ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 12, No. 6, pp. 1048-1058, June 2022
DOI: https://doi.org/10.17507/tpls.1206.04

Clarifying Learner Englishes From Greater China Using Native Language Identification — A Pilot Study

Xiaoyun Li
Department of Theoretical Linguistics, University of Szeged, Szeged, Hungary

***Abstract*—The purpose of this paper is to identify the characteristics of learner Englishes from the three major regions of Greater China, namely, Mainland China, Hong Kong, and Taiwan. To achieve this aim, a comparative study is conducted into the three learner Englishes via Native Language Identification (NLI). The average identification accuracy yielded in this study is 60 % on spoken monologues and 59.8 % on written essays. With these two satisfactory accuracies, this paper profiles the three learner Englishes by probing into their best-identifying indicators. The results show that learner English from Mainland China are characteristic for high degree of collectivistic involvement and uncertainty, low informativeness, and underuse of conjunctions; learner English from HKG is highly informative and impersonal; the two types of learner English from Taiwan are similar in that they share an individualistically involved style but differ in that the English essays by Taiwan L2 learners are found to be high on uncertainty and negation but low on informativeness and the usage of conjunctions..**

***Index Terms*—learner English, greater China, native language identification, spoken monologue, written essays**

I. Introduction

The integration of Greater China (i.e., mainland China, Hong Kong, Taiwan, and Macau) into globalization in the last two centuries has witnessed the exponential growth of English-language learners in this area. The penetration of English in the societies of Greater China is shown by the unprecedented passion of people to learn English (Feng, 2012), and as a result, learner English of L1 Chinese has received enormous attention from Chinese and international researchers. It must be admitted that plenty of current research focusing on learner Englishes from the different areas of Greater China is comprehensive and illuminating. There still are two major research voids waiting for filling.

The first void is that the research regarding L1 Chinese learners English appears to be fragmented despite the fact that learners share the same first language (L1) and cultural background and that the three major regions of Greater China (viz. Mainland China, Hong Kong, and Taiwan) are geographically close to each other. The majority of the learner English research which claims to be Chinese-targeted, is more likely mainland Chinese-centered, while researchers from Hong Kong or Taiwan often put particular concerns on learners from the region where they originate. Although a separate analysis could offer a more in-depth look, it might not be able to distinguish whether one linguistic phenomenon is shared by learners from other regions or not. Their similarities and differences thus remain unclear.

The second void lies in the fact that in learner English corpus research, corpus linguists invariably construct separate corpora for the three regions with no substantial linguistic evidence offered for doing such a division. Indeed, the societies of the three regions are "hugely different" (Feng, 2012, p.363) and their language policy, planning and pedagogical implementation therefore also show great discrepancies (Crystal, 2011). For English-language learners from mainland China and Taiwan, English is a foreign language (EFL) mostly spoken and learned in the classroom, while in Hong Kong, English is a second language (ESL) and serves as an official language other than Chinese. Another major difference regarding language is that Greater China is diversified in terms of dialects: Putonghua (or Mandarin) is vastly spoken in both Mainland and Taiwan while in Hong Kong, Cantonese, one major dialect of Chinese that is hardly intelligible to people from the other two regions, prevails among the local inhabitants. This study therefore hopes to provide some linguistic evidence for doing the division.

Given the above research voids, the present study is dedicated to profiling the learner Englishes (both written and spoken) from three regions of Greater China by employing Native Language Identification (NLI).

II. NLI and Learner Language Study

NLI refers to "the task of automatically identifying the first language of a language user based on the person's production of the target language" (Jarvis & Paquot, 2015, p.605). Its identification function is achieved through categorizing the certain traits transferred from a language user's mother tongue in his or her target language productions. The emergence of NLI is motivated by the need for solving a variety of practical issues. One of the major benefits

brought by NLI is that it offers an essential method for mining the information about the authorship of a text produced by an anonymous author, which is crucial for countering security threats in real life (Koppel et al., 2005), especially in today's world where cybercrimes become increasingly covert. Another primary application of NLI is that it can enhance the robustness of NLP tools and techniques on performing tasks that might be related with non-native English varieties, including parsing, speech recognition, and information extraction as currently a notable of NLP tools and techniques are designed solely based on native speaker data (Jarvis & Paquot, 2015).

Since NLI is guided by the theory of language transfer in SLA research, detecting L2 learners' mother tongue might offer pedagogical insights for language learning and teaching (Malmasi, 2016). According to the transfer theory, there are roughly two types of transfer - positive transfer and positive transfer, with the former being the transfer of L1 that might boosts L2 learning and the latter as the transfer that might hinder or pose difficulties for language learning. Adopting NLI into pedagogical settings thus could provide certain feedbacks, especially on the negative transfer for L2 learners and educators. For example, fossilization, a typical learner language phenomenon that might be amplified by negative L1 transfer (Tarone, 2012), can be further determined through NLI conducted between different learner groups from varied L1s. Also, NLI can contribute to the SLA theory building (Jarvis & Paquot, 2015) and Forensic Linguistic studies (Perkins, 2015).

So far, English has dominated the NLI research, which is not surprising at all given that English is the lingua franca of the world. However, such a situation is under changing recently due to the increasing international migration and the consequent need for learning or acquiring other languages. Malmasi and Dras (2014c) use a Finnish learner corpus which is composed of L2 Finnish writings by learners from 9 L1s to determine the availabilities of the NLI techniques which are developed based on L2 English in the context of Finnish. They achieve a 69.5 % accuracy against the baseline of 20% in predicting learners' L1 and 97 % in distinguishing non-native from native writings. In their another study on Chinese learner corpus (Malmasi & Dras, 2014b), they find that the models using part-of-speech tags, context-free grammar production rules, and function words identify Chinese L2 learners from 11 L1s with 71 % accuracy and the same models can also achieve a similar accuracy on English L2 learners. Other major NLI research that deals with non-English language learners' production includes Malmasi and Dras (2014a) on Arabic, del Río et al. (2018) on Portuguese, Malmasi et al. (2015) on Norwegian, to name but a few.

