceptible to topic bias, unlike word and character n-grams (Brooke and Hirst, 2011).

An essay will be represented through various combinations of the feature sets we consider: POS & FW n-grams; n-grams from POS & FW sequences including word-level L1 information; character n-grams that represent misspelled words.

3.3.1 Part-of-speech tags and function words

POS features capture the morpho-syntactic patterns in a text, and are indicative of the L1, especially when used in combination with other types of features (Cimino and Dell'Orletta, 2017; Markov et al., 2017). POS tags were obtained with TreeTagger (Schmid, 1999), which uses the Penn Treebank tagset (36 tags).

FWs clarify the relationships between the content-carrying elements of a sentence, and introduce syntactic structures like verbal complements, relative clauses, and questions (Smith and Witten, 1993). The FW feature set consists of 318 English FWs from the scikit-learn package (Pedregosa et al., 2011).

3.3.2 Misspelled cognates, L2-ed words and other misspellings

We build features that gather information from misspelled words in the essays in the data. The information about which L1 a cognate or L2-ed word hints to is used as an attribute of the word.

Misspelled cognates. Several studies applied discriminative string similarity to the task of cognate identification (Mann and Yarowsky, 2001; Bergsma and Kondrak, 2007; Nicolai et al., 2013). Following the work by Nicolai et al. (2013), we detect cognates by identifying the cases where the closest correctly spelled L2 word w_e to the misspelled word w_m has a translation in an L1 w_f to which it is close in form, and w_m is closer 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.[3]
2. For each L1:
   (a) Look up the translation w_f of the intended word w_e in L1.[4]
   (b) Replace diacritics in w_f with the corresponding Latin equivalent (e.g., "é" → "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.[5]
   (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.[6]

L2-ed words. To identify the L2-ed, in our case anglicized, words we take a misspelled word 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.

We use the freely available lists of expressions provided by the OmegaWiki project[7] and extract vocabularies for each of the L1 languages represented in our datasets. The statistics for each language in terms of the number of expressions and the extracted vocabularies is provided in Table 2.

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" → "e").
   (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

[3] 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.

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

[5] Following Mann and Yarowsky (2001) we consider a word pair (w_e, w_f) to be cognate if their Levenshtein distance (Levenshtein, 1966) is less than three.

[6] 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.

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

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**	

Table 1: 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.

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

Table 2: Statistics of the number of expressions and the extracted vocabularies for each of the languages.

of w_f, and consider w_m as a clue for the native language of the author. [8]

Table 1 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 (Mann and Yarowsky, 2001). In TOEFL4, the cognate and the L2-ed word lists have 350 elements in common (310 of which have the same identified L1), while there are 230 cognates that were not identified as L2-ed words and 5,404 L2-ed words that 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), while there are 148 cognates that were not identified as L2-ed words and 2,504 L2-ed words that were not identified as cognates.

We combine the L1s of misspelled cognates and L2-ed words with the POS & FW representation. As an example consider the two phrases: `have a happy ancianity` and `a good inocent man`.[9] The identified L2-ed words and cognates are *ancianity* (ENG. old age) → SPA. ancianidad → L2-ed and *inocent* (ENG. innocent) → SPA. inocente → cognate. The phrases are represented through POS & FW & cognates & L2-ed words as `have a JJ SPA-L2-ed` and `a JJ SPA-cognate NN`, respectively. Then n-grams (n = 1–3) from this representation are extracted.

Spelling errors. Spelling errors may capture language specific transcriptions of sound sequences, as influenced by the native language (Chen et al., 2017): e.g., Spaniards often use *c* instead of *q*, writing *cuestion* instead of *question*. Following (Chen et al., 2017) we represent misspelled words through character n-grams (n = 1–3). When used, these features are added as a separate subset of the feature vector representing an essay.

4 Results and Discussion

The impact of features based on misspelled cognates, L2-ed words and character n-grams from all misspellings is evaluated using the NLI task. We report accuracy on 10-fold cross-validation experiments on the full data sets. The set-ups consist of various combinations of these features. Tests on the TOEFL dataset split by proficiency levels will allow us to assess how these features change with higher language competency.

