Punctuation as Native Language Interference

Ilia Markov
INRIA
Paris, France
ilia.markov@inria.fr

Vivi Nastase
University of Heidelberg
Heidelberg, Germany
nastase@
cl.uni-heidelberg.de

Carlo Strapparava
Fondazione Bruno Kessler
Trento, Italy
strappa@fbk.eu

Abstract

In this paper, we describe experiments designed to explore and evaluate the impact of punctuation marks on the task of native language identification. 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. Our experiments are organized in various set-ups: the usual multi-class classification for individual languages, also considering classification by language groups, across different proficiency levels, topics and even cross-corpus. The results support our hypothesis that punctuation marks are persistent and robust indicators of the native language of the author, which do not diminish in influence even when a high proficiency level in a non-native language is achieved.

1 Introduction

Native Language Identification (NLI) – identifying the native language (L1) of a person based on his/her writing in the second language (L2) – is useful for a variety of purposes, including security, marketing, and educational applications. The effect of native language phenomena seeping into texts produced in a different language is known as language transfer (Odlin, 1989). Numerous aspects of the language have been explored for NLI – character-level language models (Ionescu et al., 2014), lexical choice (Brooke and Hirst, 2012; Lahiri and Mihalcea, 2013), grammar (Nagata and Whittaker, 2013), spelling errors (Chen et al., 2017; Koppel et al., 2005), cognates (Nicolai et al., 2013), and general etymology (Nastase and Strapparava, 2017).

While punctuation has been included in some of these studies (e.g., in character-level models), its impact has not been studied. It is however an important, and often revealing, aspect of written language. For example, punctuation is a strong indicator of authorship, and has been used successfully in stylometric analysis for authorship attribution (Markov et al., 2017b; Grieve, 2007). More generally, from a linguistic point of view, punctuation has been disputed as following prosodic principles or as a clarifier of grammatical structure (Baron, 2001; Bruthiaux, 1993). Moore (2016) 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. As an example, consider the following English sentence, written by a native German speaker[1]:

> I think the biggest question is , how to defin an “ enjoyed life ” .

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

[1] Extracted from one of the training essays in the data we work with (NLI: 10086.txt), with the author’s misspelling of ‘define’.

Proceedings of the 27th International Conference on Computational Linguistics, pages 3456–3466
Santa Fe, New Mexico, USA, August 20-26, 2018.

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

We propose the hypothesis that punctuation usage from the native language appears in a speaker’s new language, and is distinctive enough to contribute to the native language identification task. To investigate this hypothesis we perform a series of experiments that measure the impact of punctuation on the NLI task, which we approach from a machine-learning perspective, as a multi-class classification problem of English (as L2) written documents, using two representations – word n-grams and part-of-speech (POS) tags – and analyze the performance of these feature sets when they include or not punctuation marks (PMs). We perform one-step classification – classify the L1 of the author directly – and also a two-step classification where in the first step we classify the language family/geographical group, and then the actual L1. This experiment shows that there are commonalities across language families, and also particularities that distinguish punctuation usage for specific languages within a family/group. We evaluate the use of punctuation within different proficiency levels, to test whether with better L2 skills, the speakers also adopt punctuation usage closer to a native speaker’s. Surprisingly, the results indicate that punctuation usage remains influenced by L1 even for learners with high proficiency in L2.

In the previously mentioned experiments we focused on more abstract features (POS tags), which we combine with punctuation marks to capture a higher-level representation of punctuation usage. However, BoW/word n-gram features are the ones that lead to the best results on the task, even though they have been disputed as being useful (Brooke and Hirst, 2012), but potentially overfitting (Brooke and Hirst, 2011). To test whether punctuation marks are robust and can compensate some shortcomings of lexical features, we perform a final set of experiments in which we produce word and punctuation n-grams and test them in cross-dataset classification (training and testing on different datasets). The high drop in performance when using word n-gram features – consistent with the hypothesis that this representation leads to models that overfit the training data – is partly countered when punctuation marks are added to the mix.

The consistent and substantial improvement in native language identification brought by including punctuation marks in the different set-ups explored indicates that punctuation usage is a robust and persistent indicator of the native language of the author.

