Exploring the ICNALE

—How to make the most of its design features—

Yukio TONO
Tokyo University of Foreign Studies

Abstract

This paper aims to explore the ICNALE for its unique design features. Especially, detailed learner proficiency information such as TOEIC scores and CEFR levels would be very useful in classifying texts. As the breakdown of subcorpora based on various factors is examined, the issue of data size across subcorpora will be revisited. In the latter half of the paper, more exploratory searches will be done in order to examine what kind of observations can be made about the nature of texts written by a variety of English users in the Asian regions.

Keywords

Learner profile, Text features, NS vs NNS comparison

I Introduction

Recently, more and more attention has been paid to the refinement of the learner profile information accompanied with learner corpora. Whilst Granger stressed the importance of strict design criteria at a very early stage of development in learner corpus research (Granger 1998), some of the external criteria such as university years could be misleading when it comes to describing learner proficiency levels (cf. Thewissen 2012). The newly developed corpus, the International Corpus Network of Asian Learners of English (ICNALE), is unique in this sense. It has not only conventional learner variables such as age, sex, major at university, years of L2 learning, but also a number of other numerical data such as standardized test scores (e.g. TOEIC or TOEFL), the Vocabulary Size Test, corresponding CEFR levels, motivation index scores by a questionnaire, among others. The project leader, Shin Ishikawa, claims that learner corpus comparisons across subcorpora will be made in a more sophisticated way, using these various learner-related factors.

In this paper, I will explore the ICNALE by first examining the learner proficiency

level information in detail and then looking into the similarities and differences across subcorpora using detailed learner profile information. The study is rather explorative in nature, and there is no specific research question to be addressed, but I hope that this explorative data analysis will reveal the wealth and breadth of the ICNALE as well as some areas which need some caution.

II Exploring the writers' proficiency level variables

Let me first look at the breakdown of subcorpora classified by their countries and their CEFR levels. See Table 1 for the results. The number is based on the latest release of the data. There are a few interesting points. Whilst it is normal that native speakers (ENS) are all classified into XX_0, i.e. above C level, there are cases in which speakers in an ESL environment, e.g. Singapore, are classified into either B1_2 or B2, which seems a bit odd, considering the status of English in their country.

Table 1 The ICNALE subcorpus breakdown (Country x CEFR)

Country	A2_0	B1_1	B1_2	B2_0	XX_0
China	50	232	105	13	0
ENS_Australia	0	0	0	0	17
ENS_Canada	0	0	0	0	28
ENS_NZ	0	0	0	0	13
ENS_UK	0	0	0	0	28
ENS_USA	0	0	0	0	114
Hong Kong	1	30	52	17	0
Indonesia	32	82	83	3	0
Japan	154	179	49	18	0
Korea	75	61	88	76	0
Pakistan	18	91	88	3	0
Philippines	2	11	176	11	0
Singapore	0	0	134	66	0
Taiwan	29	87	61	23	0
Thailand	119	179	100	2	0

The CEFR level was determined by the standardized test scores, but a closer look at the learner profile information revealed that there were more than 20 different test types shown in Table 2:

Table 2 Standardized test scores available for the ICNALE

A Level (General Paper)	Cambridg O Level	Cambridge	CEE
144	1	2	192
CSEPT	HKALE	IELTS	IEPT
4	14	15	1
N/A	NAT	NCAE	NCAT
1301	1	1	1
NMET	O Level (Eng Lang)	ONET	SAT
5	56	388	1
TEPS	TOEFL (iBT)	TOEFL (PBT)	TOEIC
17	42	27	586
UPCAT			
1			

If the learners' proficiency scores were not available or their test scores do not have any reported compatibility with the CEFR levels, a decision was made based on the scores of Paul Nation's Vocabulary Size Test (VST) and the pseudo CEFR levels were estimated using a linear model (Ishikawa, personal communication). Figure 1 shows the histogram of the VST results.

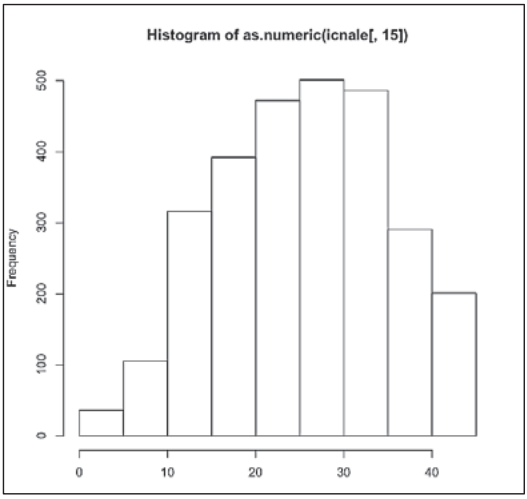

Fig. 1 The results of the Vocabulary Size Test scores

The highest score is 50, which is approximately 10,000 word level (cf. Nations' VST manual). Since Nation does not provide any alignment formula to convert the VST results to the corresponding CEFR levels, Ishikawa first converted the scores to TOEIC using a linear model prediction. Then the converted TOEIC scores are further aligned to the corresponding CEFR levels. Here we need to be cautious in assigning CEFR levels using TOEIC and VST. First, the VST is not claimed to be a valid measure in any sense

in terms of its compatibility with the CEFR levels. Table 3 indicates that if all the samples were assigned by the VST scores only, there are some mismatches between the CEFR levels by VST scores and those based on other external measures such as TOEIC (e.g. Group No. 13 underlined).