It might be arguable, however, that NLI, in theory, targets at L2 learner groups with varied L1 backgrounds, whereas the three L2 learner groups chosen in this study share the same mother tongue and cultural background. The consideration for this study not to follow prior NLI research is that it is the classification mechanism of NLI instead of its theoretical framework that is adopted. To be more specific, in common NLI research L2 learner groups from various L1 backgrounds are assumed to be distinctive to each other by default and the reached results are termed as the traits left by corresponding L2 learner's L1 regardless of the fact that L2 learners' target language productions, in addition to L1, are also subject to influences of other factors including, but not limited to, social context (Tarone, 2007), target language environment (Håkansson & Norrby, 2010), language varieties of L1 (Nisioi, 2015). Similarly, this study, with the assumption that the three learner Englishes from Greater China are distinct to each other, attempts to investigate the extent to which they can be distinguished and what the most discriminative features are.

TABLE 1
SELECTED CORPORA

L2 Corpus	Register	Document
Mainland China	Spoken	200
	Written	200
Taiwan	Spoken	200
	Written	200
Hong Kong	Spoken	200
	Written	200
Singapore	Spoken	200
	Written	200
Japan	Spoken	200
	Written	200
Native English speaker	Spoken	200
	Written	200
Total		2400

III. CORPUS AND METHOD

The present study is conducted based on the texts extracted from the International Corpus Network of Asian Learners of English (ICNALE hereinafter) - a corpus developed by Ishikawa (2013, p.91) for "a contrastive interlanguage analysis of varied English-language learners in Asia". This corpus consists of both spoken and written data which are collected under well-controlled conditions, including prompts, speaking or writing time, genres, and so forth. The

newest version of ICNALE (until August 2019) contains 15160 texts produced by English-language learners at college level from 11 Asian countries or regions and 370 texts by native English speakers from UK, Australia, US, and New Zealand. There are four modules in ICANLE: Spoken Monologue, Spoken Dialogue, written essays, and edited essays. In this study, the Spoken Dialogue module and Edited Essays module are ruled out due to their limited number of texts (425 and 640) in comparison with the other two modules. Out of the extracted texts from the Spoken Monologue module and Written Essay module of ICNALE, six corpora are built, namely, mainland China (CHN), Taiwan (TWN), Hong Kong (HKG), Singapore (SIN), Japan (JPN) and native English speakers (ENS). The later three subsets are chosen to provide references for the former three subsets since they represent respectively three different types of English status: English as second language (ESL), English as foreign language (EFL), and English as native language (ENL). Each corpus consists of 200 spoken monologue texts and 200 written essays that are randomly selected from corresponding corpus of ICNALE (See Table 1).

TABLE 2
SELECTED INDICES

Statistical Indices	Average number and rate of tokens and lemmas. Average length and number of sentences.
Morphologic Indices	Average number and rate of unknown words, verbs, nouns, adjectives, adverbs, pronouns, connectives, number words, proper words, singular 1st person pronouns, 1st person plural pronouns, demonstrative pronouns, past and present tenses, and punctuations.
Syntactic Indices	Average number and rate of active and passive subjects and objects. Average number and rate of attributives, adverbials and coordinations.
Semantic Indices	Average number and rate of uncertain words, negation words, epistemic modals, investigation words, weasel words and peacock words. Average number and rate of emotional words including joy, fear, sorrow, anger, positive, negative. Average number and rate of content words, function words, private verbs, public verbs and suasive verbs.

There were two steps carried on in this study. The first was to perform a NLI experiment to determine the extent to which the English corpora from the three regions of Greater China can be identified. UDpipe developed by Straka and Strakov á(2017) was applied to conduct POS tagging on the 12 chosen corpora (6 spoken corpora and 6 written corpora). In total, 85 linguistic indices that can be divided into statistical, morphological, syntactic and semantic categories were used (See Table 2). Once the corpora were tagged, a SVM model (Cortes & Vapnik, 1995) was selected to train the language data on Weka 3.8.4 (Hall et al., 2009) under the conditions of Weka's default settings and ten-fold cross-validation. It is expected that after this step, the identification accuracies of the investigated corpora can be obtained.

Considering that the starting point of this study is to profile the three learner Englishes from Greater China, the second step focused on features that contribute most to the discriminations of the corpora, by which the three corpora were profiled. This step was implemented through calculating ANOVA F-scores for the 85 investigated indices from each corpus once the corpora were confirmed to be effectively identified. The calculation of ANOVA F-scores was undertaken between the corpus in discussing and a reference corpus that is comprised of the rest corpora. The statistically significant features from each corpus were then retained for an in-depth discussion.

IV. RESULTS AND DISCUSSION

Table 3 summarizes the identification results of the corpora investigated in this study. It can be seen that the average identification accuracies of spoken and written corpora are both around 60% - an accuracy that is well above the baseline of 16.67%, indicating that the corpora can be effectively classified. Besides, it should be noted in particular that although the identification accuracies of the three learner Englishes from Greater China appear to be relatively low, they in fact can be higher if the reference corpora (i.e., ENS, JPN and SIN) are removed. In sum, it is safe to say that the three learner Englishes are distinctive to each other.

TABLE 3
IDENTIFICATION ACCURACY

	Spoken			Written		
Corpus	Precision	Recall	F-score	Precision	Recall	F-score
CHN	0.614	0.580	0.596	0.507	0.525	0.516
ENS	0.597	0.583	0.583	0.757	0.725	0.738
HKG	0.477	0.474	0.474	0.505	0.550	0.526
JPN	0.743	0.754	0.754	0.697	0.770	0.732
SIN	0.610	0.610	0.610	0.600	0.600	0.600
TWN	0.562	0.610	0.585	0.528	0.425	0.471
Average	0.600	0.601	0.600	0.598	0.599	0.597

Given that the three learner Englishes from Greater China can be correctly identified with relatively satisfactory accuracies, it is necessary find out the major features that primarily contribute to the classification of the three groups. The present study conducts ANOVA F-score calculations for each significant feature ($p<0.05$) from each corpus. Due to the limited space, only the top 15 significant indices from each corpus are selected and discussed.

