Native Language Identification and Reconstruction of Native Language Relationship Using Japanese Learner Corpus

Mitsuhiro Nishijima[1,2] and **Ying Liu**[1]
[1]Department of Chinese Language and Literature, Tsinghua University, China
[2]Department of Industrial Engineering and Economics, Tokyo Institute of Technology, Japan
nishijima.m.ae@m.titech.ac.jp, yingliu@tsinghua.edu.cn

Abstract

We investigated the influence of the native language (L1) and its relationship preserved in second language (L2) Japanese texts by machine learning methods. We firstly performed native language identification (NLI) on L2 Japanese texts and proposed the character type as a new Japanese-specific feature in NLI. Combining it with other features increases the identification accuracy, and we obtained the maximum accuracy of 72.2%. We also confirmed the L1 relationship from the confusion matrix of NLI. To investigate the L1 relationship in more detail, we secondly applied hierarchical clustering to Japanese texts. Based on the results, we proposed a new hypothesis about the L1 relationship preserved in L2 texts.

1 Introduction

When we write a text in a second language (L2), the native language (L1) has more or less positive or negative impact on the L2 text, and traces of the L1 remain on the text. Therefore, many studies have been conducted to identify the L1 of the writer from L2 texts using machine learning methods—this is native language identification (NLI).

NLI can be applied to automatic error correction on texts written by L2 learners. It has been pointed out that the quality of the correction can be improved by taking the writers' L1 into account (Rozovskaya and Roth, 2011). If the writers' L1 is automatically identified by NLI, then the correction system can provide better quality feedback to the L2 learners. NLI can also be regarded as a kind of author profiling, and it is expected to use for criminal investigations and marketing (Estival et al., 2007). The number of L2 Japanese learners is increasing every year (Japan Foundation, 2018). Therefore, it is important to extend NLI, which has been conducted mainly on L2 English, to L2 Japanese.

In general, it can be observed from the identification results of NLI that L1s sharing a geographical or genealogical relationship are easily misidentified. It means that L2 texts preserve information not only on L1 itself but also on the L1 relationship. Some studies have successfully reconstructed the L1 relationship from L2 English texts (see Section 2 in detail). However, all of them used L2 English texts, and thus whether the preservation of the L1 relationship on L2 texts depends on the L2 has not been discussed yet. Although Nagata (2014) discussed the dependence on L1, the argumentation was incomplete.

Based on the above problems, we firstly perform NLI on L2 Japanese texts in Section 3. We propose the character type as a Japanese-specific feature in NLI. We confirm that combining it with other features improves the identification accuracy. The maximum accuracy reaches 72.2% against a baseline of 8.3%. We also confirm the L1 relationship from the confusion matrix of NLI.

Then, we apply hierarchical clustering to Japanese texts in Section 4. The results show that in most cases, Asian and European languages formed a cluster, respectively. In addition, the genealogical relationship within European languages is also preserved, though to a limited extent. Based on the re-

sults, we propose a new hypothesis which is a modification of the claim by Nagata (2014).

2 Related Work

2.1 Native Language Identification

The NLI researches have mainly focused on L2 English texts. The climax of the NLI research is NLI Shared Task held in 2013. Tetreault et al. (2013) reported that word, part of speech (POS), and character n-grams were widely used as features, and the support vector machine (SVM) was widely used as a classifier.

Considering the application of NLI to other fields, it is also important to be able to identify L1s from L2 texts other than L2 English. Therefore, researchers started to verify whether NLI techniques used for English L2 texts were also valid for other L2 languages, such as Arabic, Chinese, Finnish, Norwegian, Spanish, Italian, Portuguese, and Russian (Malmasi et al., 2015; Malmasi and Dras, 2017; Malmasi et al., 2018; Remnev, 2019). On the other hand, there is no NLI research on L2 Japanese as long as the authors know.

In addition, few NLI studies have used language-specific features to improve the accuracy of NLI. There are indeed some NLI studies on L2 English that used English-specific features. For example, Wond and Dras (2009) used English-specific errors such as subject-verb disagreement as features. However, NLI studies on L2 languages other than L2 English have only used the features found to be effective in NLI studies on L2 English. On the other hand, several studies about stylometry and author identification have claimed the effectiveness of language-specific features, such as n-gram based on the finals of the sentence-final character in Chinese (He and Liu, 2014) and syllabic writings usage patterns in Japanese (Maeshiro et al., 2014). Therefore, there is still room to utilize language-specific features to increase the identification accuracy in NLI.