In a nutshell, the research questions addressed in this work are the following: (i) evaluate the contribution of PMs to NLI, (ii) examine how robust and persistent their contribution is with respect to levels of proficiency and topic/corpus variation.

2 Related Work

Numerous aspects of language, possibly all, can insinuate themselves from a language learner’s native language to the targeted L2.

Numerous studies (Brooke and Hirst, 2012; Lahiri and Mihalcea, 2013; Tsur and Rappoport, 2007) show that the *lexical choices* of non-native speakers are strong indicators of their native language, and so are the spelling errors (Chen et al., 2017; Koppel et al., 2005). Lexical features and character n-grams (Ionescu et al., 2014) are considered the most indicative feature types for the NLI task. Nicolai et al. (2013) and Nastase and Strapparava (2017) find that lexical choice is partly influenced by cognates or more generally, by etymologically related ancestor languages.

Both Wong and Dras (2009) and Nagata and Whittaker (2013) investigate the influence of *grammar* as grammatical structures transfer from L1 to L2, whether they are legitimate patterns in L2 as well (Nagata and Whittaker, 2013), or are erroneous with respect to L2 (Wong and Dras, 2009).

The interest in the NLI task led to the organization of several NLI shared tasks (Tetreault et al., 2013; Malmasi et al., 2017), where participating teams used a variety of features: character n-grams of different types, n-grams of lexical features (words, lemmas, POS tags), grammatical information (parse tree rules, syntactic dependency-based n-grams), spelling errors, function words, among others. The top two teams in the recent edition of the NLI shared task (Cimino and Dell’Orletta, 2017; Markov et al., 2017a) used a combination of these features, achieving accuracy close to 90% for this task.

As linguistic studies show (Moore, 2016), punctuation is an important aspect of language. For the written mode of language, it serves to reveal/emphasize the information structure of a sentence, partly

	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%

Table 1: Data statistics for the three English proficiency levels 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: Distribution of topics in TOEFL11.

sharing the sentence structuring function of grammar. In this paper, we propose to investigate punctuation usage as a source of native language information, and understand to what degree it is reflected in writing in a new language. We do this through a suite of experiments described in Sections 4 and 5.

3 Data

3.1 Datasets

To investigate the impact of the punctuation usage on native language identification (NLI), we conducted experiments on two datasets commonly used in NLI research:

TOEFL11 (Blanchard et al., 2013): the ETS Corpus of Non-Native Written English (TOEFL11) contains 1,100 essays in English (with an average of 348 tokens per 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 dataset also contains information regarding the proficiency level (low, medium, high) of the essay authors. Dataset statistics in terms of proficiency levels and writing prompts are presented in Tables 1 and 2, respectively.

ICLEv2 (Granger et al., 2009): the ICLEv2 dataset (henceforth, ICLE) consists of essays written by highly-proficient non-native college-level students of English. We used a 7-language subset of the corpus normalized for topic and character encoding (Tetreault et al., 2012; Ionescu et al., 2014)[2]. This subset contains 110 essays (with an average of 747 tokens per essay after tokenization and removal of metadata) for each of the 7 languages: Bulgarian, Chinese, Czech, French, Japanese, Russian, and Spanish.

[2] The authors express their gratitude to A. Cahill for providing the list of documents used in their paper.

3.2 Features

The suite of experiments we report were designed to investigate the impact of punctuation-based features on native language identification. The hypothesis we are testing is whether patterns of punctuation usage – possibly motivated by prosody or grammatical structure in the native language – are indicative of the native language of the author of written essays in English. Lexicalized representation of documents (through word/character n-grams of different sizes) have been criticized as overfitting the data (Brooke and Hirst, 2012). Because of this we use more abstract POS n-gram features to represent the documents (essays) in our data, to which we add punctuation marks to capture punctuation usage in the analyzed texts.

Part-of-speech (POS) tags POS features capture the morpho-syntactic patterns in a text. They are considered indicative features for NLI, especially when used in combination with other features, such as word and character n-grams (Cimino and Dell'Orletta, 2017; Markov et al., 2017a). POS tags were obtained with the TreeTagger software package (Schmid, 1999), which uses the Penn Treebank tagset (36 POS tags). We used POS n-grams with n = 1–3.