A. Spoken Monologue

This section is to profile the characteristics of monologues included in the three corpora from Greater China based on the indices proven to be statistically significant for identifying corresponding corpora. In the following part, the frequency information of the statistically significant indices from each corpus is compared and discussed in detail. In addition, to make findings more reliable, indices with ratio values are focused given that the corpora investigated vary considerably with respect to average monologue length (*Mean = 119, SD = 31.2*).

1. Mainland China

The best indicator for CHN is associated with two indices related to 1st person plural pronouns (Pl1PronNum & Pl1PronRate). As Table 4 shows, learners from CHN use Pl1Prons far more frequently than other learners/native speakers of English, suggesting that they prefer a highly involved style and tend to think and speak from a collective perspective. Besides, the overuse of Pl1Prons is found to be persistent in CHN across proficiency-levels. Scrutiny of the monologue texts from CHN shows that upper-level learners even use more Pl1Prons: 1.52 (per 100 words) for A2 learners, 2.0 for B1 learners, 1.96 for B1_2 learners, and 2.86 for B2 learners.

TABLE 4
TOP 15 INDICATORS FOR CHN (SPOKEN MONOLOGUE)

Rank	Feature	F-score	CHN	ENS	HKG	TWN	JPN	SIN
1	Punct	143.14	**24.550**	20.870	18.580	14.880	12.460	18.080
2	Pl1PronNum	97.32	**2.915**	0.505	1.160	1.135	1.515	0.850
3	Pl1PronRate	90.02	**0.212**	0.023	0.064	0.087	0.185	0.041
4	ConjRate	54.43	**0.066**	0.080	0.084	0.077	0.079	0.087
5	ConjNum	52.81	**7.295**	12.710	10.760	7.370	5.340	13.180
6	UncertainRate	30.90	**0.014**	0.008	0.010	0.011	0.008	0.007
7	PresentNum	30.57	**5.270**	9.080	6.475	5.920	4.500	7.990
8	VerbNum	29.29	**14.275**	22.840	17.770	14.255	9.975	20.51
9	PeacockRate	29.25	**0.129**	0.142	0.138	0.151	0.147	0.135
10	AdverbNum	25.74	**8.315**	13.990	10.465	8.055	5.680	14.545
11	AdvNum	24.63	**7.515**	12.465	9.975	6.920	4.650	14.285
12	UncertainNum	21.89	**1.530**	1.285	1.245	0.980	0.580	1.115
13	WeaselRate	18.92	**0.113**	0.102	0.106	0.106	0.084	0.102
14	AdjRate	17.43	**0.067**	0.059	0.058	0.054	0.062	0.062
15	DemPronNum	16.10	**0.385**	0.745	0.640	0.290	0.735	0.930

The massive use of Pl1Pron in CHN might be a carefully designed speaking technique for increasing the persuasion of the statement. Both *I* and *we* are categorized as ego-involvement (Chafe, 1985), with the difference that the first one is ubiquitously "referring to the speaker/writer" (Biber et al., 2007, p.93) while the latter one includes both the speaker/writer and the addressees. *We* is shown to be a powerful linguistic device to establish the group identity (Íñigo-Mora, 2004) by involving the speaker and the addresses together. As can be seen from the following text, apart from the first sentence where the author uses 1st person singular pronoun to highlight his/her central opinion, the rest part of the whole monologue is switched to a collective angle. The speaker seems to use 1st person plural pronouns to shorten his/her distance to the possible listeners, thus creating an impression to the possible addressees that the central topic is associated with not only his/her warfare but also theirs.

> *I agree that it's important for college students to have a part-time job. The reasons are as follows. First, have a part – having a part-time job can help a – help us earn some money by our own which can share some economic burden with our parents and, uh, if we have some money of our own we can spend – spend the extra and something we – well, like we haven't something that's generally or buy something we wish to. Second, we can get prepared for entering the society after we graduate from the university. We want to be more experienced. (SM_CHN_PTJ1_088_B2_0)*

An alternative possibility is the influence by the unique social system of Mainland China. It should be noted that societies of Mainland China, Hong Kong, Taiwan, Singapore, and Japan are shown to be preferring a collective pattern (Hofstede, 1984; Dadkhah et al., 1999). Given that learners from Hong Kong and Taiwan share the same culture and L1 background with those from Mainland China and that there is a strong relationship between collectivism and socialism, the abnormally frequent 1st person plural pronouns in CHN might be ascribed to the unique social ideology of Mainland China, i.e., the ideology of socialism strengthens the collectivistic sense of the English-language learners from mainland China, resulting in their heavy use of plural 1st person pronouns.

The next best indicators for CHN are two indices of Conjunction (ConjRate & ConjNum). Different from Pl1Pron, conjunctions in CHN occur infrequently in comparison with other corpora. This contradicts Liu and Miao's (2011) finding that Chinese students appear to overuse connectives in their oral English. To take the casual conjunction because as an example, L2 learners from CHN seem to be underusing because when expressing a causal relationship in their monologues. In the two hundred texts of CHN, only 109 are found while this number is 205 in ENS, 275 in HKG, 161 in TWN, 180 in JPN and 305 in SIN.

Other indicators that are worth noting in Table 4 are three indices related to uncertainty, including UncertainRate, PeacockRate and WeaselRate. The three types of indices are discussed together here as they both can be used to mark

the uncertain epidemic stance of language users (Vincze, 2013), especially when referring to a piece of unknown or vague information. The frequent uncertainty-related words in CHN may demonstrate that English learners from Mainland China tend to produce ambiguous information regarding the prompts compared with learners from other groups. The low uncertainty of learners from CHN can be partly observed from the above text that is cited for the discussion of *we*. In this text, the author seems to favor the use of hedge *can* and weasel words *some* and *something*.

