Reconstructing an Indo-European Family Tree from Non-native English texts

Ryo Nagata[1,2] **Edward Whittaker**[3]
[1]Konan University / Kobe, Japan
[2]LIMSI-CNRS / Orsay, France
[3]Inferret Limited / Northampton, England
nagata-acl@hyogo-u.ac.jp, ed@inferret.co.uk

Abstract

Mother tongue interference is the phenomenon where linguistic systems of a mother tongue are transferred to another language. Although there has been plenty of work on mother tongue interference, very little is known about how strongly it is transferred to another language and about what relation there is across mother tongues. To address these questions, this paper explores and visualizes mother tongue interference preserved in English texts written by Indo-European language speakers. This paper further explores linguistic features that explain why certain relations are preserved in English writing, and which contribute to related tasks such as native language identification.

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 such as Swan and Smith (2001), Aarts and Granger (1998), Davidsen-Nielsen and Harder (2001), and Altenberg and Tapper (1998) work on mother tongue interference to reveal overused/underused words, part of speech (POS), or grammatical items.

In contrast, very little is known about how strongly mother tongue interference is transferred to another language and about what relation there is across mother tongues. At one extreme, one could argue that it is so strongly transferred to texts in another language that the linguistic relations between mother tongues are perfectly preserved in the texts. At the other extreme, one can counter it, arguing that other features such as non-nativeness are more influential than mother tongue interference. One possible reason for this is that a large part of the distinctive language systems of a mother tongue may be eliminated when transferred to another language from a speaker's mother tongue. For example, Slavic languages have a rich inflectional case system (e.g., Czech has seven inflectional cases) whereas French does not. However, the difference in the richness cannot be transferred into English because English has almost no inflectional case system. Thus, one cannot determine the mother tongue of a given non-native text from the inflectional case. A similar argument can be made about some parts of gender, tense, and aspect systems. Besides, Wong and Dras (2009) show that there are no significant differences, between mother tongues, in the misuse of certain syntactic features such as subject-verb agreement that have different tendencies depending on their mother tongues. Considering these, one could not be so sure which argument is correct. In any case, to the best of our knowledge, no one has yet answered this question.

In view of this background, we take the first step in addressing this question. We hypothesize that:

Hypothesis: Mother tongue interference is so strong that the relations in a language family are preserved in texts written in another language.

In other words, mother tongue interference is so strong that one can reconstruct a language fam-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1137–1147,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

ily tree from non-native texts. One of the major contributions of this work is to reveal and visualize a language family tree preserved in non-native texts, by examining the hypothesis. This becomes important in native language identification[1] which is useful for improving grammatical error correction systems (Chodorow et al., 2010) or for providing more targeted feedback to language learners. As we will see in Sect. 6, this paper reveals several crucial findings that contribute to improving native language identification. In addition, this paper shows that the findings could contribute to reconstruction of language family trees (Enright and Kondrak, 2011; Gray and Atkinson, 2003; Barbançon et al., 2007; Batagelj et al., 1992; Nakhleh et al., 2005), which is one of the central tasks in historical linguistics.

The rest of this paper is structured as follows. Sect. 2 introduces the basic approach of this work. Sect. 3 discusses the methods in detail. Sect. 4 describes experiments conducted to investigate the hypothesis. Sect. 5 discusses the experimental results. Sect. 6 discusses implications for work in related domains.

2 Approach

To examine the hypothesis, we reconstruct a language family tree from English texts written by non-native speakers of English whose mother tongue is one of the Indo-European languages (Beekes, 2011; Ramat and Ramat, 2006). If the reconstructed tree is sufficiently similar to the original Indo-European family tree, it will support the hypothesis. If not, it suggests that some features other than mother tongue interference are more influential.

The approach we use for reconstructing a language family tree is to apply agglomerative hierarchical clustering (Han and Kamber, 2006) to English texts written by non-native speakers. Researchers have already performed related work on reconstructing language family trees. For instance, Kroeber and Chriétien (1937) and Ellegård (1959) proposed statistical methods for measuring the similarity metric between languages. More recently, Batagelj et al. (1992) and Kita (1999) proposed methods for reconstructing language family trees using clustering. Among them, the most related method is that of Kita (1999). In his method, a variety of languages are modeled by their spelling systems (i.e., character-based n-gram language models). Then, agglomerative hierarchical clustering is applied to the language models to reconstruct a language family tree. The similarity used for clustering is based on a divergence-like distance between two language models that was originally proposed by Juang and Rabiner (1985). This method is purely data-driven and does not require human expert knowledge for the selection of linguistic features.

