Language Family Relationship Preserved in Non-native English

Ryo Nagata
Konan University
8-9-1 Okamoto, Higashinada, Kobe, Hyogo 658-8501, Japan
nagata-coling@hyogo-u.ac.jp

Abstract

Mother tongue interference is the phenomenon where linguistic systems of a mother tongue are transferred to another language. Recently, Nagata and Whittaker (2013) have shown that language family relationship among mother tongues is preserved in English written by Indo-European language speakers because of mother tongue interference. At the same time, their findings further introduce the following two research questions: (1) Does the preservation universally hold in non-native English other than in English of Indo-European language speakers? (2) Is the preservation independent of proficiency in English? In this paper, we address these research questions. We first explore the two research questions empirically by reconstructing language family trees from English texts written by speakers of Asian languages. We then discuss theoretical reasons for the empirical results. We finally introduce another hypothesis called *the existence of a probabilistic module* to explain why the preservation does or does not hold in particular situations.

1 Introduction

Transfer of linguistic systems of a mother tongue to another language, namely *mother tongue interference*, is often observable in the writing of non-native speakers. The reader may be able to determine the mother tongue of the writer of the following sentence from the underlined article error: *The alien wouldn't use my spaceship but the hers*. The answer would probably be French or Spanish; the definite article is allowed to modify possessive pronouns in these languages, and the usage is sometimes negatively transferred to English writing.

Researchers in corpus linguistics including Swan and Smith (2001), Aarts and Granger (1998), and Altenberg and Tapper (1998) have been working on mother tongue interference to reveal overused/underused words, part of speech (POS), or grammatical items. Recently, Nagata and Whittaker (2013) have shown that language family relationship between mother tongues is preserved in English written by Indo-European language speakers; because of the preservation, one can reconstruct a language family tree similar to the canonical Indo-European family tree (Beekes, 2011; Ramat and Ramat, 2006) from their English writings. They have further revealed factors contributing to the preservation of the language family relationship, which they show is useful for related natural language processing (NLP) tasks such as grammatical error detection/correction and native language identification (Wong and Dras, 2009).

At the same time, their findings further introduce the following two research questions: (1) Does the preservation universally hold in non-native English? (2) Is the preservation independent of proficiency in English? The results (Nagata and Whittaker, 2013) for English written by Indo-European language speakers suggest that the answer to question (1) is *yes*. Based on this, we hypothesize that:

Hypothesis I: The preservation of language family relationship universally holds in non-native English.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1940–1949, Dublin, Ireland, August 23-29 2014.

However, one can counter **Hypothesis I**, arguing that the preservation holds only in English written by Indo-European language speakers because Indo-European languages share large part of linguistic properties with English which is a member of the Indo-European languages, which contributes to the preservation. Apparently, this is not the case in languages in other language families. In these languages, other properties than language family relationship may be more dominant. Furthermore, Kachru's Three Circles of English (Kachru, 1992) raises a question. In Kachru's model, world Englishes are classified into the inner, outer, and expanding circles. The inner circle roughly corresponds to the traditional native speakers of English. The outer circle refers to the non-native varieties in the regions where English serves as a useful lingua franca. The expanding circle roughly corresponds to the other non-native speakers of English. Then, it would be difficult to answer question (1) for the outer circle of English (e.g., English in Hong Kong). For example, on one hand, English in Hong Kong is expected to have mother tongue interference from Chinese language. From this point of view, it is expected to have the family relationship with the Sino-Tibetan language family. On the other hand, one can point out that the outer circle of English should be closer to native English than the expanding circle of English (e.g., English in China) is. This implies that English in Hong Kong might have some other relationship with the members in the outer circle. For question (2), the answer is likely *no* considering that theoretically, the higher one's proficiency is, the closer to native English his or her English becomes; it would be difficult to distinguish between native English and English of non-native speakers whose proficiency is very high. With this reason, we hypothesize that:

Hypothesis II: The preservation of language family relationship is dependent on proficiency in English.

In view of this background, we address these research questions in this paper. We first examine the two hypotheses empirically by reconstructing language family trees from English texts written by speakers of Asian languages, including the outer and expanding circles of English. If we can reconstruct language family trees similar to their canonical family trees from these English texts, it will be a good piece of evidence for **Hypothesis I**. Similarly, to examine **Hypothesis II**, we reconstruct a language family tree from the English texts using the information about their proficiency levels. If we cannot reconstruct language family trees similar to the canonical trees, **Hypothesis II** will be accepted. We then explore theoretical reasons for the empirical results. We finally introduce another hypothesis called *the existence of a probabilistic module* to explain why the preservation does or does not hold in particular situations.