2. Hong Kong

Table 5 shows that statistical indices are the most effective indicators for detecting monologue texts of HKG. We can see that HKG stands out with its long SentenceLength (this feature is included here as it is not directly affected by the length of monologue) and high TokenRate (per sentence). This might reflect the high informativeness of the monologue texts from HKG. The high ratios of coordination (CoordRate), condition clause (ConditionRate), conjunction (ConjRate) in HKG also reinforce the above argument since in English those features are commonly used for packing information. AttRate (attributive rate) is another indicator that is relevant to the high informativeness of monologues of HKG given that its major function is to add extra information to a linguistic form that is being modified. High ratios of passive subjects (PassiveSubjRate & PassiveSubjNum), and objects (ObjRate), but low ratios of 1st person singular pronouns (Sg1PronRate) and general personal pronouns (PronRate) indicate that HKG, in comparison with the other corpora, display an impersonal and low author-involved style.

TABLE 5
TOP 15 INDICATORS FOR HKG (SPOKEN MONOLOGUE)

Rank	Feature	F-score	CHN	HKG	TWN	JPN	SIN	ENS
1	SentenceLength	48.89	23.19	**33.96**	29.51	16.99	30.04	26.03
2	TokenRate	46.64	28.32	**38.77**	33.92	19.96	33.51	29.34
3	AttRate	43.26	0.91	**1.25**	0.76	0.61	1.12	0.893
4	PassiveSubjRate	34.29	0.04	**0.12**	0.04	0.05	0.08	0.05
5	ObjRate	27.25	1.48	**2.00**	1.92	1.01	1.59	1.41
6	NounNum	23.46	22.80	**26.89**	18.76	16.14	30.21	29.90
7	LemmaNum	22.58	61.31	**70.44**	53.71	42.03	81.40	80.87
8	ConjNum	19.34	7.30	**10.76**	7.37	5.34	13.18	12.71
9	Sg1PronRate	18.44	0.20	**0.16**	0.31	0.30	0.14	0.21
10	SentenceNum	18.37	5.78	**4.66**	4.20	4.70	5.57	6.82
11	SubjNum	17.99	12.52	**13.84**	12.35	9.20	15.81	18.83
12	PronRate	16.83	0.13	**0.12**	0.15	0.14	0.12	0.14
13	PassiveSubjNum	16.68	0.22	**0.47**	0.16	0.20	0.44	0.35
14	ConditionNum	16.04	0.84	**1.39**	0.65	0.26	1.89	1.37
15	CoordRate*	14.06	1.14	**1.53**	1.52	0.89	1.22	1.22

The most noticeable feature of the monologue text below is the long sentences with low repetition (compared with monologue text cited in Mainland China section): the author expresses his/her 157-words-long opinion towards the topic "*to ban smoking in the public place*" within only four sentences by using coordinate and, conditional if and conjunction but. Besides, compared with the texts cited in the "Mainland China" section, the author, in stating the supporting arguments, seems to favor an impersonal and objective tone. As can be seen from the text, the author only uses I once to express his/her claim, whereas in the rest of the monologue, s/he makes heavy use of inanimate subjects (*cigarette*, *smoking*, *the government*, *other crimes*, etc) or animate subjects with no specific referents (*some people*, *people*, *no one*, *the public*, etc). Moreover, the author is prone to using passive voice without concerning the negative effect brought by the repeated expressions though there are alternative options, for example, active voice. As can be seen from the text, the author uses the same passive construction three times with minor alteration (*smoking should be completely banned*, *if smoking is completely banned and smoking cannot be completely banned*). In general, the author speaks the way s/he writes since even if we remove the repetitions and pauses, the monologue text still will be a traditional informational argumentative essay.

> *These days some people say that smoking should be completely banned at all restaurants in country, and I disapprove with this statement. First of all, cigarette is a part of the tax income of the government; if smoking is completely banned people may somehow see – uh – may still smoke secretly and they will try to buy the cigarette from the black market, and therefore the government may – uh – gain a less – uh – tax income and also there will be – uh – other crimes rising. Apart from that also smoking is one of the source of the poor air quality – to the air pollution and it may also cause damage to our health, but there is no right for the government to deprive people from – uh – smoking. It is the right of the public to choose their own habits and no one can stop them from this one right, and for smoking cannot be completely banned at all public.... (SM_HKG_SMK2_032_B1_2)*

The high informativeness in HKG is likely due to the influence of learners' high language proficiency as it is not surprising that advanced learners are more capable of making long and informative sentences, which, to some extent, can only be achieved on the basis of mastering certain degree of the target language. Moreover, the low rate of pronouns of HKG can be explained in the same vein. Pronouns are learnt at an early stage and massively utilized in daily communication. Advanced learners might be motivated to replace these early adopted lexicons to avoid repetition

when they have alternatives. For the salient impersonal and informative style manifested by low frequency of pronouns, learners from Hong Kong might ignore the spoken nature of monologue and mix it with argumentative writing. As a matter of fact, Hong Kong learners have long been proved to prefer low impersonality, especially in writing, which has been documented and criticized by Hyland (2002a & 2002b). The monologue text cited above, to some extent, highlights that the author transplants the ways of composing argumentative writing into performing monologue.

3. Taiwan

The monologues contained in TWN are relatively short. In Table 6, TWN and JPN, two corpora that are close in terms of average monologue length (69.42 for TWN, 58.14 for JPN), are quite similar on a number of indicators but are distinct from the other corpora which contains longer monologues, implying that pure frequencies are severely skewed by monologue length. Therefore, the discussion to Taiwan spoken monologues excludes AttNum, NounNum, AdjNum, HedgeNum, ConjNum, AdvNum, Punct, AdvNum, AdverbNum, WeaselNum, DemPronNum, VerbNum and EpitemicNum though they appear in the forefront of the top 15 indices of TWN, and concentrates on indicators with ratio values, viz, PronRate (pronoun rate), Sg1PronRate (1st person singular pronoun rate), SubjRate (active subject rate), and PrivateVerbRate (private verb rate).