2.2 Preservation of the L1 Relationship

We can observe the L1 relationship from the tendency of misidentification in NLI. For example, Wond and Dras (2009) showed that Indo-European languages were more easily misidentified with each other than non-Indo-European languages. The geographical relationship, as well as the genealogical relationship, can lead to misclassification. For example, Gebre et al. (2013) reported that geographically close languages such as Hindi and Telugu, which are widely used in India, were easy to be misidentified with each other. The same was true of Japanese, Korean, and Chinese mainly used in East Asia. Note that Hindi and Telugu, plus Japanese, Korean, and Chinese belong to different language families, or at least there is no consensus that they belong to the same language family. NLI studies on L2 Norwegian, Portuguese, and Russian also reported similar phenomena.

Several studies attempted to reconstruct the L1 relationship from L2 English texts. Nagata and Whittaker (2013) performed hierarchical clustering on English texts written by L2 English learners of 11 Indo-European languages. As a result, languages belonging to the Italic, Germanic, and Slavic branches formed a cluster, respectively. However, the L1s and the L2 belong to the same language family, which could contribute to the preservation of the L1 relationship in L2 texts. Therefore it was not clear whether the preservation of the L1 relationship in L2 texts does not depend on the L1. To resolve the problem, Nagata (2014) performed hierarchical clustering on English texts written by Asian L2 English learners and L1 English speakers. As a result, mainland China and Taiwan, Japan and South Korea, and Thailand and Indonesia, which all belong to the expanding circle of Kachru's three circles of English (Kachru, 1992), formed a cluster, respectively. Based on the result, he claimed that "the preservation of language family relationship universally holds in the expanding circle of English."[1] Other studies such as Berzak et al. (2014), Nastase and Strapparava (2017), and Rabinovich et al. (2018) also attempted to reconstruct the L1 relationship from L2 English texts.

However, all of these focused on L2 English. Thus they did not discuss whether the preservation of the L1 relationship depended on L2 at all[2]. The dependence on L1 was discussed in Nagata (2014), but there seemed to be some problems in the study.

[1] The underline is based on Nagata (2014).

[2] However, the confusion matrices in NLI seem to provide useful insights into this issue.

Indeed, it makes sense that Japan and South Korea formed a cluster because of the geographical closeness and the similarities in grammar and vocabulary between them. However, the language commonly used both in mainland China and in Taiwan is Chinese. Mainland China and Taiwan clustered could be based on a stronger relationship that the learners' L1s are the same rather than they belong to the same language family. Nagata also explained that Thailand and Indonesia were clustered because Thai has a relationship with the Austronesian family, to which Indonesian belongs. However, it is also true that the two languages belong to different language families, and one could argue that Thailand and Indonesia formed a cluster as "leftovers" after mainland China and Taiwan plus Japan and South Korea had formed a cluster. Therefore, it seems that Nagata (2014) does not completely remove the problem in Nagata and Whittaker (2013).

3 Japanese Native Language Identification

In this section, we perform NLI on L2 Japanese texts. We propose the character type as a Japanese-specific feature and confirm its effectiveness. We also check whether the confusion matrix in NLI reflects the L1 relationship.

3.1 Data and Method

(1) Data

We used *International Corpus of Japanese as a Second Language*[3] as a Japanese learner corpus in this paper. Among the corpus, we specifically used the texts produced in the story-writing task in which subjects described the story of two comic strips. The two texts produced from the comic strips were combined for each subject and treated as one text. In this section, we used data produced by L2 Japanese learners with twelve L1 backgrounds who studied Japanese in a foreign classroom. Table 1 shows the correspondence among their L1, its language family, and the country or region where they live (Sakoda et al., 2020). Because we also use texts produced by L1 Japanese speakers in Section 4, the information about Japanese is included in Table 1. Since there are 50 texts for each of the nine languages other than Korean, Chinese, and English, we used all of them. On the other hand, since the number of texts produced by L2 Japanese learners of Korean, Chinese, and English is more than 50, we selected 50 texts for each of them. In other words, we prepared 50 texts for each L1, totaling 600 texts. We did not strictly control proficiency in this study because Nagata (2014) claimed that the preservation of the L1 relationship was independent of English proficiency. We performed the morphological analysis and the dependency structure analysis for these texts using the Japanese natural language processing library GiNZA[4].