Punctuation marks usage (PM) Because linguistic theories propose that punctuation emphasizes the information structure of a sentence, and different languages structure information differently, the usage of punctuation for an author's L1 is distinct from norms of L2. To encode punctuation usage we incorporate them in the POS and word n-gram features. We use a set of 36 punctuation marks (only 31 of which appear in the ICLE dataset).[3]

3.3 Experiment setup

We used the tokenized version of the TOEFL11 dataset and performed tokenization of the ICLE dataset using the Natural Language Toolkit (NLTK)[4] tokenizer. We performed lowercasing and removed the text surrounded by <. . . > in the ICLE dataset, since it corresponds to the dataset metadata. Each essay was represented through the sets of features described above. We used 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. The effectiveness of SVM has been proven by numerous experiments on text classification tasks. 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

We have designed a suite of experiments to help clarify the role/impact of punctuation usage as indicators of the author's native language. We start with the usual multi-class classification setting, then investigate the punctuation usage within the represented language families/geographical groups, with respect to the levels of proficiency, and finally with respect to the topics represented in the data.

Multi-class classification This 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.

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

[3] Punctuation marks used as separate features would not capture their usage and role in the sentence, which is what we aim to represent. For the curious reader, however, we can report that used in isolation, punctuation marks achieve 18.74% accuracy on the TOEFL11 dataset under 10-fold cross-validation and 33.25% on ICLE, which are twice as high as the baselines (which are included in Table 4).

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

Based on the languages represented in each dataset, we group the languages either by language family or by geographical location[5]. The language grouping used is the following:

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

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

Proficiency-level classification As students master a language better and better, 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 1. 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, in italics in Table 1).

We use both the imbalanced and balanced subsets to perform multi-class classification based on the proficiency level using POS n-grams with and without PM features, to determine the impact the punctuation has within each proficiency level.

Cross-topic and cross-corpus classification Brooke and Hirst (2012) 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 capturing 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) We split the TOEFL dataset into folds based on the topics – a topic will be present in only one fold (8 topics → 8-fold cross-validation).

(2) We used 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 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). For example, there are 140 essays of Chinese learners on P0, so we used 70 for training and 70 for testing, etc.

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 ICLE dataset and used all the essays for each of the selected languages. The number of essays per class in the ICLE dataset is shown in Table 3.

5 Results and Discussion

5.1 Multi-class classification

Table 4 shows the multi-class classification results (column *1-step*) in terms of accuracy (%) for POS n-grams (n = 1, 2, 3, and 1–3) with and without PMs. As a reference point we provide the random baseline (also used in the NLI shared tasks 2013 (Tetreault et al., 2013) and 2017 (Malmasi et al., 2017)): 9.09% for 11 classes in the TOEFL11 dataset and 14.29% for 7 classes in the ICLE dataset. We include

[5] 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 in Section 5).

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

Table 3: Number (No.) of essays per class in the ICLE dataset used for the cross-corpus experiment.

the improvement (as absolute percentage points) when using PM features over the setting when PMs are omitted. In this and further experiments, the number of features (No.) is provided; the improvements are shown in bold typeface.

Features	TOEFL11 dataset			ICLE dataset		
	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 w/o PMs	12.72	15.67	35	32.34	34.68	35
POS w/ PMs	17.50	19.12	71	48.05	46.49	66
Improvement:	**4.78**	**3.45**		**15.71**	**11.81**	
POS 2-grams w/o PMs	32.40	32.43	923	52.34	51.04	826
POS 2-grams w/ PMs	43.11	41.93	2,262	67.14	65.32	1,678
Improvement:	**10.71**	**9.50**		**14.80**	**14.28**	
POS 3-grams w/o PMs	37.99	38.19	14,036	55.19	52.73	9,455
POS 3-grams w/ PMs	46.88	47.31	27,431	65.45	61.69	16,850
Improvement:	**8.89**	**9.12**		**10.26**	**8.96**	
POS 1–3-grams w/o PMs	38.88	39.08	14,993	59.87	51.04	10,316
POS 1–3-grams w/ PMs	48.83	48.43	29,763	69.48	65.19	18,594
Improvement:	**9.95**	**9.35**		**9.61**	**14.15**	

Table 4: 10-fold cross-validation results (accuracy, %); POS n-grams with and without PMs; 1- and 2-step approaches.