Our work closely follows Kita's work. However, it should be emphasized that there is a significant difference between the two. Kita's work (and other previous work) targets clustering of a variety of languages whereas our work tries to reconstruct a language family tree preserved in non-native English. This significant difference prevents us from directly applying techniques in the literature to our task. For instance, Batagelj et al. (1992) use basic vocabularies such as *belly* in English and *ventre* in French to measure similarity between languages. Obviously, this does not work on our task; *belly* is *belly* in English writing whoever writes it. Kita's method is also likely not to work well because all texts in our task share the same spelling system (i.e., English spelling). Although spelling is sometimes influenced by mother tongues, it involves a lot more including overuse, underuse, and misuse of lexical, grammatical, and syntactic systems.

To solve the problem, this work adopts a word-based language model in the expectation that word sequences reflect mother tongue interference. At the same time, its simple application would cause a serious side effect. It would reflect the topics of given texts rather than mother tongue interference. Unfortunately, there exists no such English corpus that covers a variety of language speakers with uniform topics; moreover the availability of non-native corpora is still somewhat limited. This also means that available non-native corpora may be too small to train reliable word-based language models. The next section describes two methods (language model-based and vector-based), which address these problems.

3 Methods

3.1 Language Model-based Method

To begin with, let us define the following symbols used in the methods. Let D_i be a set of English

[1] Recently, native language identification has drawn the attention of NLP researchers. For instance, a shared task on native language identification took place at an NAACL-HLT 2013 workshop.

texts where i denotes a mother tongue i. Similarly, let M_i be a language model trained using D_i.

To solve the problems pointed out in Sect. 2, we use an n-gram language model based on a mixture of word and POS tokens instead of a simple word-based language model. 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, as desired. 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 first 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) following Kita's work and use Kneser-Ney (KN) smoothing (Kneser and Ney, 1995) to estimate its conditional probabilities.

With M_i and D_i, we can naturally apply Kita's method to our task. The clustering algorithm used is agglomerative hierarchical clustering with the average linkage method. The distance[2] 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

$$\begin{aligned}\Pr(D_i|M_i) \approx\\ &\Pr(w_{1,i})\Pr(w_{2,i}|w_{1,i})\\ &\times \prod_{t=3}^{|D_i|} \Pr(w_{t,i}|w_{t-2,i}, w_{t-1,i}) \qquad (1)\end{aligned}$$

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

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

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:

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

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.

3.2 Vector-based Method

We also examine a vector-based method for language family tree reconstruction. As we will see in Sect. 5, this method allows us to interpret clustering results more easily than with the language model-based method while both result in similar language family trees.

In this method, D_i is modeled by a vector. The vector is constructed based on the relative frequencies of trigrams. As a consequence, the distance is naturally defined by the Euclidean distance between two vectors. The clustering procedure is the same as for the language model-based method except that M_i is vector-based and that the distance metric is Euclidean.

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

4 Experiments

We selected the ICLE corpus v.2 (Granger et al., 2009) as the target language data. It consists of English essays written by a wide variety of non-native speakers of English. Among them, the 11 shown in Table 1 are of Indo-European languages. Accordingly, we selected the subcorpora of the 11 languages in the experiments. Before the experiments, we preprocessed the corpus data to control the experimental conditions. Because some of the writers had more than one native language, we excluded essays that did not meet the following three conditions: (i) the writer has only one native language; (ii) the writer has only one language at home; (iii) the two languages in (i) and (ii) are the same as the native language of the subcorpus to which the essay belongs[3]. After the selection, markup tags such as essay IDs were removed from the corpus data. Also, the symbols ‘ and ’ were unified into ’[4]. For reference, we also used native English (British and American university students’ essays in the LOCNESS corpus[5]) and two sets of Japanese English (ICLE and the NICE corpus (Sugiura et al., 2007)). Table 1 shows the statistics on the corpus data.

Native language	# of essays	# of tokens
Bulgarian	294	219,551
Czech	220	205,264
Dutch	244	240,861
French	273	202,439
German	395	236,841
Italian	346	219,581
Norwegian	290	218,056
Polish	354	251,074
Russian	255	236,748
Spanish	237	211,343
Swedish	301	268,361
English	298	294,357
Japanese1 (ICLE)	171	224,534
Japanese2 (NICE)	340	130,156
Total	4,018	3,159,166

Table 1: Statistics on target corpora.