(2) Classifier

Many existing NLI studies selected linear SVM as a classifier and reported its higher accuracy compared to other classifiers (Malmasi, 2016). Therefore, we also used linear SVM as a classifier in this paper. We performed 10×10 nested cross-validation. The hyperparameter C of linear SVM took a value from $10^{-1}, 10^{0}, \ldots, 10^{7}$, and was determined by the grid search in each inner 10-fold cross-validation. Then, the average of the 10 accuracies obtained in the outer 10-fold cross-validation was used as the final accuracy.

(3) Linguistic Features

The linguistic features and their explanations on which we focused in creating a feature vector for each text are as follows. In the following, "1–n-gram" refers to everything from 1-gram to n-gram. The value of n for each feature was set so that the accuracy would be the highest. The accuracy for each feature is shown in square brackets.

① The lemma of morphemes 1–5-gram [69.2%]
② The character 1–4-gram [71.5%]
③ The lemma of particles and auxiliary verbs 1–5-gram: corresponding to function words in English. [40.7%]
④ POS 1–5-gram: the deepest level of POS tags. [50.5%]
⑤ The dependency label [30.2%]
⑥ The triple of the dependency label and the two lemmas [63.0%]

[3] https://chunagon.ninjal.ac.jp/static/ijas/about.html

[4] https://megagonlabs.github.io/ginza/

Table 1: The correspondence among L1, its language family, and the country or region where the subjects with the L1 background live

L1	Language family	Country or Region
Korean (KOR)	unknown (Altaic family?)	South Korea
Chinese (CHI)	Sino-Tibetan family	mainland China and Taiwan
Vietnamese (VIE)	Austroasiatic family	Vietnam
Thai (THA)	Tai-Kadai family	Thailand
Indonesian (IND)	Austronesian family	Indonesia
Turkish (TUR)	Altaic family	Turkey
Hungarian (HUN)	Uralic family	Hungary
Russian (RUS)	Indo-European family, Slavic branch	Russia
German (GER)	Indo-European family, Germanic branch	Germany and Austria
English (ENG)	Indo-European family, Germanic branch	US,UK, Australia, and New Zealand
French (FRE)	Indo-European family, Italic branch	France
Spanish (SPA)	Indo-European family, Italic branch	Spain
Japanese (JPN)	unknown (Altaic family?)	Japan

⑦ The triple of the dependency label and the two parts of speech: in which the lemmas in feature ⑥ are replaced with their parts of speech. [45.7%]

⑧ The character type: the frequency of *kanji*, *hiragana*, and *katakana*. [14.2%]

Feature ① to ⑦ have already been widely used in many NLI studies (Malmasi, 2016). In addition to them, We propose the character type as a Japanese-specific feature in NLI. The Japanese writing system is mainly composed of three character types: *kanji*, *hiragana*, and *katakana*. The rationale for using it as a feature is that L2 Japanese learners from countries using the Chinese character tend to use Sino-Japanese words written in *kanji* (Sakoda, 2020). We counted the absolute frequency of each feature to create a feature vector.

There are two possible ways to combine multiple feature vectors:

Method 1: Each feature vector is $l2$-normalized first and then connected.

Method 2: Each feature vector is connected first and then $l2$-normalized.

We created all the combined feature vectors for all cases ($(2^8 - 1) \times 2 = 510$ cases) and measured the identification accuracy.

3.2 Experiment

The result of the experiment is shown in Table 2, where ①–⑧ represent the eight features of Section 3.1, part (3). Due to paper limitations, using the character 1–4gram, which had the highest accuracy of feature ① to ⑧, as a baseline, we only show the eight feature combinations of 510 cases that had a higher accuracy than 71.5% in Table 2. We can see from Table 2 that we achieved the highest accuracy of 72.2%.