Results on the TOEFL4 and ICLE4 datasets We first examine only the features obtained from misspelled words – cognates, L2-ed, spelling error (SE) character n-grams – and verify whether they are informative for NLI: (i) we use just the aggregated information about identified L1s as features; (ii) 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 and FW features, to which we add the cognates, L2-ed words, and spelling error character

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

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

Features	TOEFL4 Acc.%	TOEFL4 diff	TOEFL4 No.	ICLE4 Acc.%	ICLE4 diff	ICLE4 No.
Majority baseline	25.00			30.62		
Cognates	37.34	**12.34***	4	38.55	**7.93***	4
L2-ed	36.05	**11.05***	4	44.85	**14.23***	4
Cognates & L2-ed	39.84	**14.84***	8	46.18	**15.56***	8
Cognates & L2-ed & SE	54.55	**29.55***	7,347	56.33	**25.71***	6,391
POS & FW 1–3-grams	74.45		231,737	80.58		189,622
POS & FW 1–3-grams & cognates	75.50	**1.05***	236,716	80.72	**0.14**	192,572
POS & FW 1–3-grams & L2-ed	75.80	**1.35***	247,814	81.56	**0.98**	198,469
POS & FW 1–3-grams & cognates & L2-ed	76.20	**1.75***	253,175	81.77	**1.19**	201,623
POS & FW 1–3-grams & SE	78.23	**3.78***	238,929	82.75	**2.17***	195,869
POS & FW 1–3-grams & cognates & L2-ed & SE	78.80	**4.35***	260,367	82.61	**2.03***	207,870

Table 3: 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. Diff stands for difference: gain/drop; '*' marks statistically significant differences.

Features	Low Acc.%	Low diff	Low No.	Medium Acc.%	Medium diff	Medium No.	High Acc.%	High diff	High No.
Majority baseline	51.09			28.64			35.35		
Cognates	56.49	**5.40***	4	39.81	**11.17***	4	40.23	**4.88***	4
L2-ed	58.12	**7.03***	4	38.39	**9.75***	4	36.24	**0.89**	4
Cognates & L2-ed	59.24	**8.15***	8	42.57	**13.93***	8	40.18	**4.83***	8
Cognates & L2-ed &SE	60.79	**9.70***	3,241	55.26	**26.62***	6,031	45.95	**10.60***	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	**–0.54**	35,609	75.57	**1.24***	152,158	68.08	**0.37**	154,318
POS & FW 1–3-grams & L2-ed	65.16	**2.24**	37,214	76.17	**1.84***	159,508	68.03	**0.32**	160,025
POS & FW 1–3-grams & cognates & L2-ed	64.54	**1.62**	37,922	77.09	**2.76***	163,057	68.55	**0.84**	162,419
POS & FW 1–3-grams & SE	66.09	**3.17**	38,114	78.14	**3.81***	154,774	70.07	**2.36***	157,346
POS & FW 1–3-grams & cognates & L2-ed & SE	69.13	**6.21***	41,066	79.25	**4.92***	168,953	71.28	**3.57***	167,660

Table 4: 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. Diff stands for difference: gain/drop; '*' marks statistically significant differences.

L1	Low No.	Low %	Medium No.	Medium %	High No.	High %
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**

Table 5: Data statistics for the three English proficiency levels in TOEFL4.

n-grams. The POS tags of the cognates and L2-ed words are replaced by the identified L1, and we then build n-grams from this representation. SE character n-grams are represented through separate feature vectors (as explained in Section 3).

The result for this experiment is shown in Table 3. The number of features (No.) is included. Statistically significant gains with respect to the baseline according to McNemar's statistical significance test (McNemar, 1947) with $\alpha < 0.05$ are marked with '*'.

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 enhance the obtained results. Replacing the POS tags of the misspelled words by the corresponding L1s, and using word n-grams of such features (n = 1–3) provides improvement on both datasets.