Performance of POS tagging is an important factor in our methods because they are based on word/POS sequences. Existing POS taggers might not perform well on non-native English texts because they are normally developed to analyze native English texts. Considering this, we tested CRFTagger[6] on non-native English texts containing various grammatical errors before the experiments (Nagata et al., 2011). It turned out that CRFTagger achieved an accuracy of 0.932 (compared to 0.970 on native texts). Although it did not perform as well as on native texts, it still achieved a fair accuracy. Accordingly, we decided to use it in our experiments.

Then, we generated cluster trees from the corpus data using the methods described in Sect. 3. We used the Kyoto Language Modeling toolkit[7] to build language models from the corpus data. We removed n-grams that appeared less than five times[8] in each subcorpus in the language models. Similarly, we implemented the vector-based method with trigrams using the same frequency cutoff (but without smoothing).

Fig. 1 shows the experimental results. The tree at the top is the Indo-European family tree drawn based on the figure shown in Crystal (1997). It shows that the 11 languages are divided into three groups: Italic, Germanic, and Slavic branches. The second and third trees are the cluster trees generated by the language model-based and vector-based methods, respectively. The number at each branching node denotes in which step the two clusters were merged.

The experimental results strongly support the hypothesis we made in Sect. 1. Fig. 1 reveals that the language model-based method correctly groups the 11 Englishes into the Italic, Germanic, and Slavic branches. It first merges Norwegian-English and Swedish-English into a cluster. The two languages belong to the North Germanic branch of the Germanic branch and thus are closely related. Subsequently, the language model-based method correctly merges the other languages into the three branches. A dif-

[3] For example, because of (iii), essays written by native speakers of Swedish in the Finnish subcorpus were excluded from the experiments. This is because they were collected in Finland and might be influenced by Finnish.

[4] The symbol ‘ is sometimes used for ’ (e.g., *I‘m*).

[5] The LOCNESS corpus is a corpus of native English essays made up of British pupils’ essays, British university students’ essays, and American university students’ essays: `https://www.uclouvain.be/en-cecl-locness.html`

[6] Xuan-Hieu Phan, “CRFTagger: CRF English POS Tagger,” `http://crftagger.sourceforge.net/`, 2006.

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

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

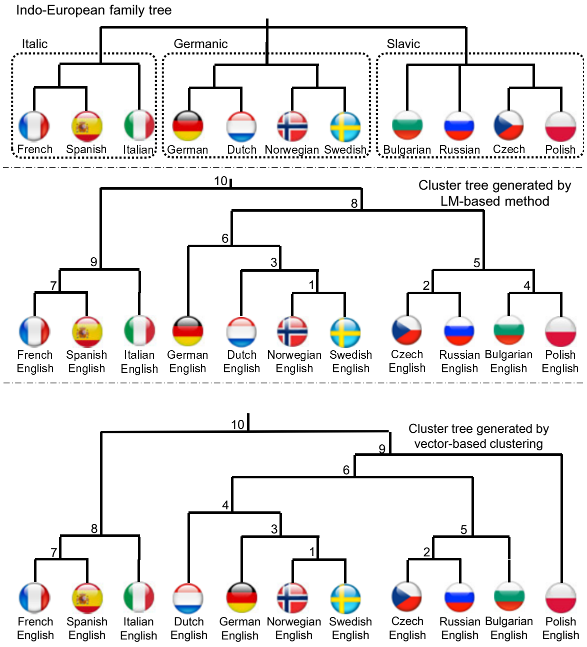

Figure 1: Experimental results.

ference between its cluster tree and the Indo-European family tree is that there are some mismatches within the Germanic and Slavic branches. While the difference exists, the method strongly distinguishes the three branches from one another. The third tree shows that the vector-based method behaves similarly while it mistakenly attaches Polish-English into an independent branch. From these results, we can say that mother tongue interference is transferred into the 11 Englishes, strongly enough for reconstructing its language family tree, which we propose calling *the interlanguage Indo-European family tree* in English.