Table 2: The identification accuracy (%) of SVM

①	②	③	④	⑤	⑥	⑦	⑧	connection method	Acc.
✓	✓			✓	✓		✓	1	72.2
	✓			✓			✓	2	72.0
	✓				✓	✓	✓	2	72.0
✓	✓				✓	✓	✓	2	72.0
	✓			✓	✓	✓	✓	2	71.8
✓	✓			✓	✓	✓	✓	2	71.8
	✓				✓		✓	2	71.7
✓	✓					✓	✓	2	71.7
				Baseline 1 : only ②					71.5
				Baseline 2 : random					8.3

It is noteworthy that the character type, which itself only has the accuracy of 14.2%, appears in all combinations in Table 2. It implies that the character type can increase the accuracy when combined with other features. In fact, we fixed the features other than the character type as shown in Table 2 and compared the accuracy with and without the character type. Table 3 shows its result, and we can see from Table 3 that the accuracy increased by up to 1.3%. We also performed a binomial test on the sum of eight 2×2 cross-tabulation tables represent-

Table 3: The comparison of the accuracy (%) with and without the character type (CType) (Because the values were rounded off after subtraction, the third column is not necessarily equal to the values obtained by subtracting the second column from the first column.)

with CType	w/o CType	Diff.
72.2	71.3	0.8
72.0	70.8	1.2
72.0	71.5	0.5
72.0	71.5	0.5
71.8	70.5	1.3
71.8	71.3	0.5
71.7	71.3	0.3
71.7	70.8	0.8

Table 4: The confusion matrix corresponding to the combination with the highest accuracy in Table 2 (Cells with a frequency of three times and more are painted gray.)

		Predicted L1											
		KOR	CHI	VIE	THA	IND	TUR	HUN	RUS	GER	ENG	FRE	SPA
True L1	KOR	47	0	0	0	1	0	0	1	1	0	0	0
	CHI	0	40	5	1	0	1	0	1	1	1	0	0
	VIE	0	3	39	3	1	0	0	2	0	0	0	2
	THA	0	1	7	33	3	0	0	1	3	2	0	0
	IND	0	2	1	1	42	0	0	1	0	1	1	1
	TUR	0	0	0	0	0	46	2	1	0	0	1	0
	HUN	0	0	1	0	0	3	33	6	0	2	1	4
	RUS	0	0	1	0	0	1	5	36	0	3	1	3
	GER	2	1	0	0	1	1	6	1	24	7	4	3
	ENG	1	1	1	1	1	1	6	0	7	23	3	5
	FRE	1	0	0	1	2	1	2	2	4	3	31	3
	SPA	1	1	1	1	0	1	1	2	0	3	0	39

ing whether the identification was correct or not with and without the character type. The result was statistically significant with a p-value of 1.22×10^{-2}.

The feature diversity can explain the fact that combining the character type with other features increased the accuracy. Malmasi and Cahill (2015) pointed out that combining features with different properties can increase the accuracy. In our case, the character type is different from typical lexical features such as the character n-gram, and it is also different from syntactic features such as the dependency. Therefore, it can be inferred that the information captured by the character type complemented the information captured by other features, and consequently the accuracies were increased.

One of the purposes of Japanese NLI is to check whether the confusion matrix reflects the L1 relationship. We used the combination with the highest accuracy in Table 2. In the nested cross-validation, a total of 10 confusion matrices obtained from the outer 10-fold cross-validation were added up to yield the matrix shown in Table 4. We can see from Table 4 that Hungarian, Russian, German, English, French, and Spanish were easily misidentified with each other. Russian, German, English, French, and Spanish belong to the Indo-European family. Although Hungarian is not an Indo-European language but a Uralic language, Hungary is also a European country. Also, we can find that Chinese, Vietnamese, Thai, and Indonesian were also easily misidentified. Although the language families they belong to are different from each other, these languages are mainly spoken in Asian countries. Therefore, we can say that the confusion matrix for L2 Japanese texts also reflects the L1 relationship, i.e., the conflict between Asian and European languages.

4 Reconstruction of the L1 Relationship

In this section, we investigate whether the L1 relationship is preserved in L2 Japanese texts by hierarchical clustering.