The rest of this paper is structured as follows. Sect. 2 introduces the basic approach of this work. Sect. 3 and Sect. 4 examine **Hypothesis I** and **Hypothesis II**, respectively. Sect. 5 describes theoretical reasons for the experimental results.

2 Approach

2.1 Data Set

Through this paper, we use the International Corpus Network of Asian Learners of English (ICNALE) (Ishikawa, 2011) as the target data to examine the two hypotheses. ICNALE consists of English essays of the outer and expanding circles of English in Asia together with those of native speakers of English. Table 1 (a) shows the statistics on ICNALE.

In ICNALE, each essay, except native essays, is annotated with a proficiency level of the writer, ranging from A_2 (lowest) B1_1, B1_2, to B2+ (highest); Table 1 (b) shows the correspondence between these four proficiency levels and TOEIC scores. We use this information to examine **Hypothesis II**.

2.2 Method for Reconstructing Language Family Trees

We employ the method proposed by Nagata and Whittaker (2013) for reconstructing language family trees, which in turn is based on the method proposed by Kita (1999). In this method, each group of the essays in ICNALE is modeled by an n-gram language model. Then, agglomerative hierarchical clustering (Han and Kamber, 2006) is applied to the language models to reconstruct a language family tree. The distance used for clustering is based on a divergence-like distance between two language models that was originally proposed by Juang and Rabiner (1985).

Category	# of essays	# of tokens
Native	400	88,792
Outer Circle		
Hong Kong	200	46,111
Pakistan	400	93,100
Philippines	400	96,586
Singapore	400	96,733
Expanding Circle		
China	800	194,613
Indonesia	400	92,316
Japan	800	176,537
Korea	600	130,626
Thailand	800	176,936
Taiwan	400	89,736

(a) Statistics on ICNALE

Level	A2	B1_1	B1_2	B2+
Score	225-549	550-669	670-784	785+

(b) Correspondence between the Proficiency Levels and TOEIC Score

Table 1: Summary of ICNALE.

To explain the method in more detail, let us define the following symbols used in the method. Let D_i be a set of English texts where i denotes a mother tongue i. Similarly, let M_i be a language model trained using D_i.

To reduce the influences from the topics of the data set, we use an n-gram language model based on a mixture of word and POS tokens. In this language model, content words in n-grams are replaced with their corresponding POS tags. This greatly decreases the influence of the topics of texts. It also decreases the number of parameters in the language model.

To build the language model, the following three preprocessing steps are applied to D_i. First, texts in D_i are split into sentences. Second, each sentence is tokenized, POS-tagged, and mapped entirely to lowercase. For instance, the example sentence in Sect. 1 would give:

> the/DT alien/NN would/MD not/RB use/VB my/PRP$ spaceship/NN but/CC the/DT hers/PRP ./.

Finally, words are replaced with their corresponding POS tags; for the following words, word tokens are used as their corresponding POS tags: coordinating conjunctions, determiners, prepositions, modals, predeterminers, possessives, pronouns, question adverbs. Also, proper nouns are treated as common nouns. At this point, the special POS tags *BOS* and *EOS* are added at the beginning and end of each sentence, respectively. For instance, the above example would result in the following word/POS sequence:

> BOS the NN would RB VB my NN but the hers . EOS.

Note that the content of the original sentence is far from clear while reflecting mother tongue interference, especially in *the hers*.

Now, the language model M_i can be built from D_i. We set $n = 3$ (i.e., trigram language model) and use Kneser-Ney (KN) smoothing (Kneser and Ney, 1995) to estimate its conditional probabilities.

The clustering algorithm used is agglomerative hierarchical clustering with the average linkage method. The distance[1] between two language models is measured as follows. The probability that M_i generates D_i is calculated by $\Pr(D_i|M_i)$. Note that

$$\Pr(D_i|M_i) \approx \Pr(w_{1,i}) \Pr(w_{2,i}|w_{1,i}) \prod_{t=3}^{|D_i|} \Pr(w_{t,i}|w_{t-2,i}, w_{t-1,i}) \quad (1)$$

[1]It is not a distance in a mathematical sense. However, we will use the term *distance* following the convention in the literature.

where $w_{t,i}$ and $|D_i|$ denote the tth token in D_i and the number of tokens in D_i, respectively, since we use the trigram language model. Then, the distance from M_i to M_j is defined by

$$d(M_i \to M_j) \equiv \frac{1}{|D_j|} \log \frac{\Pr(D_j|M_j)}{\Pr(D_j|M_i)}. \quad (2)$$

In other words, the distance is determined based on the ratio of the probabilities that each language model generates the language data. Because $d(M_i \to M_j)$ and $d(M_j \to M_i)$ are not symmetrical, we define the distance between M_i and M_j to be their average:

$$d(M_i, M_j) \equiv \frac{d(M_i \to M_j) + d(M_j \to M_i)}{2}. \quad (3)$$

Equation (3) is used to calculate the distance between two language models for clustering.