TABLE 6
TOP 15 INDICATORS FOR TWN (SPOKEN MONOLOGUE)

Rank	Feature	F-score	CHN	HKG	TWN	JPN	SIN	ENS
1	LemmaNum	101.16	61.31	70.44	**53.70**	42.00	81.40	80.87
2	AttNum	93.86	4.42	4.745	**2.56**	2.51	5.675	5.51
3	NounNum	92.49	18.76	22.80	**26.89**	16.14	30.21	29.90
4	TokenNum	90.67	110.43	128.10	**94.56**	68.88	151.86	160.27
5	AdjNum	84.29	7.43	7.45	**5.06**	4.28	9.43	4.28
6	SentenceNum	56.99	5.78	4.66	**4.20**	4.70	5.57	6.82
7	HedgeNum	56.99	10.41	11.96	**8.24**	6.23	14.00	13.85
8	PronRate	49.82	0.13	0.12	**0.15**	0.14	0.12	0.14
9	ConjNum	49.04	7.30	10.76	**7.37**	5.34	13.18	12.71
10	AdvNum	44.48	7.52	9.98	**6.92**	4.65	14.29	12.47
11	Sg1PronRate	39.79	0.20	0.16	**0.31**	0.30	0.14	0.21
12	Punct	37.21	24.55	18.58	**14.88**	12.46	18.08	20.87
13	SubjRate	35.34	2.61	3.61	**3.83**	2.24	3.10	2.96
14	AdverbNum	33.64	8.32	10.47	**8.06**	5.68	14.55	13.99
15	PrivateVerbRate	33.52	0.22	0.20	**0.25**	0.18	0.18	0.20

As in HKG, PronRate and Sg1PronRate are also listed as the best indicators of TWN, with the difference that they appear here due to their high values. More specifically, TWN is distinctive from the other corpora for its relatively high percentage of pronouns and 1st person singular pronouns. This reflects Taiwan learners' overt involved style in monologue, and different from the involved style of their peers from Mainland China, they rather speak from an individualistic perspective. Taiwan learners' involved style is also manifested by the high rate of another top indicator – PrivateVerbRate. Being a special type of verb that are specifically used for "overt expression of private attitudes, thoughts, and emotions" (Biber, 1988, p.105), Private verbs are viewed as an involvement feature that is typically found in spoken registers. In the text cited below, 1st person singular pronoun I collocates frequently with private verbs such as think, want, know to express author' opinions, feelings, and beliefs. Unlike the author of the Hong Kong monologue cited previously, the author here deeply involves him/herself to the discussion.

> *I think smoking should be banned at all restaurants in this country because I am a nonsmoking people, and I really don't like the smell of smoke and if in – in the inside, I – inside the room of a restaurant I don't want to have second smoke – second-hand smoke because I know it's bad for my health and people smoking – people who smoke it – it – they don't – they don't have – they don't know how the smell – how bad....(SM_TWN_SMK1_040_B2_0)*

The last indicator of TWN with ratio value is SubjRate, an index concentrating on the active subjects in a syntactic structure. The high rate of this indicator probably also relates to the involved style of Taiwan learners. From above two texts, it can be seen that the central opinions and supporting argument are full of subjectivity of the authors, which is in a stark contrast to the impersonality of the Hong Kong monologue text cited before.

B. Written Essay

This section focuses on describing the characteristics of the three corpora. Due to the fact that the six corpus shows a minor difference in terms of writing length (*Mean = 234.3, DS = 7.9*), indices are attached with equal importance no matter they are with frequency information or ratio values.

1. Mainland China

The written essays in CHN show certain similarities to the monologues in CHN with regard to their best indicators. As is evident from Table 7, CHN stands out among the surveyed corpora mostly because of its frequent 1st person plural pronouns (Pl1PronRate and Pl1PronNum) and features that mark authors' uncertainty including peacocks

(PeacockRate and PeacockNum), weasels (WeaselRate & WeaselNum), and hedges (HedgeNum). Likewise, the writings contained in CHN, therefore, can be characterized by collectivism-oriented, involved, and lacking certainty.

TABLE 7
TOP 15 INDICATORS FOR CHN (WRITTEN ESSAYS)

Rank	Feature	F-score	CHN	HKG	TWN	JPN	SIN	ENS
1	Pl1PronRate	119.07	**0.165**	0.053	0.087	0.110	0.028	0.031
2	Pl1PronNum	112.70	**5.425**	1.600	2.895	3.675	0.885	5.425
3	PeacockNum	54.27	**8.535**	7.005	7.280	6.330	6.065	6.610
4	WeaselNum	46.19	**30.560**	28.100	28.565	23.725	27.815	24.220
5	PeacockRate	42.82	**0.035**	0.030	0.032	0.028	0.024	0.029
6	SentenceNum	42.49	**14.820**	13.890	13.345	16.400	11.350	9.365
7	SentenceLength	40.21	**16.837**	17.835	18.334	14.294	22.722	25.900
8	WeaselRate	39.30	**0.128**	0.118	0.125	0.106	0.113	0.107
9	SuasiveVerbNum	37.77	**0.865**	1.245	1.335	1.220	1.840	1.595
10	TokenRate	37.51	**18.933**	19.945	20.657	16.148	25.086	28.096
11	SuasiveVerbRate	36.89	**0.027**	0.040	0.042	0.038	0.058	0.052
12	Punct	36.20	**30.375**	28.630	29.470	29.855	26.095	20.010
13	CoordRate	33.87	**0.519**	0.548	0.674	0.459	0.761	1.062
14	ConjRate	30.43	**0.057**	0.055	0.059	0.063	0.064	0.078
15	HedgeNum	28.83	**22.265**	20.110	20.045	18.375	20.035	20.905

As for the last two indices relating with suasive verb (SuasiveVerbRate & SuasiveVerbNum), CHN shows a lower rate than other corpora. The retrieving results indicate that this low rate of suasive verb of CHN is mainly contributed by the infrequent use of suasive verb agree. Although the essays contained in investigated corpora share a similar size, only 50 agree are found in CHN while the frequencies of agree are 71 in ENS, 62 in HKG, 112 in TWN, 138 in JPN, and 80 in SIN. Besides, suasive verbs are "those which intend to effect a change of some sort" (Grant & Ginther, 2000, p.131), whereas most of the agree found in investigated corpora are used in a way similar to private verbs since they frequently collocate with personal pronouns to express author's personal opinion: to agree with the propositions in discussing (as shown in the following essay). Therefore, the rare presence of suasive verb in CHN is more likely the result of learners' frequent use of private verbs which mark the involvement of the author.