On the TOEFL4 dataset, the result for the combination of the proposed features is similar to

the performance of the bag-of-words (BoW) approach, while on the ICLE4 dataset the BoW approach outperforms our representation by around 5% accuracy. The BoW approach covers a multitude of linguistic particularities, while the goal of this work is to identify which particular characteristics skew the language production in an L2.

As mentioned above, 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 experiments The TOEFL dataset contains information concerning the proficiency levels of the students (low, medium, high). We evaluated 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 5. 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 Figure 1. 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.

The results for each proficiency level when cognates and L2-ed words are evaluated separately and in combination with spelling error (SE) 1–3-grams, as well as when these features are combined with the POS & FW representation, are presented in Table 4.

The results presented in Table 4 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 lexical item is readily available to them.

High improvement achieved for medium proficiency can be related to a larger number of essays for this level.[10] Moreover, it can be noted that higher results are usually achieved when these features are combined, regardless of the proficiency level.

Discussion 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 (de Melo and Weikum, 2010). 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.

When the L1s of the identified correct cognates are used as features in isolation, they perform by around 3 percentage points 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 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 top-

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

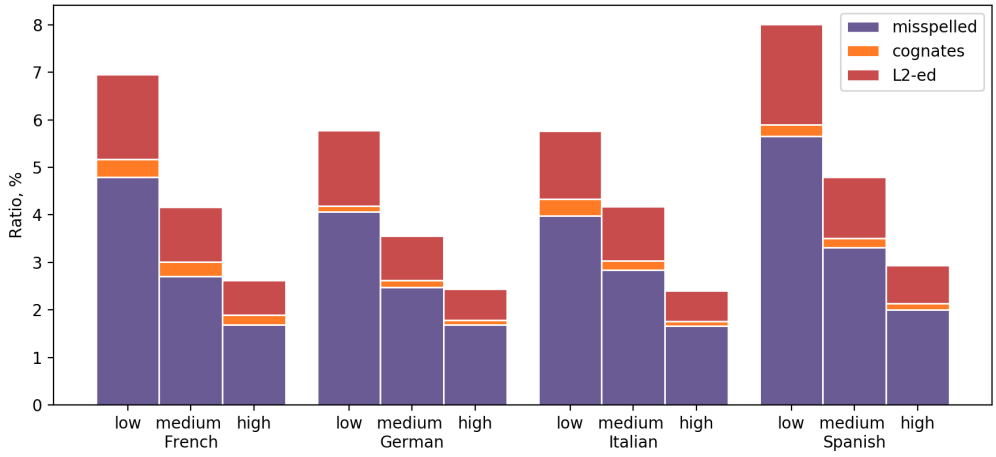

Figure 1: Ratio (%) of the misspelled words, cognates, and L2-ed words to the total number of words for each language within each proficiency level.

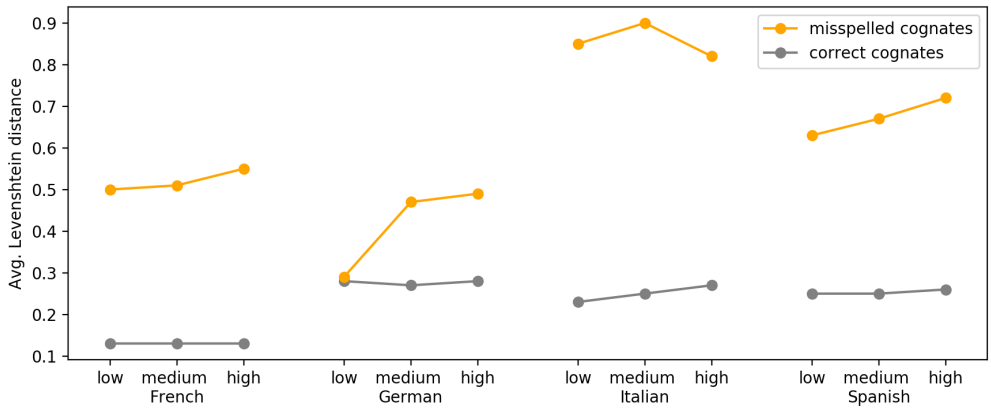

Figure 2: Average Levenshtein distances for correct and misspelled cognates for each language within each proficiency level.

ics 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 (Figure 2), 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 though is a more 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 Conclusions

In this paper, we analyzed misspellings for particular clues about an essay author's native language. In particular, we identified misspelled cognates and L2-ed (here, anglicized) words and analyzed

the information they provided separately and combined with other misspellings. Experiments on native language identification (NLI) 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.