To sum up, the procedure of the language family tree construction method is as follows: (i) Preprocess each D_i; (ii) Build M_i from D_i; (iii) Calculate the distances between the language models; (iv) Cluster the language data using the distances; (v) Output the result as a language family tree.

3 Reconstructing Language Family Trees from Asian English

We used the whole ICNALE as the target data. We used a POS-tagger with the Penn Treebank Tagset (Santorini, 1990), which we had specially developed for analyzing non-native English; we trained it on native and non-native corpora we had manually annotated with POS tags, part of which is available to the public as the Konan-JIEM (KJ) learner corpus (Nagata et al., 2011). Then, we generated a cluster tree from the corpus data using the method described in Subsect. 2.2. We used the Kyoto Language Modeling toolkit[2] to build language models from the corpus data. We removed n-grams that appeared less than five times[3] in each subcorpus in the language models.

Fig. 1 shows the resulting cluster tree. The number at each branching node denotes in which step the two clusters were merged.

The cluster tree supports **Hypothesis I** that the preservation of language family relationship universally holds in non-native English. Although the detailed language family relationship is less well-known in these Asian languages than in the Indo-European languages, still the cluster tree shown in Fig. 1 reflects a rational interpretation of their language family relationship. In the cluster tree, Taiwanese and Chinese Englishes are first merged into a cluster. This perfectly agrees with the fact that their mother tongues are primarily Chinese and thus both should belong to the Sino-Tibetan language family. In turn, Japanese and Korean Englishes are merged into a cluster. Their mother tongues are said to be a member of the Altaic language family. Admittedly, it is still controversial whether the two languages belong to the Altaic language family or not. However, the current research often treats them as a member of the Altaic language family (Crystal, 1997). After Japanese and Korean Englishes, Thai and Indonesian Englishes are merged in to a cluster of which mother tongues belong to different language families; the former belong to the Thai language family while the latter mostly belong to the Austronesian language family. Having said that, it has been pointed out that Thai has some language family relationship with the Austronesian language family (Crystal, 1997). All these observations support **Hypothesis I**.

Interestingly, the cluster tree shown in Fig. 1 preserves, together with language family relationship, the three circles of English, namely, the inner (native), outer, and expanding circles of English with an exception of Pakistani English. This can be interpreted as that some other properties are more dominant than language family relationship in the outer circle of English. An implication from this is that we should not treat the outer and expanding circles as a group of non-native speakers of English but separately as different groups in the related NLP tasks such as grammatical error correction. For example, a method performing well on the outer circle of English (e.g., the NUS corpus (Dahlmeier et al., 2013)) does not necessarily perform equally well on the expanding circle of English (e.g., the CLC corpus) and vice versa. Similarly, a model trained on English written by Indo-European language speakers may perform

[2]The Kyoto Language Modeling toolkit: `http://www.phontron.com/kylm/`

[3]We found that the results were not sensitive to the value of frequency cutoff so long as we set it to a small number.