As the results from Table 4 show, the inclusion of PMs improves the results for all the considered settings. The improvement in results brought by including the punctuation marks in the representation shown in Table 4 and throughout this section are statistically significant (unless explicitly mentioned otherwise) according to McNemar's statistical significance test (McNemar, 1947) with an α value of 0.05.

5.2 2-step classification

As explained in Section 4, the 2-step classification set-up would be useful to determine whether there are commonalities in punctuation usage across languages within the same family/geographical group. This would be reflective of 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 across the 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.

5.3 Proficiency-level classification

We investigate whether higher proficiency levels lead to punctuation usage closer to an L2 native speaker. Should that be the case, we should note lower performance in native language identification with higher proficiency levels, and in particular lower improvement in performance 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 Tables 5 and 6, respectively. The results are provided for 1- and 2-step approaches. Here, and in further experiments, the impact of PMs is evaluated using POS 1–3-gram features with and without PMs.

Features	**1-step acc.**	**2-step acc.**	**No.**
Low proficiency			
POS 1–3-grams w/o PMs	41.47	39.32	8,681
POS 1–3-grams w/ PMs	46.71	45.49	13,311
Improvement:	**5.24**	**6.17**	
Medium proficiency			
POS 1–3-grams w/o PMs	42.60	42.48	13,259
POS 1–3-grams w/ PMs	51.34	51.51	24,800
Improvement:	**8.74**	**9.03**	
High proficiency			
POS 1–3-grams w/o PMs	32.39	33.58	12,480
POS 1–3-grams w/ PMs	42.05	43.93	23,006
Improvement:	**9.66**	**10.35**	

Table 5: *Imbalanced setting*: 10-fold cross-validation results (accuracy, %) for each proficiency level.

Features	**1-step acc.**	**2-step acc.**	**No.**
Low proficiency			
POS 1–3-grams w/o PMs	37.87	38.22	8,471
POS 1–3-grams w/ PMs	44.08	42.76	12,900
Improvement:	**6.21**	**4.54**	
Medium proficiency			
POS 1–3-grams w/o PMs	36.00	35.52	8,953
POS 1–3-grams w/ PMs	43.36	44.11	13,996
Improvement:	**7.36**	**8.59**	
High proficiency			
POS 1–3-grams w/o PMs	31.93	31.31	9,367
POS 1–3-grams w/ PMs	37.25	39.39	14,992
Improvement:	**5.32**	**8.08**	

Table 6: *Balanced setting*: 10-fold cross-validation results (accuracy, %) for each proficiency level.

It is interesting to note that while the L1 classification results based on POS n-grams go down for high proficiency levels, the impact of adding the punctuation marks is higher for each proficiency level compared to the lower ones. According to the study conduced by Hirvela et al. (2012), 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 L1 punctuation style even when achieving high English proficiency.

5.4 Cross-topic experiments

The cross-topic and cross-corpus experiments were performed to show that the influence of punctuation from the native language transcends topics and corpora, through features that capture PM usage, can partly compensate for the loss in performance under cross-topic or cross-corpus conditions.

	TOEFL11 (10FCV)		**TOEFL11 (topic = fold)**	
Features	**Acc.**	**No.**	**Acc.**	**No.**
POS 1–3-grams w/o PMs	38.88	14,993	33.88	14,636
POS 1–3-grams w/ PMs	48.83	29,763	43.21	28,635
Improvement:	**9.95**		**9.33**	

Table 7: 10-fold cross-validation and one fold/topic setting results.

	TOEFL11 (mixed-topic)		TOEFL11 (cross-topic)	
Features	**Acc.**	**No.**	**Acc.**	**No.**
POS 1–3-grams w/o PMs	36.63	13,174	32.27	13,042
POS 1–3-grams w/ PMs	45.95	24,215	40.74	23,950
Improvement:	**9.32**		**8.47**	

Table 8: Mixed- and cross-topic settings results.

The results for cross-topic classification are presented in Tables 7–8. 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.

5.5 Cross-corpus experiments

The cross-corpus experiments explore further the robustness of the punctuation-based features. An over-fitting 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).