Fig. 2 shows the experimental results with native and Japanese Englishes. It shows that the same interlanguage Indo-European family tree was reconstructed as before. More interestingly, native English was detached from the interlanguage Indo-European family tree contrary to the expectation that it would be attached to the Germanic branch because English is of course a member of the Germanic branch. This implies that non-nativeness common to the 11 Englishes is more influential than the intrafamily distance is[9]; otherwise, native English would be included in the German branch. Fig. 2 also shows that the two sets of Japanese English were merged into a cluster and that it was the most distant in the whole tree. This shows that the interfamily distance is the most influential factor. Based on these results, we can further hypothesize as follows: interfamily distance > non-nativeness > intrafamily distance.

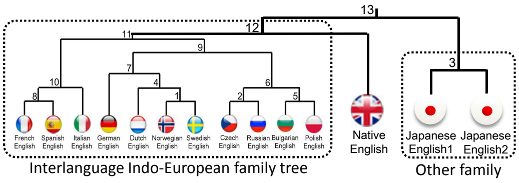

Figure 2: Experimental results with native and Japanese Englishes.

5 Discussion

To get a better understanding of the interlanguage Indo-European family tree, we further explore linguistic features that explain well the above phenomena. When we analyze the experimental results, however, some problems arise. It is almost impossible to find someone who has a good knowledge of the 11 languages and their mother language interference in English writing. Besides, there are a large number of language pairs to compare. Thus, we need an efficient and effective way to analyze the experimental results.

To address these problems, we did the following. First, we focused on only a few Englishes out of the 11. Because one of the authors had some knowledge of French, we selected French-English as the main target. This naturally made us select the other Italic Englishes as its counterparts. Also, because we had access to a native speaker of Russian who had a good knowledge of English, we included Russian-English in our focus. We analyzed these Englishes and then examined whether the findings obtained apply to the other Englishes or not. Second, we used a method for extracting interesting trigrams from the corpus data. The method compares three out of the 11 corpora (for example, French-, Spanish-, and Russian-Englishes). If we remove instances of a trigram from each set, the clustering tree involving

[9] Admittedly, we need further investigation to confirm this argument especially because we applied CRFTagger, which is developed to analyze native English, to both non-native and native Englishes, which might affect the results.

the three may change. For example, the removal of *but the hers* may result in a cluster tree merging French- and Russian-Englishes before French- and Spanish-Englishes. Even if it does not change, the distances may change in that direction. We analyzed what trigrams had contributed to the clustering results with this approach.

To formalize this approach, we will denote a trigram by t. We will also denote its relative frequency in the language data D_i by r_{ti}. Then, the change in the distances caused by the removal of t from D_i, D_j, and D_k is quantified by

$$s = (r_{tk} - r_{ti})^2 - (r_{tj} - r_{ti})^2 \qquad (4)$$

in the vector-based method. The quantity $(r_{tk} - r_{ti})^2$ is directly related to the decrease in the distance between D_i and D_k and similarly, $(r_{tj} - r_{ti})^2$ to that between D_i and D_j in the vector-based method. Thus, the greater s is, the higher the chance that the cluster tree changes. Therefore, we can obtain a list of interesting trigrams by sorting them according to s. We could do a similar calculation in the language model-based method using the conditional probabilities. However, it requires a more complicated calculation. Accordingly, we limit ourselves to the vector-based method in this analysis, noting that both methods generated similar cluster trees.

Table 2 shows the top 15 interesting trigrams where D_i, D_j, and D_k are French-, Spanish-, and Russian-Englishes, respectively. Note that s is multiplied by 10^6 and r is in % for readability. The list reveals that many of the trigrams contain the article *a* or *the*. Interestingly, their frequencies are similar in French-English and Spanish-English, and both are higher than in Russian-English. This corresponds to the fact that French and Spanish have articles whereas Russian does not. Actually, the same argument can be made about the other Italic and Slavic Englishes (e.g., *the JJ NN*: Italian-English 0.82; Polish-English 0.72)[10]. An exception is that of trigrams containing the definite article in Bulgarian-English; it tends to be higher in Bulgarian-English than in the other Slavic Englishes. Surprisingly and interestingly, however, it reflects the fact that Bulgarian does have the definite article but not the indefinite article (e.g., *the JJ NN*: 0.82; *a JJ NN*: 0.60 in Bulgarian-English).