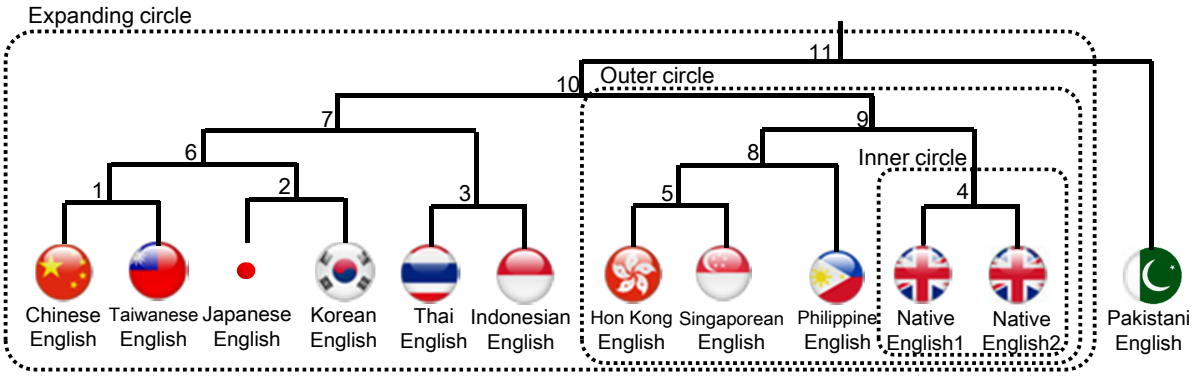


Figure 1: Cluster Tree Reconstructed from Asian Englishes (ICNALE).

better on Chinese English than a model trained on Hon Kong English does. Above all, the subtree for the outer circle of English is a piece of evidence that partly denies **Hypothesis I**.

We further reconstructed a clustering tree from the same data set using 5-gram language models so that the resulting clustering reflects longer-distance syntactic relations. Fig. 2 shows the resulting cluster tree, which reveals that the tree is almost the same as in Fig. 1 with an exception of the Philippine English.

After having observed all these, it would be rational to partly accept **Hypothesis I** and to modify it as follows:

Hypothesis I′: The preservation of language family relationship universally holds in the expanding circle of English.

4 Exploring Correlation between the Preservation and Proficiency

The simplest way to examine **Hypothesis II** would be clustering that uses only either high-proficiency or low-proficiency essays. However, it is not so straightforward because the distribution of each proficiency level varies depending on the English groups. Particularly, some of the 10 non-native Englishes contains no or very few low-proficiency essays[4].

As a simple solution, we first generated a clustering tree from only the high-proficiency essays (B1_2 and B2+) with the same conditions as in Sect. 3. As a more sophisticated solution, we created a new data set from ICNALE so that one of the two Englishes merged into a cluster in Fig. 1 consists of only low-proficiency essays and the other of only high-proficiency essays. For instance, we used only low-proficiency essays (A_2 and B1_1) for Chinese English and only high-proficiency essays (B1_2 and B2+) for Taiwanese English. Then, we generated another cluster tree from the new data set again with the same conditions as in Sect. 3. In addition, as a reference, we generated a cluster tree only using the information about the proficiency levels. In this clustering, we created a vector for each English whose elements and values corresponded to the four proficiency levels and the relative frequencies of the essays falling into the corresponding proficiency level[5]. In this method, we defined the distance for clustering by the Euclidean distance between two vectors.

The idea behind this experiment is as follows. If the preservation is completely independent of proficiency, we will obtain the exact same tree as in Fig.1 both from the only-high-proficiency data set and the high-low proficiency-paired data set. Otherwise, the cluster tree will result in a different form, similar to the one obtained by the vector-based method solely relying on the information about proficiency.

Fig. 1 and Fig. 3 show the cluster trees obtained from the only-high-proficiency data set and the high-low proficiency-paired data set, respectively. In the case of the only-high-proficiency data set, the resulting tree is the exact same as in the one generated from the original data set. Fig. 3 also shows that the cluster tree is very similar to that in Fig. 1. Besides, both tree are far from the cluster tree obtained by the

[4]For instance, Singapore English contains no low-proficiency essays (A2 and B1_1), and Philippine English 26 essays out of 400. See `http://language.sakura.ne.jp/icnale/` for the complete list of the distribution.

[5]We create vectors for the native English essays by setting 1.0 to the element corresponding to B2+ and 0.0 to the others because proficiency levels are not available for the native English essays in ICNALE.