> *In my opinion, I am strongly agree with the idea that, it is important for our college students to have a part-time job. Now we can see ... So why they do these jobs? What advantages benefit us? I think the main reason is money. With the improvement of our living standards, a lot of study material are expensive than before … By doing so, we can also have the ability to travel or buy some items we like. We also hope that through this way, we can no longer dependent on our parents. In addition, we can also accumulate some social experience. From kindergarten to high school, what we learned is totally the knowledge from books. In this way, it does a lot benefits to our future jobs. In a word, there are many advantages for our college students to do some jobs. Not only make money, but also develop our independence... (W_CHN_PTJ0_002_B1_1)*

This 201-word essay includes 13 1st person plural pronouns (we and our) but only three 1st person singular pronouns (I and my) and one 3rd person pronoun, indicating a pronounced collectivistic sense as well as a highly involved writing style. Concerning the terms that mark uncertainty, this essay shows a heavy use: weasel terms - *a lot*, *many*, *some* and *a lot of*, peacock terms - *important*, *everywhere* and *totally*, and hedge – *can*.

2. Hong Kong

The top indicators of HKG overlap largely with the features included in the first dimension of Biber's multidimensional model (1988) on explaining the variation across spoken and written language. This dimension is a continuum ranging from the informational production pole to the involved production pole. Written genres are characterized for high-frequent informational features with negative loading and therefore have a relatively low score on this dimension while spoken genres usually contain more involvement features with positive loading and thus often obtain a high score. In the present study, Table 8 shows that features with positive weights, including pronouns (PronRate, PronNum, Sg1PronRate & Sg1Pron), present tense (PresentNum & PresentRate), and private verbs (PrivateVerbRate & PrivateVerbNum), appear less frequently in HKG than in other corpora, whereas Nouns (NounNum & NounRate), a typical feature marking informativeness, are found more in HKG, thus indicating that HKG are more informative than other corpora. Besides, the left indicators also add evidence to this finding. In Table 8, HKG, compared with other corpora, is low on the rate of Subjective Sentences (SubjRate), function words (FunctionRate) and doxastic verbs (DoxiasticRate), but high on the rate of content words (ContentRate), which mirrors that in writing L2 learners of English from HKG, in general, favor an impersonal and informative writing style. This finding is exactly in tune with the observation by Hyland (2002a, 2002b) and Kobayashi and Abe (2016) on the English writings of Hong Kong Chinese learners. Again, a text is cited for a better explanation:

> *Recently, banning smoking at all restaurants becomes a hot topic of discussion. Government also implants the law to ban smoking in public area. The supporter points out that banning smoking in restaurants can ... Other people such as restaurant owners argue that this can cause loss in term of*

profit. As a whole, the benefit of banning smoking in all restaurants outweighs its drawback … Customers still inhale second hand smoke when sitting at non-smoking area due to the air-conditioner system of restaurant. Passive smoking can even greater harm to health of non-smokers than active smoking. That explains why smoking should be banned in all restaurants. Some people argue that some smokers may not go to restaurants after banning smoking and result in the loss of profit. They also add that some restaurants such as discos and bars as most customers at such places are smokers … Passive smoking in bars and discos still cause harm to their workers. (W_HKG_SMK0_036_B1_1).

This essay demonstrates a highly informative and impersonal style – a similar style found in the monologue text of HKG. We can see that except for they and their which respectively occur only once, no other pronouns are used in this 235-word-long essay, indicating that the author tends to establish a detachment instead of an involvement relationship with the discussion. Besides, as they are in the cited HKG monologue text, inanimate subjects or subjects without definite referents (*the government*, *banning smoking*, *passive smoking*, *people*, *supporters*, etc.) are used in this essay with a high frequency. At the same time, features associated with subjective sentiments or opinions like subjective objects and private verbs that are common in other corpora, seem to be avoided by the author. Moreover, reporting verbs or verb phrases that are typical in academic writing such as *point out* and *argue*, make the essay academic-like.

TABLE 8
TOP 15 INDICATORS FOR HKG (WRITTEN ESSAY)

Rank	Feature	F-score	CHN	HKG	TWN	JPN	SIN	ENS
1	PronRate	119.03	0.106	**0.078**	0.107	0.113	0.080	0.105
2	PresentRate	94.77	0.310	**0.240**	0.302	0.403	0.247	0.300
3	PresentNum	92.53	10.050	**7.545**	9.910	13.115	7.880	9.360
4	PronNum	89.95	25.565	**18.505**	24.74	25.295	19.66	23.475
5	Sg1PronRate	88.34	0.090	**0.058**	0.144	0.174	0.061	0.169
6	SubjRate	87.39	1.600	**1.501**	1.822	1.597	1.741	2.422
7	NounRate	82.54	0.243	**0.267**	0.242	0.247	0.255	0.214
8	Sg1Pron	79.94	2.855	**1.795**	4.710	5.665	1.920	5.175
9	ConjRate	68.11	0.057	**0.055**	0.059	0.063	0.064	0.078
10	PrivateVerbNum	66.72	5.310	**3.605**	5.730	5.710	3.660	5.110
11	PrivateVerbRate	57.81	0.163	**0.115**	0.173	0.175	0.115	0.167
12	ContentRate	56.61	0.532	**0.550**	0.531	0.533	0.533	0.512
13	FunctionRate	56.43	0.467	**0.449**	0.468	0.466	0.466	0.488
14	NounNum	55.37	58.625	**63.775**	55.71	55.415	63.025	48.890
15	DoxasticRate	39.93	0.011	**0.008**	0.011	0.017	0.006	0.011