Table 3 shows that the differences in article use exist even between the Italic and Germanic branches despite the fact that both have the indefinite and definite articles. The list still contains a number of trigrams containing articles. For a better understanding of this, we looked further into the distribution of articles in the corpus data. It turns out that the distribution almost perfectly groups the 11 Englishes into the corresponding branches as shown in Fig. 3. The overall use of articles is less frequent in the Slavic-Englishes. The definite article is used more frequently in the Italic-Englishes than in the Germanic Englishes (except for Dutch-English). We speculate that this is perhaps because the Italic languages have a wider usage of the definite article such as its modification of possessive pronouns and proper nouns. The Japanese Englishes form another group (this is also true for the following findings). This corresponds to the fact that the Japanese language does not have an article system similar to that of English.

s	Trigram t	r_{ti}	r_{tj}	r_{tk}
5.14	the NN of	1.01	0.98	0.78
4.38	a JJ NN	0.85	0.77	0.62
2.74	the JJ NN	0.87	0.86	0.71
2.30	NN of the	0.49	0.52	0.33
1.64	. . .	0.22	0.12	0.05
1.56	NNS . EOS	0.77	0.70	0.92
1.31	NNS and NNS	0.09	0.13	0.21
1.25	BOS RB ,	0.25	0.22	0.14
1.22	of the NN	0.42	0.44	0.30
1.17	VBZ to VB	0.26	0.22	0.14
1.09	BOS i VBP	0.07	0.05	0.17
1.03	NN of NN	0.74	0.70	0.63
0.88	NN of JJ	0.15	0.15	0.25
0.67	the JJ NNS	0.28	0.28	0.20
0.65	NN to VB	0.40	0.38	0.31

Table 2: Interesting trigrams (French- (D_i), Spanish- (D_j), and Russian- (D_k) Englishes).

Another interesting trigram, though not as obvious as article use, is *NN of NN*, which ranks 12th and 2nd in Table 2 and 3, respectively. In the Italic Englishes, the trigram is more frequent than the other non-native Englishes as shown in Fig. 4. This corresponds to the fact that noun-noun compounds are less common in the Italic languages than in English and that instead, the *of*-phrase (*NN of NN*) is preferred (Swan and Smith, 2001). For

[10] Due to the space limitation, other lists were not included in this paper but are available at `http://web.hyogo-u.ac.jp/nagata/acl/`.

s	Trigram t	r_{ti}	r_{tj}	r_{tk}
21.49	the NN of	1.01	0.98	0.54
5.70	NN of NN	0.74	0.70	0.50
3.26	NN of the	0.49	0.52	0.30
3.10	the JJ NN	0.87	0.86	0.70
2.62	. . .	0.22	0.12	0.03
1.53	of the NN	0.42	0.44	0.29
1.50	NN , NN	0.30	0.30	0.18
1.50	BOS i VBP	0.07	0.05	0.19
0.85	NNS and NNS	0.09	0.13	0.19
0.81	JJ NN of	0.40	0.39	0.31
0.68	. . EOS	0.13	0.06	0.02
0.63	a JJ NN	0.85	0.77	0.73
0.63	RB . EOS	0.21	0.16	0.31
0.56	NN , the	0.16	0.16	0.08
0.50	NN of a	0.17	0.09	0.06

Table 3: Interesting trigrams (French- (D_i), Spanish- (D_j), and Swedish- (D_k) Englishes).

instance, *orange juice* is expressed as *juice of orange* in the Italic languages (e.g., *jus d'orange* in French). In contrast, noun-noun compounds or similar constructions are more common in Russian and Swedish. As a result, *NN of NN* becomes relatively frequent in the Italic Englishes. Fig. 4 also shows that its distribution roughly groups the 11 Englishes into the three branches. Therefore, the way noun phrases (NPs) are constructed is a clue to how the three branches were clustered.

This finding in turn reveals that the consecutive repetitions of nouns occur less in the Italic Englishes. In other words, the length tends to be shorter than in the others where we define the length as the number of consecutive repetitions of common nouns (for example, the length of *orange juice* is one because a noun is consecutively repeated once). To see if this is true, we calculated the average length for each English. Fig. 5 shows that the average length roughly distinguishes the Italic Englishes from the other non-native Englishes; French-English is the shortest, which is explained by the discussion above, while Dutch- and German-Englishes are longest, which may correspond to the fact that they have a preference for noun-noun compounds as Snyder (1996) argues. For instance, German allows the concatenated form as in *Orangensaft* (equivalently *orangejuice*). This tendency in the length of noun-noun compounds provides us with a crucial insight for native language identification, which we will come back to in Sect. 6.