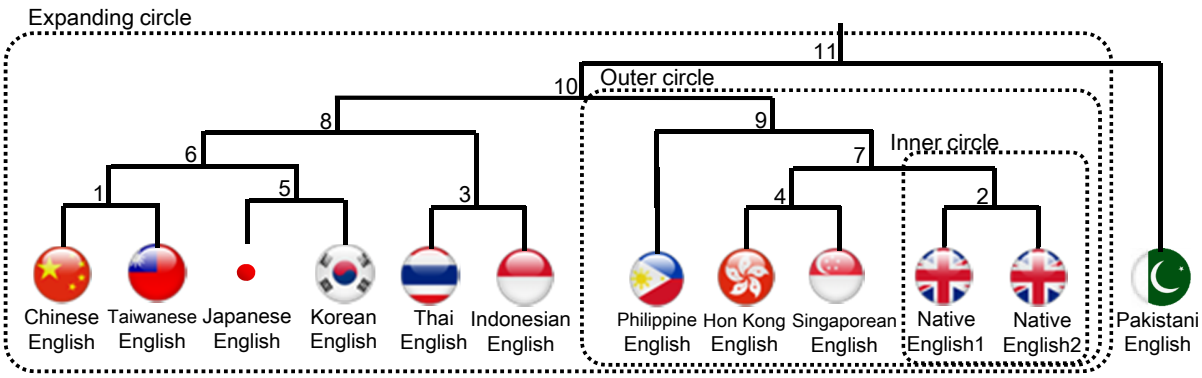


Figure 2: Cluster Tree Reconstructed from Asian Englishes (ICNALE) using 5-gram language models.

vector-based method solely relying on the information about proficiency as shown in Fig. 4. In summary, Fig. 1 to 4 show that the preservation of language family relationship holds in the expanding circle of English regardless of proficiency in English.

These results deny **Hypothesis II** that the preservation of language family relationship is dependent on proficiency in English. Contrary to our expectation, they support[6]:

Hypothesis II′: The preservation of language family relationship is independent of proficiency in English.

5 Discussion

The experiments show that the tree generation method relying on the distributions of word/POS sequences reconstructs from Asian Englishes cluster trees reflecting the family relationship in the Asian languages. These empirical findings, together with those about English written by Indo-European language speakers (Nagata and Whittaker, 2013), support **Hypothesis I′**.

In order to explain theoretically **Hypothesis I′**, we introduce another hypothesis called *the existence of a probabilistic module*, that is, that a probabilistic module that stores the distributional information exists in the human brain. We hypothesize that the probabilistic module consists of sets of probabilities where each set corresponds to a linguistic item which has arbitrariness in its use; the arbitrariness is expressed by means of the probabilities that one of the candidates allowed in the linguistic item is chosen in one's mother tongue. An example of such a linguistic item would be the position of adverb in English where the probabilities in this case represent how likely adverbs appear in certain positions (e.g., the beginning, middle, and end of a sentence). The probabilistic module is equipped with the values of the probabilities which are set according to one's mother tongue. To be precise, in our hypothesis, the probabilities are adapted as follows: (1) proto-languages had developed their values of the probabilities and handed them down to their descendants; (2) over the time, some of the values changed and the others remained unchanged; (3) in turn, the decedent languages handed their values of the probabilities to their descendants with the changes. An example of this would be as follows. The proto-Indo-European language handed down its values of the probabilities to, for example, the Proto-Germanic language and the Proto-Italic language with some changes in the values. Then the Proto-German language handed them down to the Germanic languages such as German and Dutch, again with some changes. So did the Proto-Italic language to the romance languages such as French and Italian. Therefore, the values of the probabilities in German should be more similar to those in Dutch than to those in French or Italian.

With this probabilistic module in the human brain, we can naturally explain the preservation of language family relationship. When non-native speakers use English, the candidates of the arbitrary linguistic items in English are chosen according to the probabilistic module adapted to their mother tongue.

[6]It would be worth while to see if **Hypothesis II** holds in the case of Indo-European Englishes. The difficult part is that there are only a few data annotated with proficiency levels.

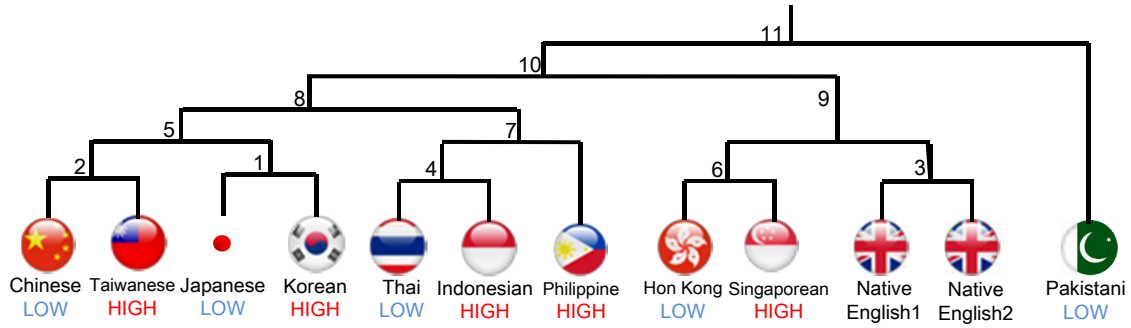


Figure 3: Cluster Tree Reconstructed from the High-low Proficiency-paired ICNALE Data Set (HIGH: high proficiency; LOW: low proficiency).