4.1 Data and Method

(1) Data

We again used the texts prepared in Section 3. However, we target only 11 languages, excluding Turkish, as L1s. Since Turkey is located on the border between Asia and Europe, from a geographical point of view, Turkish can neither be considered an Asian nor a European language. Also, Table 4 shows that Turkish was not so much misidentified with other languages. For this reason, Turkish was excluded from the experiments in this section. In addition, we used 50 story-writing texts produced by L1 Japanese speakers in this section.

(2) Method for Reconstructing the L1 Relationship

The method of this paper followed the study by Nagata and Whittaker (2013) and Nagata (2014). First, a morpheme whose first level of the POS tag was not a particle nor an auxiliary verb was replaced with the deepest level of the POS tag. A morpheme

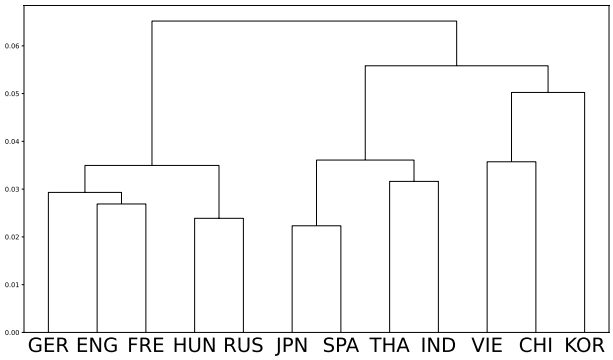

Figure 1: Dendrogram (600 texts, n=1, and threshold=0)

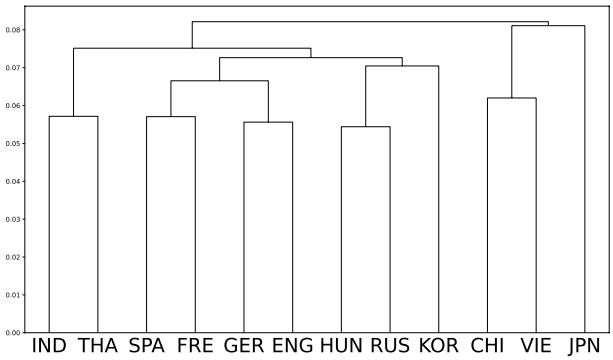

Figure 2: Dendrogram (600 texts, n=4, and threshold=2)

whose first level of the POS tag was a particle or an auxiliary verb was replaced with the lemma of the morpheme. Next, special tags "BOS" and "EOS" were added to the beginning and end of each sentence, respectively. In this paper, BOS, EOS, POS tags, particles, and auxiliary verbs are collectively called "quasi-POS."

For each L1 group, an absolute frequency vector of quasi-POS n-grams was created from quasi-POS texts. A feature item was deleted when the minimum absolute frequency of the feature item across the twelve L1 groups is less than a threshold value. We created a relative frequency vector from the new absolute frequency vector obtained by the above operation. This relative frequency vector was used for hierarchical clustering[5]. The Euclidean distance was used to measure the similarity between each data, and the group average method was used to measure the similarity between clusters.

4.2 Result

First, the result for the case where n=1 and threshold=0 (i.e., feature items were not deleted at all) is shown in Figure 1. We can see from Figure 1 that, except for Spanish, European languages were clustered on the left, and Asian languages were clustered on the right. However, the preservation of the genealogical relationship within European languages cannot be confirmed. For example, English, which belongs to the Germanic branch, was not clustered with German, which also belongs to the Germanic branch first; instead, English was clustered with French, which belongs to the Italic branch first. One might wonder why Spanish was clustered with Japanese. We speculate that the reason is due to the small number of texts used in this experiment, as described in Section 4.3.

The result for the case where n=4 and threshold=2 is shown in Figure 2. Unlike Figure 1, Asian and European languages were not dichotomized. However, all the European languages, including Spanish, formed a cluster surrounded by Indonesian and Thai. In particular, Spanish and French, which belong to the Italic branch, and German and English, which belong to the Germanic branch, formed a cluster, respectively. This indicates that the genealogical relationship within European languages was reflected in the result of hierarchical clustering.

However, very few cases showed the same trend as Figure 2. When n was varied from 1 to 4 and threshold from 0 to 3, only the case where n=4 and threshold=1 showed the same trend as Figure 2. Almost all other results showed the same trend as Figure 1.