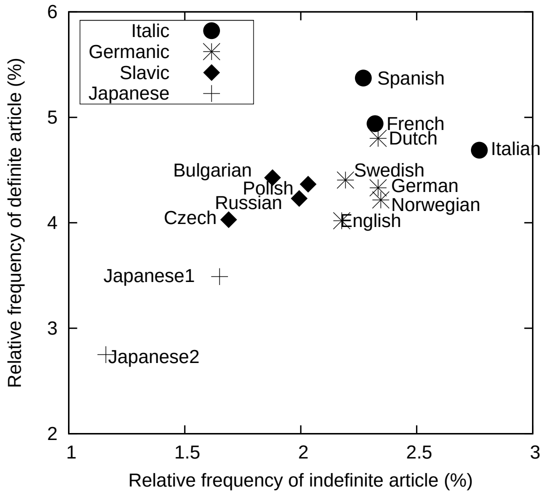

Figure 3: Distribution of articles.

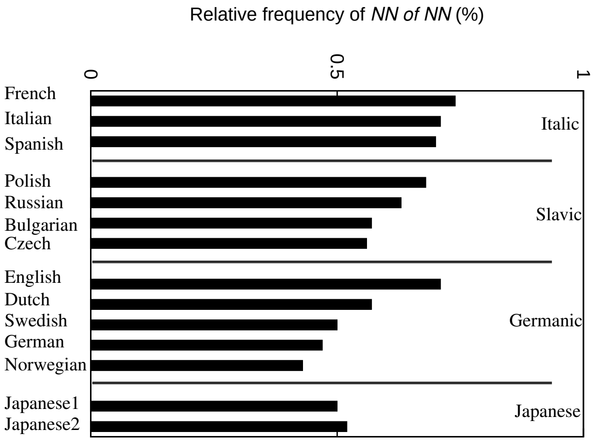

Figure 4: Relative frequency of *NN of NN* in each corpus (%).

The trigrams *BOS RB ,* in Table 2 and *RB . EOS* in Table 3 imply that there might also be a certain pattern in adverb position in the 11 Englishes (they roughly correspond to adverbs at the beginning and end of sentences). Fig. 6 shows an insight into this. The horizontal and vertical axes correspond to the ratio of adverbs at the beginning and the end of sentences, respectively. It turns out that the German Englishes form a group. So do the Italic Englishes although it is less dense. In contrast, the Slavic Englishes are scattered. However, the ratios give a clue to how to distinguish Slavic Englishes from the others when combined with other

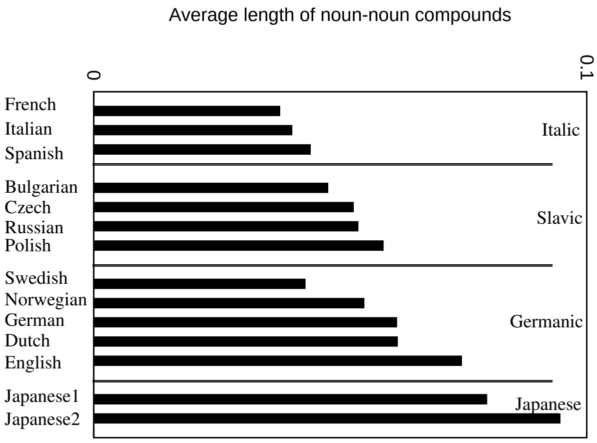

Figure 5: Average length of noun-noun compounds in each corpus.

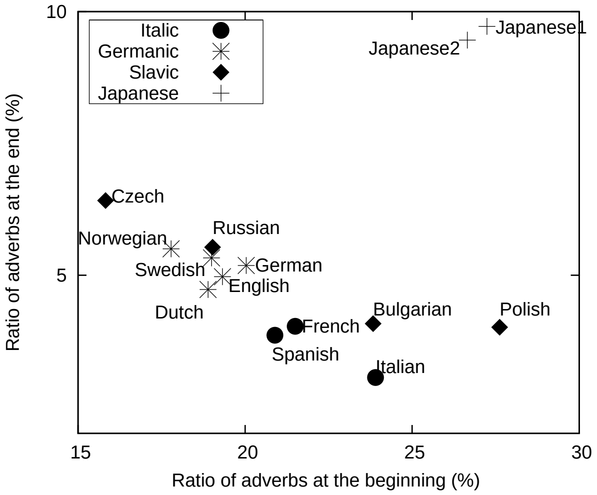

Figure 6: Distribution of adverb position.

trigrams. For instance, although Polish-English is located in the middle of Swedish-English and Bulgarian-English in the distribution of articles (in Fig. 3), the ratios tell us that Polish-English is much nearer to Bulgarian-English.