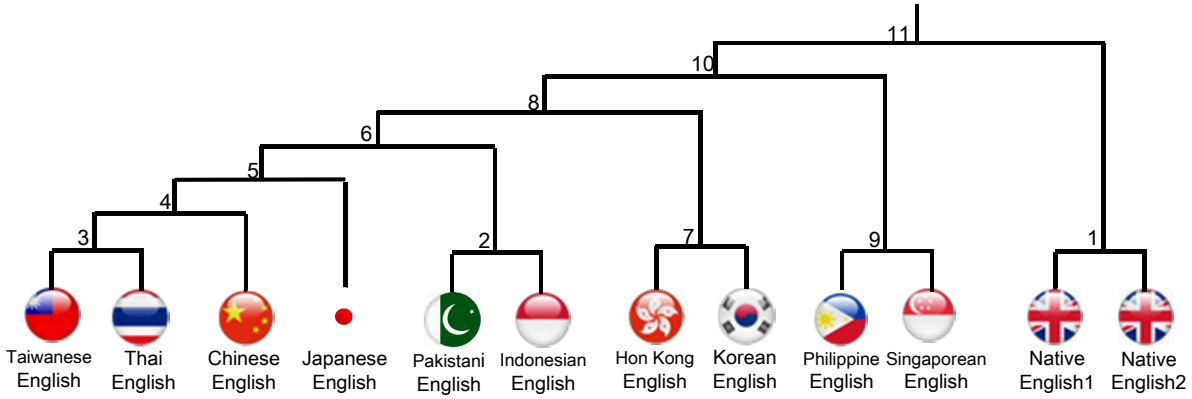


Figure 4: Cluster Tree Generated Based on Only Proficiency Levels.

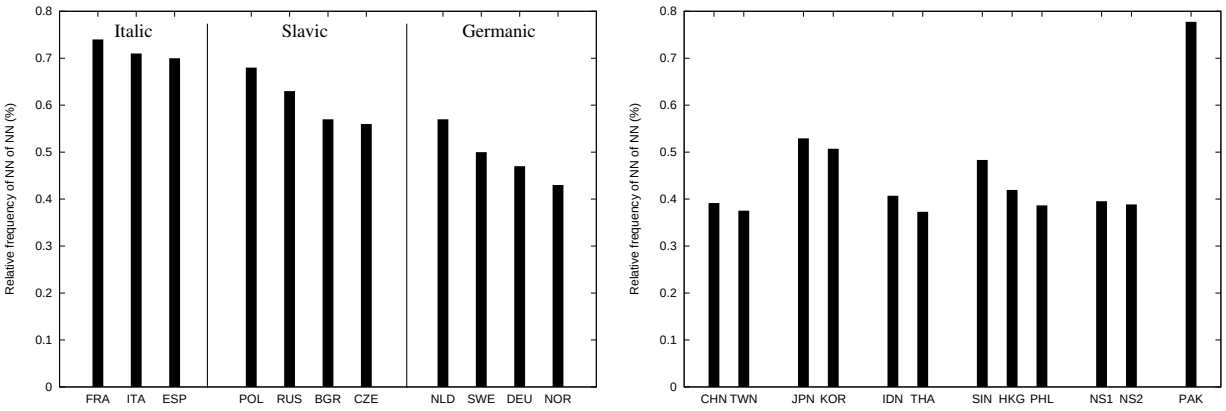
For example, speakers of languages which have a preference for sentence-beginning adverbs would also prefer sentence-beginning adverbs in English writing. Accordingly, the values of the probabilities are implicitly encoded in word/POS sequences such as *BOS RB ,* and *NN RB .*[7] in their English writings, and thus the tree generation method can recognize language family relationship as language family trees via the trigram language model. Provided that the probabilistic module exists in the human brain, this argument can be made about any mother tongues and the target language (not only English) as long as they have arbitrary linguistic items in their language systems, which should be the case in most languages.