Training on TOEFL, testing on ICLE

	10FCV		Test Set		
Features	**Acc.**	**F1**	**Acc.**	**F1**	**No.**
POS 1–3-grams w/o PMs	48.62	48.51	43.27	40.70	13,587
POS 1–3-grams w/ PMs	60.29	60.22	54.50	53.05	25,870
Improvement:	**11.67**	**11.71**	**11.23**	**12.35**	
Word 1–3-grams w/o PMs	80.91	80.87	73.81	71.27	1,904,839
Word 1–3-grams w/ PMs	83.52	83.47	74.47	72.05	1,806,102
Improvement:	**2.61**	**2.60**	**0.66**	**0.78**[6]	

Training on ICLE, testing on TOEFL

	10FCV		Test Set		
Features	**Acc.**	**F1**	**Acc.**	**F1**	**No.**
POS 1–3-grams w/o PMs	79.47	75.21	34.22	32.26	13,730
POS 1–3-grams w/ PMs	86.67	83.82	41.64	39.72	26,890
Improvement:	**7.20**	**8.61**	**7.42**	**7.46**	
Word 1–3-grams w/o PMs	92.47	90.85	43.66	42.41	1,706,554
Word 1–3-grams w/ PMs	94.11	93.04	47.19	45.26	1,644,978
Improvement:	**1.64**	**2.19**	**3.53**	**2.85**	

Table 9: Cross-corpus classification results for POS and word n-grams with and without PMs.

The results for cross-corpus experiments (training on TOEFL11 and testing on ICLE, and vice versa) are shown in Table 9. 10FCV stands for 10-fold cross-validation on the training data (accuracy, % and F1 macro, %). We note that 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.

In a detailed, per-language view, presented through the confusion matrices (Figure 1), we can note that the highest improvement when including PM features in this cross-corpus study was achieved for German. There is also a high improvement for Turkish and Italian. They indicate that these language have stronger, or maybe more consistent, punctuation styles that interfere in the production of L2. Personal

[6] This improvement is not statistically significant.