6 Implications for Work in Related Domains

Researchers including Wong and Dras (2009), Wong et al. (2011; 2012), and Koppel et al. (2005) work on native language identification and show that machine learning-based methods are effective. Wong and Dras (2009) propose using information about grammatical errors such as errors in determiners to achieve better performance while they show that its use does not improve the performance, contrary to the expectation. Related to this, other researchers (Koppel and Ordan, 2011; van Halteren, 2008) show that machine learning-based methods can also predict the source language of a given translated text although it should be emphasized that it is a different task from native language identification because translation is not typically performed by non-native speakers but rather native speakers of the target language[11].

The experimental results show that n-grams containing articles are predictive for identifying native languages. This indicates that they should be used in the native language identification task. Importantly, all n-grams containing articles should be used in the classifier unlike the previous methods that are based only on n-grams containing article errors. Besides, no articles should be explicitly coded in n-grams for taking the overuse/underuse of articles into consideration. We can achieve this by adding a special symbol such as ϕ to the beginning of each NP whose head noun is a common noun and that has no determiner in it as in "I like ϕ orange juice."

In addition, the length of noun-noun compounds and the position of adverbs should also be considered in native language identification. In particular, the former can be modeled by the Poisson distribution as follows. The Poisson distribution gives the probability of the number of events occurring in a fixed time. In our case, the number of events in a fixed time corresponds to the number of consecutive repetitions of common nouns in NPs, which in turn corresponds to the length. To be precise, the probability of a noun-noun compound with length l is given by

$$\Pr(l) = \frac{\lambda^l}{l!}\mathrm{e}^{-\lambda}, \quad (5)$$

where λ corresponds to the average length. Fig. 7 shows that the observed values in the French-English data very closely fit the theoretical proba-

[11] For comparison, we conducted a pilot study where we reconstructed a language family tree from English texts in European Parliament Proceedings Parallel Corpus (Europarl) (Koehn, 2011). It turned out that the reconstructed tree was different from the canonical tree (available at `http://web.hyogo-u.ac.jp/nagata/acl/`). However, we need further investigation to confirm it because each subcorpus in Europarl is variable in many dimensions including its size and style (e.g., overuse of certain phrases such as *ladies and gentlemen*).

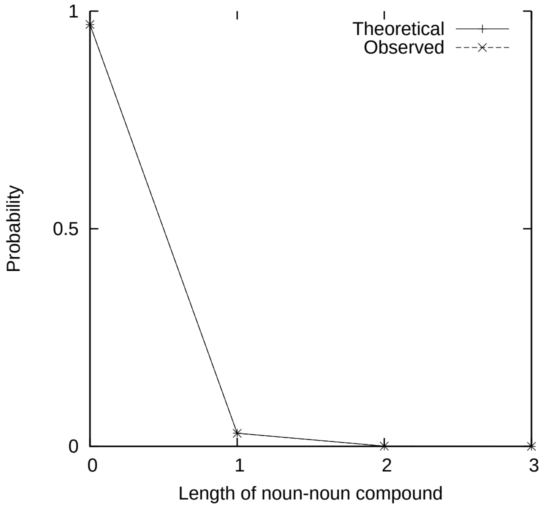

Figure 7: Distribution of noun-noun compound length for French-English.

bilities given by Equation (5)[12]. This holds for the other Englishes although we cannot show them because of the space limitation. Consequently, Equation (5) should be useful in native language identification. Fortunately, it can be naturally integrated into existing classifiers.

In the domain of historical linguistics, researchers have used computational and corpus-based methods for reconstructing language family trees. Some (Enright and Kondrak, 2011; Gray and Atkinson, 2003; Barbançon et al., 2007; Batagelj et al., 1992; Nakhleh et al., 2005) apply clustering techniques to the task of language family tree reconstruction. Others (Kita, 1999; Rama and Singh, 2009) use corpus statistics for the same purpose. These methods reconstruct language family trees based on linguistic features that exist within words including lexical, phonological, and morphological features.

The experimental results in this paper suggest the possibility of the use of non-native texts for reconstructing language family trees. It allows us to use linguistic features that exist between words, as seen in our methods, which has been difficult with previous methods. Language involves the features between words such as phrase construction and syntax as well as the features within words and thus they should both be considered in reconstruction of language family trees.