This is of course another hypothesis and we need more data and evidence to examine the hypothesis. Nagata and Whittaker (2013) show some evidence that implies the existence of a probabilistic module. They reveal that Englishes written by Indo-European language speakers exhibit certain probabilistic patterns at least in the way of constructing noun phrases (NPs), adverb positions, and article use, reflecting the Italic, Germanic, and Slavic branches of the Indo-European family. Take as an example Fig. 5 (i) which shows frequencies of the trigram *NN of NN* in English written by Indo-European language speakers[8]. Here, note that English language has arbitrariness between the noun-noun compound and the *NN of NN* construction to form an NP (e.g., *education system* vs. *system of education*). Fig. 5 (i) reveals that speakers of the Italic languages (French, Italian, and Spanish) which have a preference for the *NN of NN* construction over the noun-noun compound exhibit relatively high frequencies of the trigram *NN of NN* in English writing. Conversely, speakers of the Germanic languages (Dutch, Swedish, German, and Norwegian) have a preference for the noun-noun compound over the *NN of NN* construction accordingly exhibit lower frequencies of the trigram *NN of NN*. In total, the frequencies roughly classify the 11 Englishes into three groups corresponding to the Italic, Slavic, and Germanic branches of the

[7]These two trigrams roughly correspond to adverbs at the beginning and end of a sentence, respectively.

[8]The ICLE corpus (Granger et al., 2009) was used to calculate the frequencies. The three letters such as FRA in Fig. 5 and Fig. 6 denote the ISO 31661 alpha-3 codes except NS1 (Native Speaker 1) and NS2 (Native Speaker 2).

Indo-European language family.



(i) Indo-European language speakers (ICLE) (ii) Asian language speakers (ICNALE)

Figure 5: Relative Frequency of *NN of NN* in English Texts Written by Non-native Speakers of English.

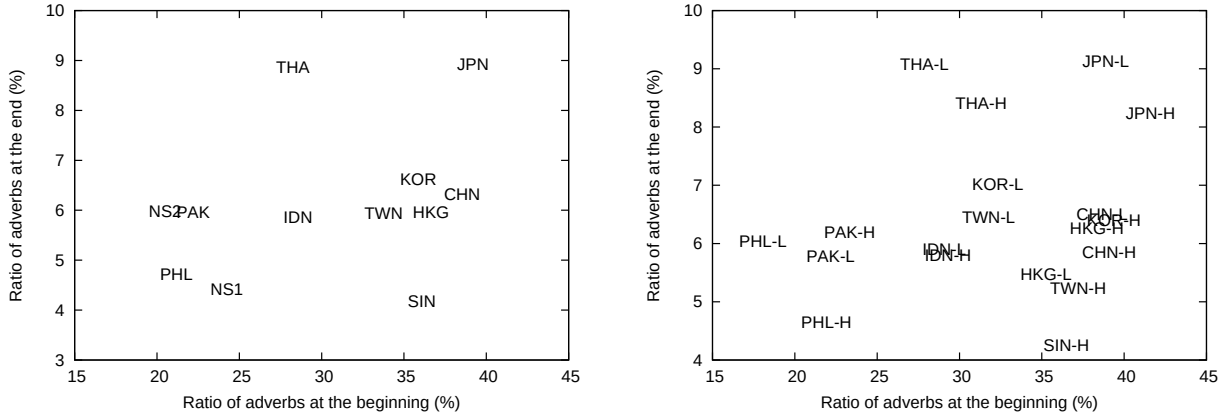
The data of Asian Englishes we used in the experiments exhibit similar tendencies. Fig. 5 (ii) shows frequencies of the trigram *NN of NN* for the Asian Englishes together with the native Englishes (denoted as NS1 and NS2). Fig. 5 (ii) reveals that the pairs of Englishes which share language family relationship each other exhibit similar frequencies of the trigram *NN of NN* as in Fig. 5. Furthermore, Fig. 6 (i) shows a similar tendency in the distribution of adverb positions. The horizontal and vertical axes of Fig. 6 correspond to the ratios of adverbs at the beginning and the end of sentences, respectively, in the Asian and native Englishes. It turns out that the pairs again tend to be located in near positions in the distribution. All of these imply the existence of the probabilistic module.

The probabilistic module also explains why the preservation is independent of proficiency. It is because the values of the probabilities in the probabilistic module will change quite slowly as one improves his or her proficiency. First of all, unlike grammatical errors, explicit feedback such as correction by teachers is not normally given to language learners in the case of the use of the arbitrary linguistic items since any choice among the candidates allowed in a linguistic item is normally correct, as in the adverb positions in English: *Already, I have done it.*, *I have already done it.*, and *I have done it already*, although each of which might have a slightly difference in meaning. Therefore, language learners have little opportunity to adapt the values of the probabilities in their probabilistic module to those in the target language in the first place. Even if feedback is given, it would still be difficult to do so considering that learners scarcely observe the values of the probabilities directly. This is why the values of the probabilities in the probabilistic module tend to be similar within a mother tongue regardless of one's proficiency in English. We can actually see this in Fig. 6 (ii). Fig. 6 (ii) shows the distribution of the ratios of adverbs at the beginning and the end of sentences in the high/low-proficiency essays in ICNALE where X-H and X-L denote high-proficiency and low-proficiency essays of X English, respectively (e.g., THA-H denotes the high-proficiency essays of Thai English). Fig. 6 (ii) reveals that Englishes of the same language speakers tend to remain in near positions regardless of the difference in proficiency.