4.3 Discussion

It was confirmed from the experiments of Section 4.2 that in most cases, the conflict between Asian and European languages was preserved in L2 Japanese texts. It was also confirmed that the genealogical relationship within European languages was also preserved, though to a limited extent. These results suggest that the preservation of the L1 relationship in L2 texts is independent of the L2. It should also be noted that Japanese and the European

[5]Nagata and Whittaker (2013) also experimented with a method based on a probabilistic language model, but they reported that the result was similar to that of the vector-based method. Therefore, the vector-based method was only used in this paper.

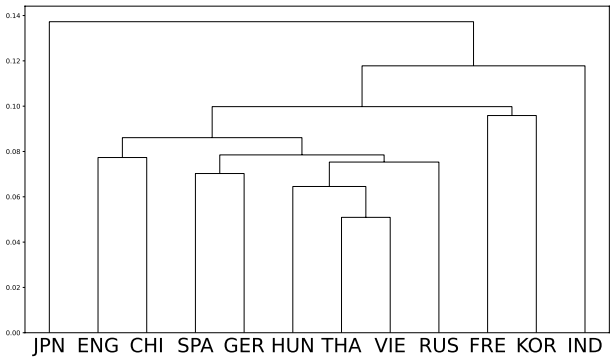

Figure 3: Dendrogram (180 texts, n=4, and threshold=2)

languages used in this paper are different in terms of language family. This fact has removed the problem in Nagata and Whittaker (2013). Therefore, these results seem to support the claim of Nagata (2014). However, the genealogical relationship of L1s was not so strongly reflected in the L2 Japanese texts used in this study. Therefore it is difficult to *fully* support Nagata (2014)'s claim only based on the results of this study.

Why, then, was the genealogical relationship of L1s not so strongly reflected in the results? One factor may be the small number of texts: Nagata and Whittaker (2013) and Nagata (2014) used at least 200 texts per L1, while our study used only 50 texts per L1. In fact, we confirmed through the experiment described below that the number of texts might affect the clustering results. We selected 15 for each L1 (180 in total) out of 600 texts[6] and experimented under the same condition as Figure 2 except the number of texts. The result is shown in Figure 3. Figure 3 is different from Figure 2 and we cannot almost find the geographical nor genealogical relationship.

Another possible explanation may come from Kellerman's psychotypology. Kellerman (1979) argues that the closer learners feel their L1 is to L2, the more likely L1 is to transfer to L2 because they create morphological and syntactic correspondences between L1 and L2 creatively. Let us assume that the genealogical relationship between L1 and L2 is directly related to their closeness that learners perceive. Then, when L2 English learners of Indo-European languages produce English, "mother tongue interference is so strong that the relations between members of the Indo-European language family are preserved in English texts written by Indo-European language speakers" (Nagata and Whittaker, 2013) because both of their L1 and English belong to the Indo-European family. At the same time, Kellerman's claim is also consistent with the present results: when L2 Japanese learners of European languages produce Japanese, the L1 does not transfer to L2 Japanese so strongly because of their linguistic distance, and thus the genealogical relationship of L1s was not so strongly confirmed. Thus, we propose the following claim as a new hypothesis, which is a modification of the claim of Nagata (2014): *the L1 relationship is preserved in L2 texts whatever L1 or L2 is, but the more distant L1 is from L2, the weaker the preservation of the L1 relationship is.*

The hypothesis is more precise than the claim of Nagata (2014) in that it refers to the dependence on L2 and the strength of the preservation. On the other hand, it is rougher than the claim of Nagata (2014) in that it does not mention what part of the three circles the relationship holds in.

5 Conclusions and Future Work

In this paper, we performed two major experiments to confirm the influence of L1 and its relationship in L2 Japanese texts. First, in Section 3, we performed a Japanese NLI. We proposed the character type as a Japanese-specific feature in NLI. It was confirmed that the accuracy was improved by combining it with other features. We were able to identify L1 from the twelve L1s with maximum accuracy of 72.2%. Moreover, we were able to confirm the conflict between Asian and European languages from the confusion matrix. Next, in Section 4, we investigated the L1 relationship preserved in Japanese texts using hierarchical clustering. Based on the results, we proposed the new hypothesis that the L1 relationship is preserved in L2 texts whatever L1 or L2 is, but the more distant L1 is from L2, the weaker the preservation of the L1 relationship is.