3. Taiwan

The highest-ranked indicators for identifying TWN are two indices relevant to attributives (AttRate andAttNum). Table 9 reveals that TWN obtains a comparatively small value on the two indicators, reflecting that the essays in TWN are not as informative as those in other corpora. The next indicator is UncertainNum. This indicator, along with another high frequent uncertainty-related indicator, WeaselRate, signals the high uncertainty of Taiwan learners. The low value of TWN on the two indicators, ProperRate and ProperNum, might also be associated with learners' high uncertainty considering that proper nouns are nouns that have specific referents or "unique denotation" (Quirk et al., 1985, p.288). The indicator next to ProperNum is Punct. TWN is shown to be overusing punctuations, and the reason behind might be the underuse of another indicator of TWN - ConjNum. The following indicators are NegationRate and NegationNum, which is quite surprising as they are the only two sentiment-related indices found to be functioning for the corpus identification. The relatively frequent negation words in TWN possibly indicate that learners of English from Taiwan prefer negative structures to highlight their statements. Other noteworthy indicators are three indicators that are often connected to involved, informal writing style: PrivateVerbRate, PrivateVerbNum and PronRate. TWN is shown to have high values on the three features and therefore might be regarded as involved.

I think it is important for college student to have a part-time job. Most of our parents don't give us too much money. For example, I can 4000NT per month from my mother. But minus the daily basic meal costs and MRT costs, I only have 500NT left to watch a movie with friends or to buy CDs or other things. Apparently, that is not enough. Also, college students will soon join into social live and finding a formal job. If we college students don't get a part-time job and experienced working life, we must tend to be frustrated after we graduate and start working...if you want to buy something, you need to pay effort on it … some people will earn even more-is enough. Through this process, college student will know that earning money is not easy, and will be precious the money their parents earn and things they have got. It is not only a good experience before graduate, but also a nice education. That's why I think part-time job is important.

Above essay reflects an overly involved and interactive style of TWN. In addition to *I* which marks the involvement of the author, the author repeatedly uses pronouns such as *you*, *we*, *our* and *us*, to connect the topic under discussion with the readers. For the use of attributives, we can see that most of the attributives are limited to simple attributives and numeral adjectives (*important*, *basic*, *enough*, etc.). Besides, the author uses several negative sentences to highlight the

necessity for students to find a part-time job (*our parents don't give us*…, *if students don't get*…, *though part-time didn't*, etc.), instead of directly demonstrating the benefits of a part-time job. Lastly, the text also displays a high degree of uncertainty, which is manifested by a number of uncertain expressions including *most of*, *much*, *other things*, etc.

TABLE 9
THE TOP 15 INDICATORS FOR TWN (WRITTEN ESSAYS)

Rank	Feature	F-score	CHN	HKG	TWN	JPN	SIN	ENS
1	AttRate	35.05	0.832	1.005	**0.809**	0.515	1.238	1.374
2	AttNum	29.13	11.750	13.290	**10.145**	8.200	13.520	12.210
3	UncertainNum	27.46	2.635	2.590	**2.660**	1.820	2.035	2.100
4	ProperRate	23.44	0.106	0.178	**0.091**	0.133	0.180	0.165
5	ProperNum	23.34	0.455	1.330	**0.415**	0.870	1.650	1.385
6	Punct	22.79	30.375	28.630	**29.470**	29.855	26.095	20.010
7	NegationRate	21.87	0.014	0.013	**0.017**	0.017	0.011	0.013
8	NegationWord	21.68	3.275	2.970	**3.855**	3.695	2.640	2.980
9	ObjNum	20.99	16.615	15.395	**16.630**	14.625	14.590	13.730
10	PrivateVerbNum	20.23	5.310	3.605	**5.730**	5.710	3.660	5.110
11	ConjNum	19.18	13.620	13.005	**13.550**	14.215	15.845	17.580
12	VerbRate	18.90	0.135	0.132	**0.143**	0.145	0.131	0.137
13	PronRate	18.28	0.106	0.078	**0.107**	0.113	0.080	0.105
14	WeaselRate	17.82	0.128	0.118	**0.129**	0.106	0.113	0.107
15	PrivateVerbRate	16.22	0.163	0.115	**0.173**	0.175	0.115	0.167

V. CONLUSION

The present study is a comparative study that is conducted into the three learner Englishes from Greater China via the approach of NLI. It yields an average classification accuracy of 60% on spoken language and 59.8% on written language. With the two relatively satisfactory accuracies, this study further probes into the most significant identifying features.

The results reached in the exploration of the most discriminating indicators of the three learner Englishes reveal that despite the commonly recognized differences between written and spoken genres, the three learner Englishes from Greater China show a high homogeneity between spoken monologues and writing essays, or in other words, learners speak the ways they write or vice versa. In general, the major characteristics of the three learner Englishes from Greater China can be summed up as follow:

1) The spoken monologues and written essays by L2 learners of English from Mainland China are characteristic for high collectivistic involvement and uncertainty, low informativeness, and underuse of conjunctions.

2) Learner English of Hong Kong is significantly distinct from that from Mainland China and Taiwan though learners from the three regions share the same L1. It is found that the spoken monologues and written essays by L2 learners from Hong Kong feature high informativeness and an impersonality.

3) Both spoken monologues and written essays by L2 learners from Taiwan reveal a high degree of individualistic involvement. Besides, their English essays are found to be high on uncertainty and negation but low on informativeness and the usage of conjunctions.

The implication of this study is that this study benefits future corpus linguists for doing the division on L1 Chinese learners of English from the three major regions of Greater China. The overall classification accuracy around 60%, along with the most discriminating features for each corpus, partly provides an empirical basis for doing such a division. Another major implication is that it will hopefully increase learners' genre awareness towards spoken and written registers. As is noted earlier, the three groups of L1 Chinese learners of English exhibit a consistent way when delivering spoken monologues and writing argumentative essays, which could be problematic in many aspects. For learners from Hong Kong, their overly informative and formal style in monologue might make their speech tedious and less interesting for the listeners, and in writing, their overuse of features marking formality and informational focus, including objectivity, impersonality, low self-involvement and passive voice, may be detrimental to gain credibility and acceptance for their arguments from readers (Hyland, 2002a). As for learners from Mainland China and Taiwan, the problems primarily lie at their salient involved style in writing. Excessive involvement will make their written productions full of subjectivity, and more importantly novice-like since it is often connected to the lack of genre awareness – an issue that novice writers often meet (Gilquin & Paquot, 2008).