All these observations would be a good place to start to explore the existence of the probabilistic module. The next step would be to name other arbitrary linguistic items concerning the probabilistic module, one of which for example might be the order of the main and subordinate clauses (e.g., *Because I did it, I did it.* vs *I did it because I did it.*), and then one can reveal their values (probabilities) depending on mother tongues.

6 Conclusions

In this paper, we examined the following two hypotheses: **Hypothesis I**: The preservation of language family relationship universally holds in non-native English; **Hypothesis II**: The preservation of language family relationship is dependent on proficiency in English. The experimental results partly accepted **Hy-**



(i) All levels (ii) High and low levels

Figure 6: Distribution of Adverb Position in Asian Englishes (ICNALE).

pothesis I and revealed that the following hypothesis fitted the data better: **Hypothesis I′**: The preservation of language family relationship universally holds in the expanding circle of English. By contrast, the experimental results denied **Hypothesis II**, supporting the counter hypothesis: **Hypothesis II′**: The preservation of language family relationship is independent of proficiency in English. We then proposed another hypothesis that a probabilistic module exists in the human brain to explain why **Hypothesis I′** and **Hypothesis II′** hold. We further introduced empirical data implying the existence of the probabilistic module.

For future work, we will examine **Hypothesis I′** and **II′** using English texts written by speakers of languages in other families to see if the preservation really universally holds. Also, we will explore the existence of the probabilistic module.

Acknowledgments

The author would like to thank the anonymous reviewers for their thoughtful comments and suggestions on this paper.

References

Jan Aarts and Sylviane Granger, 1998. *Tag sequences in learner corpora: a key to interlanguage grammar and discourse*, pages 132–141. Longman, New York.

Bengt Altenberg and Marie Tapper, 1998. *The use of adverbial connectors in advanced Swedish learners' written English*, pages 80–93. Longman, New York.

Robert S.P. Beekes. 2011. *Comparative Indo-European Linguistics: An Introduction (2nd ed.)*. John Benjamins Publishing Company, Amsterdam.

David Crystal. 1997. *The Cambridge Encyclopedia of Language (2nd ed.)*. Cambridge University Press, Cambridge.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proc. of 8th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English v2*. Presses universitaires de Louvain, Louvain.

Jiawei Han and Micheline Kamber. 2006. *Data Mining: Concepts and Techniques (2nd Ed.)*. Morgan Kaufmann Publishers, San Francisco.

Shinichiro Ishikawa, 2011. *A new horizon in learner corpus studies: The aim of the ICNALE project*, pages 3–11. University of Strathclyde Publishing, Glasgow.

Bing-Hwang Juang and Lawrence R. Rabiner. 1985. A probabilistic distance measure for hidden Markov models. *AT&T Technical Journal*, 64(2):391–408.

Braj B. Kachru, 1992. *Teaching World Englishes*, pages 355–365. University of Illinois Press, Urbana and Chicago.

Kenji Kita. 1999. Automatic clustering of languages based on probabilistic models. *Journal of Quantitative Linguistics*, 6(2):167–171.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184.

Ryo Nagata and Edward Whittaker. 2013. Reconstructing an Indo-European family tree from non-native English texts. In *Proc. of 51st Annual Meeting of the Association for Computational Linguistics*, pages 1137–1147.

Ryo Nagata, Edward Whittaker, and Vera Sheinman. 2011. Creating a manually error-tagged and shallow-parsed learner corpus. In *Proc. of 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1210–1219.

Anna Giacalone Ramat and Paolo Ramat. 2006. *The Indo-European Languages*. Routledge, New York.

Beatrice Santorini. 1990. *Part-of-speech tagging guidelines for the Penn Treebank Project*. University of Pennsylvania.

Michael Swan and Bernard Smith. 2001. *Learner English (2nd Ed.)*. Cambridge University Press, Cambridge.

Sze-Meng J. Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proc. Australasian Language Technology Workshop*, pages 53–61.