There are still two issues to be addressed. First, we mentioned that the small number of texts might

[6]Specifically, we selected the texts written by L2 Japanese learners with intermediate Japanese proficiency and L1 Japanese speakers with no experience living overseas.

also affect the results in Section 4. In light of the scale of the existing Japanese learner corpus, it is difficult to use more L2 Japanese texts collected under the same condition. In addition, the reason why we could not refer to Kachru's three circles in the proposed hypothesis is that Japan (where Japanese is spoken as the *de facto* national language) is the only country that does not belong to the expanding circle of Japanese in the experiment. Therefore, it is necessary to re-examine our claim and the claim of Nagata (2014) using more texts, including texts written by subjects from some inner circles and outer circles.

Acknowledgments

This work is supported by Project of Humanities and Social Sciences of Ministry of Education in China "Automatic extraction and Research of stylistic features" (17YJAZH056), Tsinghua University Humanities and Social Sciences Revitalization Project (2019THZWJC38), and Project of Baidu Netcom Technology Co., Ltd. Open source Course and Case Construction Based on the Deep Learning Framework PaddlePaddle (20202000291).

References

Alla Rozovskaya and Dan Roth. 2011. Algorithm Selection and Model Adaptation for ESL Correction Tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 924–933.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving Native Language Identification with TF-IDF Weighting. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 216–223.

Braj Bihari Kachru. 1992. *The Other Tongue: English Across Cultures*, chapter 19. 2nd. Ed. University of Illinois Press.

Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author Profiling for English Emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272.

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.

Eric Kellerman. 1979. Transfer and Non-Transfer: Where Are We Now? *Studies in Second Language Acquisition*, 2(1):37–57.

Japan Foundation. 2018. Survey Report on Japanese-Language Education Abroad 2018. `https://www.jpf.go.jp/j/project/japanese/survey/result/dl/survey2018/Report_all_e.pdf` (accessed: 2021-10-07).

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.

Kumiko Sakoda. 2020. *Second Language Acquisition Research to Take Advantage of the Japanese Education*（日本語教育に生かす第二言語習得研究）. Revised Ed. ALC Press Inc. (In Japanese)

Kumiko Sakoda, Shin'ichiro Ishikawa, and Jaeho Lee. 2020. *An Introduction to Japanese Learner Corpus, I-JAS: Application for Teaching and Research*（日本語学習者コーパス I-JAS 入門：研究・教育にどう使うか）. Kurosio Publishers. (In Japanese)

Nikita Remnev. 2019. Native Language Identification for Russian. In *2019 International Conference on Data Mining Workshops*, pages 1–7.

Ryo Nagata. 2014. Language Family Relationship Preserved in Non-Native English. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1940–1949.

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.

Shervin Malmasi. 2016. Native Language Identification: Explorations and Applications. Ph.D. thesis, Macquarie University.

Shervin Malmasi and Aoife Cahill. 2015. Measuring Feature Diversity in Native Language Identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 49–55.

Shervin Malmasi, Iria del Río, and Marcos Zampieri. 2018. Portuguese Native Language Identification. In *Proceedings of International Conference on the Computational Processing of Portuguese*, pages 115–124.

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 Innovative Use of NLP for Building Educational Applications*, pages 62–75.

Shervin Malmasi and Mark Dras. 2017. Multilingual Native Language Identification. *Natural Language Engineering*, 23(2):163–215.

Shervin Malmasi, Mark Dras, and Irina Temnikova. 2015. Norwegian Native Language Identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 404–412.

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.

Tetsuya Maeshiro, Hideo Joho, Shin'ichi Nakayama, and Mai Hayakura. 2014. Author Identification of Japanese Texts Based on Notations and Syllabic Writings Usage Patterns. *Journal of Japan Society of Information and Knowledge*, 24(3):342–364. (In Japanese)

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 2702–2707.

Xiangqing He and Ying Liu. 2014. Mining Stylistic Features of Rhythm and Tempo Based on Text Clustering. *Journal of Chinese Information Processing*, 28(6):194–200. (In Chinese)

Yevgeni Berzak, Roi Reichart, and Boris Katz. 2014. Reconstructing Native Language Typology from Foreign Language Usage. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 21–29.