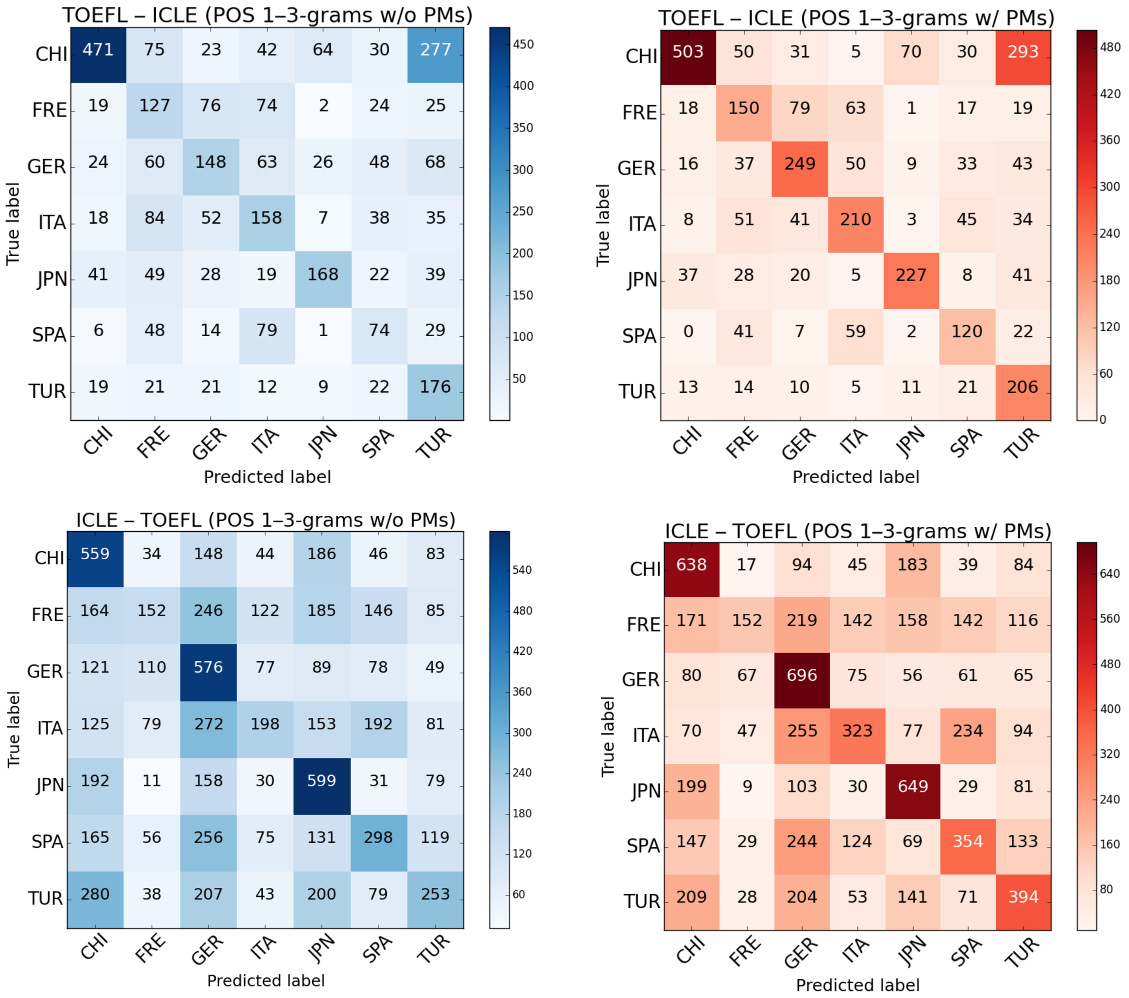

Figure 1: Training on TOEFL, testing on ICLE: POS 1–3-grams without (top left)/with PMs (top right). Training on ICLE, testing on TOEFL: POS 1–3-grams without (bottom left)/with PMs (bottom right).

experience with German and Italian punctuation confirms these findings, but our models could support deeper linguistic exploration into these phenomena.

6 Conclusions

While the role of punctuation is still disputed in linguistic theory – as a written indicator of prosody, or as grammatical features – punctuation is however linked to each language and the manner in which languages organize and convey information. We proposed the hypothesis that punctuation usage in L2 is indicative of an author's native language. We have conducted a series of experiments to investigate the impact of punctuation on native language identification. The experiments show that punctuation marks provide useful information, and when combined with POS and even word n-gram features – thus capturing their usage – lead to significant and substantial improvements. 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 seem to decrease with the level of proficiency: while we would expect that as the proficiency level increases an author's usage of punctuation will be closer to English and thus the native language will be harder to detect, this is not the case. Finally, contrary to word n-grams which necessarily capture also topic-specific information and thus tend to overfit the training data, punctuation is more abstract and as such a more robust feature, as shown by the results of cross-topic and cross-corpus experiments.

References

Naomi Baron. 2001. Commas and canaries: The role of punctuation in speech and writing. *Language Sciences*, 23(1):15–67.

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. Presses universitaires de Louvain.

Julian Brooke and Graeme Hirst. 2012. Robust, lexicalized native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 391–408. The COLING 2012 Organizing Committee.

Paul Bruthiaux. 1993. Knowing when to stop: Investigating the nature of punctuation. *Language and Communication*, 13(1):27–43.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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.

Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270.

Alan Hirvela, Alexander Nussbaum, and Herbert Pierson. 2012. ESL students' attitudes toward punctuation. *System*, 40(1):11–23.

Radu Tudor Ionescu, Marius Popescu, and Aoife Cahill. 2014. 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*, pages 1363–1373. Association for Computational Linguistics.

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. Association for Computing Machinery.

Shibamouli Lahiri and Rada Mihalcea. 2013. Using n-gram and word network features for native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 251–259. Association for Computational Linguistics.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Building Educational Applications Using NLP*, pages 62–75. Association for Computational Linguistics.

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

Ilia Markov, Efstathios Stamatatos, and Grigori Sidorov. 2017b. Improving cross-topic authorship attribution: The role of pre-processing. In *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing*. Springer, in press.

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

Nick Moore. 2016. What's the point? the role of punctuation in realising information structure in written english. *Functional Linguistics*, 3(1):6.

Ryo Nagata and Edward Whittaker. 2013. Reconstructing an indo-european family tree from non-native english texts. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1137–1147. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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.

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

Joel Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2585–2602. The COLING 2012 Organizing Committee.

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 Building Educational Applications Using NLP*, pages 48–57. Association for Computational Linguistics.

Oren Tsur and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16. Association for Computational Linguistics.

Sze-meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61. Association for Computational Linguistics.