In future work, we plan to explore deeper the usefulness of cognates and L2-ed words, by distinguishing them from false friends, which we think may be even more telling about the author's L1. We also plan to examine these phenomena – cognates, L2-ed words, and misspelled words – on datasets with other L2s, and include in the analysis languages that do not use the Latin script.

References

Shane Bergsma and Grzegorz Kondrak. 2007. Alignment-based discriminative string similarity. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 656–663, Prague, Czech Republic. ACL.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013(2):i–15.

Julian Brooke and Graeme Hirst. 2011. Native language detection with 'cheap' learner corpora. In *Proceedings of the Conference of Learner Corpus Research*, pages 37–47, Louvain-la-Neuve, Belgium. Presses universitaires de Louvain.

Lingzhen Chen, Carlo Strapparava, and Vivi Nastase. 2017. Improving native language identification by using spelling errors. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 542–546, Vancouver, Canada. ACL.

Andrea Cimino and Felice Dell'Orletta. 2017. Stacked sentence-document classifier approach for improving native language identification. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*, pages 430–437, Copenhagen, Denmark. ACL.

Brendan Flanagan and Sachio Hirokawa. 2018. An automatic method to extract online foreign language learner writing error characteristics. *International Journal of Distance Education Technologies*, 16(4):15–30.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English v2 (ICLE)*. Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. 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*, pages 624–628, New York, NY, USA. ACM.

Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Shervin Malmasi and Mark Dras. 2015. Multilingual native language identification. *Natural Language Engineering*, 23(2):163–215.

Shervin Malmasi, Joel Tetreault, and Mark Dras. 2015. Oracle and human baselines for native language identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 172–178, Denver, CO, USA. ACL.

Gideon Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 151–158, Pittsburgh, PA, USA. ACL.

Ilia Markov, Lingzhen Chen, Carlo Strapparava, and Grigori Sidorov. 2017. CIC-FBK approach to native language identification. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*, pages 374–381, Copenhagen, Denmark. ACL.

Ilia Markov, Vivi Nastase, and Carlo Strapparava. 2018a. Punctuation as native language interference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3456–3466, Santa Fe, New Mexico, USA. The COLING 2018 Organizing Committee.

Ilia Markov, Vivi Nastase, Carlo Strapparava, and Grigori Sidorov. 2018b. The role of emotions in native language identification. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 123–129, Brussels, Belgium. ACL.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Gerard de Melo and Gerhard Weikum. 2010. Towards universal multilingual knowledge bases. In *Principles, Construction, and Applications of Multilingual Wordnets. Proceedings of the 5th Global WordNet Conference*, pages 149–156, New Delhi, India. Narosa Publishing.

Vivi Nastase and Carlo Strapparava. 2017. Word etymology as native language interference. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2692–2697, Copenhagen, Denmark. ACL.

Garrett Nicolai, Bradley Hauer, Mohammad Salameh, Lei Yao, and Grzegorz Kondrak. 2013. Cognate and misspelling features for natural language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 140–145, Atlanta, GA, USA. ACL.

Terence Odlin. 1989. *Language Transfer: cross-linguistic influence in language learning*. Cambridge University Press, Cambridge, UK.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. 2018. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342.

Helmut Schmid. 1999. *Improvements In Part-of-Speech Tagging With an Application to German*, pages 13–25. Springer.

Tony C. Smith and Ian H. Witten. 1993. Language inference from function words. Working papers, https://hdl.handle.net/10289/9927.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. 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*, pages 62–72, Doha, Qatar. ACL.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57, Atlanta, GA, USA. ACL.