7 Conclusions

In this paper, we have shown that 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. To show this, we have used clustering to reconstruct a language family tree from 11 sets of non-native English texts. It turned out that the reconstructed tree correctly groups them into the Italic, Germanic, and Slavic branches of the Indo-European family tree. Based on the resulting trees, we have then hypothesized that the following relation holds in mother tongue interference: interfamily distance $>$ non-nativeness $>$ intrafamily distance. We have further explored several intriguing linguistic features that play an important role in mother tongue interference: (i) article use, (ii) NP construction, and (iii) adverb position, which provide several insights for improving the tasks of native language identification and language family tree reconstruction.

Acknowledgments

This work was partly supported by the Digiteo foreign guest project. We would like to thank the three anonymous reviewers and the following persons for their useful comments on this paper: Kotaro Funakoshi, Mitsuaki Hayase, Atsuo Kawai, Robert Ladig, Graham Neubig, Vera Sheinman, Hiroya Takamura, David Valmorin, Mikko Vilenius.

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.

François Barbançon, Tandy Warnow, Steven N. Evans, Donald Ringe, and Luay Nakhleh. 2007. An experimental study comparing linguistic phylogenetic reconstruction methods. *Statistics Technical Reports*, page 732.

Vladimir Batagelj, Tomaž Pisanski, and Damijana Keržič. 1992. Automatic clustering of languages. *Computational Linguistics*, 18(3):339–352.

[12] The theoretical and observed values are so close that it is difficult to distinguish between the two lines in Fig. 7. For example, $\Pr(l = 1) = 0.0303$ while the corresponding observed value is 0.0299.

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

Martin Chodorow, Michael Gamon, and Joel R. Tetreault. 2010. The utility of article and preposition error correction systems for English language learners: feedback and assessment. *Language Testing*, 27(3):419–436.

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

Niels Davidsen-Nielsen and Peter Harder, 2001. *Speakers of Scandinavian languages: Danish, Norwegian, Swedish*, pages 21–36. Cambridge University Press, Cambridge.

Alvar Ellegård. 1959. Statistical measurement of linguistic relationship. *Language*, 35(2):131–156.

Jessica Enright and Grzegorz Kondrak. 2011. The application of chordal graphs to inferring phylogenetic trees of languages. In *Proc. of 5th International Joint Conference on Natural Language Processing*, pages 8–13.

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

Russell D. Gray and Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, 426:435–438.

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

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

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.

Philipp Koehn. 2011. Europarl: A parallel corpus for statistical machine translation. In *Proc. of 10th Machine Translation Summit*, pages 79–86.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proc. of 49th Annual Meeting of the Association for Computational Linguistics*, pages 1318–1326.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proc. of 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 624–628.

Alfred L. Kroeber and Charles D. Chriétien. 1937. Quantitative classification of Indo-European languages. *Language*, 13(2):83–103.

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

Luay Nakhleh, Tandy Warnow, Don Ringe, and Steven N. Evans. 2005. A comparison of phylogenetic reconstruction methods on an Indo-European dataset. *Transactions of the Philological Society*, 103(2):171–192.

Taraka Rama and Anil Kumar Singh. 2009. From bag of languages to family trees from noisy corpus. In *Proc. of Recent Advances in Natural Language Processing*, pages 355–359.

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

William Snyder. 1996. The acquisitional role of the syntax-morphology interface: Morphological compounds and syntactic complex predicates. In *Proc. of Annual Boston University Conference on Language Development*, volume 2, pages 728–735.

Masatoshi Sugiura, Masumi Narita, Tomomi Ishida, Tatsuya Sakaue, Remi Murao, and Kyoko Muraki. 2007. A discriminant analysis of non-native speakers and native speakers of English. In *Proc. of Corpus Linguistics Conference CL2007*, pages 84–89.

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

Hans van Halteren. 2008. Source language markers in EUROPARL translations. In *Proc. of 22nd International Conference on Computational Linguistics*, pages 937–944.

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

Sze-Meng J. Wong, Mark Dras, and Mark Johnson. 2011. Exploiting parse structures for native language identification. In *Proc. Conference on Empirical Methods in Natural Language Processing*, pages 1600–1611.

Sze-Meng J. Wong, Mark Dras, and Mark Johnson. 2012. Exploring adaptor grammars for native language identification. In *Proc. Joint Conference on*

Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 699–709.