This study is not immune from limitations. Firstly, the corpora adopted are relatively small in size. Although the language data are highly comparable from the point of view of corpus linguistics, the limited size might confine the generalization of the findings. It is hoped that future studies could mend this limitation by conducting similar research on a larger learner corpus. Secondly, the number of features chosen in this study is relatively small in comparison with other NLI studies. Therefore, it is hoped that more features will be taken into account for future classification of the three learner Englishes from Greater China. Lastly, in discussing the top indicators of the classification of three learner Englishes, only the top 15 features are selected, which might contribute to the missing of other important indicators.

REFERENCES

[1] Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

[2] Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2007). *Longman grammar of spoken and written English (6th version)*. London: Longman.

[3] Chafe, W. (1985). Linguistic differences produced by differences between speaking and writing. *Literacy, language, and learning: The nature and consequences of reading and writing*, *105*, 105-123.

[4] Crystal, D. (2011). Foreword. In A. Feng (Eds.), *English language in education and societies across greater China* (pp. xi- xii). Bristol: St Nicholas House.

[5] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

[6] Dadkhah, A., Harizuka, S., & Mandal, M. K. (1999). Pattern of social interaction in societies of the Asia–Pacific region. *The Journal of social psychology*, *139*(6), 730-735.

[7] Feng, A. (2012). Spread of English across greater China. *Journal of Multilingual and Multicultural Development*, *33*(4), 363-377.

[8] del Río, I., Zampieri, M., Malmasi, S. (2018). A Portuguese native language identification dataset. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 291-296).

[9] Gilquin, G., & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction*, *1*(1), 41-61.

[10] Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of second language writing*, *9*(2), 123-145.

[11] Håkansson, G., & Norrby, C. (2010). Environmental influence on language acquisition: Comparing second and foreign language acquisition of Swedish. *Language learning*, *60*(3), 628-650.

[12] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.

[13] Hofstede, G. 1984. *Culture's Consequences: International Differences in Work-related Values*. Beverly Hills, CA: Sage.

[14] Hyland, K. (2002a). Authority and invisibility: Authorial identity in academic writing. *Journal of pragmatics*, *34*(8), 1091-1112.

[15] Hyland, K. (2002b). Options of identity in academic writing. *ELT journal*, *56*(4), 351-358.

[16] Íñigo-Mora, I. (2004). On the use of the personal pronoun *we* in communities. *Journal of Language and Politics*, *3*: 27-52.

[17] Ishikawa, S. I. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner corpus studies in Asia and the world*, *1*(1), 91-118.

[18] Ishikawa, S. I. (2014). Design of the ICNALE-Spoken: A new database for multi-modal contrastive interlanguage analysis. *Learner corpus studies in Asia and the world*, *2*, 63-76.

[19] Jarvis, S., & Paquot, M. (2015). Learner corpora and native language identification. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 605–628). Cambridge: Cambridge University Press.

[20] Kobayashi, Y., & Abe, M. (2016). A corpus-based approach to the register awareness of Asian learners of English. *Journal of Pan-Pacific Association of Applied Linguistics*, *20*(2), 1-17.

[21] Koppel, M., Schler, J., & Zigdon, K. (2005). Automatically determining an anonymous author's native language. In *International Conference on Intelligence and Security Informatics* (pp. 209-217). Springer, Berlin, Heidelberg.

[22] Liu, Q., & Y. Miao. A corpus-based study on connective use in oral English by Chinese science and engineering majors. *Foreign Language World*, *32*(5), 16-23.

[23] Malmasi, S. (2016). *Native language identification: explorations and applications* [Unpublished Doctoral thesis]. Macquarie University.

[24] Malmasi, S., & Dras, M. (2014a). Arabic native language identification. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)* (pp. 180-186).

[25] Malmasi, S., & Dras, M. (2014b). Chinese native language identification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers* (pp. 95-99).

[26] Malmasi, S., & Dras, M. (2014c). Finnish native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2014* (pp. 139-144).

[27] Malmasi, S., Dras, M., & Temnikova, I. (2015). Norwegian native language identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 404-412).

[28] Mu, C., & Carrington, S. (2007). An investigation of three Chinese students' English writing strategies. *The Electronic Journal for English as a Second Language*, *11*(1), 1-23.

[29] Nisioi, S. (2015). Feature analysis for native language identification. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 644-657). Springer, Cham.

[30] Perkins, R. (2015). Native language identification (NLID) for forensic authorship analysis of weblogs. In M. Dawson, & M. Omar (Eds.), *New threats and countermeasures in digital crime and cyber terrorism* (pp. 213-234). IGI Global.

[31] Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London & New York: Longman.

[32] Straka, M., & Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies* (pp. 88-99).

[33] Tarone, E. (2007). Sociolinguistic approaches to second language acquisition research (1997–2007). *The modern language journal*, *91*, 837-848.

[34] Tarone, E. (2012). Interlanguage. In K. Brown (Eds.), *The encyclopedia of language and linguistics* (pp. 747–752). Boston, MA: Elsevier.

[35] Vincze, V. (2013). Weasels, Hedges and Peacocks: Discourse-level Uncertainty in Wikipedia Articles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing* (pp. 383–391). Nagoya, Japan. Asian Federation of Natural Language Processing.

Xiaoyun Li is currently a PhD candidate from the Department of Theoretical Linguistics at the University of Szeged, Hungary. He received his M.Sc. degree in Foreign Linguistics and Applied Linguistics from the Xi'an Polytechnic University, China, in 2017. His current research interests include corpus linguistics, learner language study, discourse analysis, English for Academic Purpose (EAP) study, and Natural Language Processing (NLP).