Table 3 The mean VST scores across countries and resulting CEFR levels (A2 & B1_1)

A2 level	VST scores	B1-1 level	VST scores
1 China	11.63	10 China	21.26
2 Hong Kong	7.00	11 Hong Kong	22.13
3 Indonesia	19.75	12 Indonesia	21.77
4 Japan	15.39	13 Japan	17.26
5 Korea	17.27	14 Korea	25.43
6 Pakistan	11.22	15 Pakistan	22.17
7 Philippines	9.00	16 Philippines	23.36
8 Taiwan	14.14	17 Taiwan	22.41
9 Thailand	10.77	18 Thailand	21.21

There is a possibility that the VST may not assess the overall proficiency levels properly. Also the alignment between TOEIC and the CEFR is said to be problematical, because TOEIC only looks at listening and reading mainly. Having said that, most alignment methods between standardized tests and the CEFR levels are generally a bit shaky and this is not the only problem with the ICNALE, thus we should bear in mind that we need to make informed choice of using the CEFR level information attached to the ICNALE.

Whilst learner variables such as the standardized scores and the CEFR levels were very useful, we should be reminded that the more variables we incorporate, the smaller the subcorpus size will be. The average text length of the two types of essays (“smoking” and “part time job”) is approximately 230 words, which means that in order to have 50,000 running words (the arguably conventional minimum size of the LC subcorpora), 217 samples will be required. As Table 1 shows, most subcorpus sizes are much smaller than this, thus the evidences, which we can find from the CEFR-classified data, ought to be limited, and it would only be fair to make a claim for those language features that appear frequently enough in those subcorpora.

III Exploring the lexical profile data across the ICNALE

3.1 Token averages and TTRs

With the above limitations in mind, let us examine text characteristics of the ICNALE subcorpora based on the CEFR levels. In order to make a comparison simpler, B1_1 and B1_2 were grouped together into B1, and four levels (A2, B1, B2 and XX=NS) were compared. Figure 2 shows the boxplots across the four levels. The numbers on the outliers indicate file numbers.

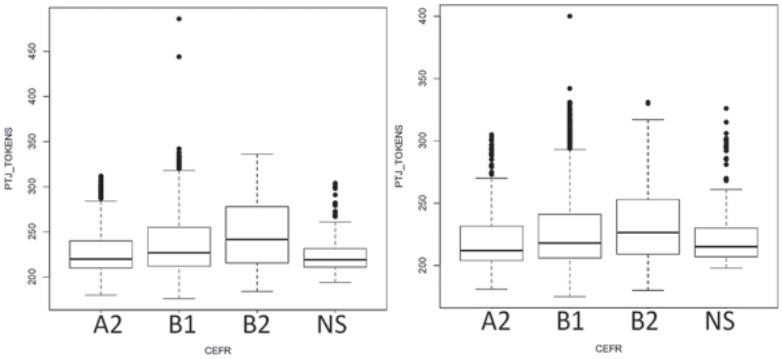

Fig. 2 Boxplots of the average tokens of the two tasks (PTJ & SMK) across CEFR levels

Since the average text length in ICNALE is strictly controlled, all the groups show approximately the same text length (220 to 240). Normally, given the same amount of time allotment (1 hour in the case of ICNALE), more advanced learners will write more. Thus fluency measures such as total text length make sense. With strict control of text length, one way to examine text characteristics is lexical richness measures such as Type/Token Ratio (TTR). Figure 3 shows the results of TTR across the groups

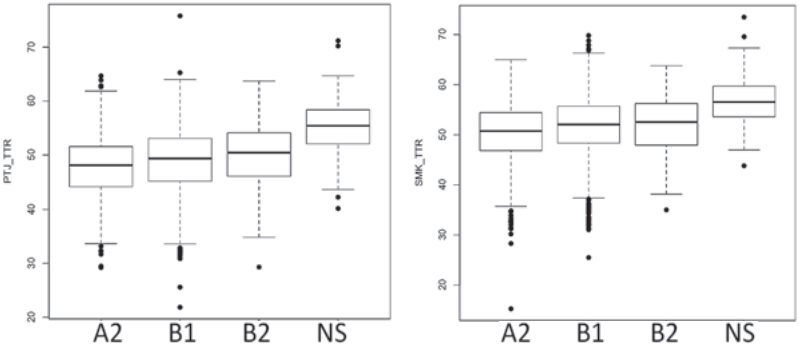

Fig. 3 Boxplots of the average TTR of the two tasks (PTJ & SMK) across CEFR levels

In both cases, Welch's one-way ANOVA showed a significant difference between NS and the other three groups (PTJ: $F(3, 540.46) = 109.67$, $p < 0.01$); SMK: $F(3, 544.52) = 86.49$, $p<0.01$).

It seems reasonable to assume that native speaker writers produced more lexically dense texts compared to the other groups. However, there is also a possibility that the control of the total text length could limit the performance of advanced users of English.

If native speakers write an essay in 250 words, they might have finished less than 30 minutes. If they had been requested to make maximum use of the allotted time, they might have been able to write much longer, elaborated texts.

Figure 4 shows plot of average tokens across different regions/countries. Overall, the average tokens range from 220 to 280, but a closer look at each distribution, there are major differences among countries. For example, A2-level Hong Kong learners tend to write much longer essays than the other higher levels, whereas most A2-level groups in the other regions/countries wrote less. English native speakers (ENS) seem to strictly keep the maximum number of words in the text.

Figure 5 shows the average tokens across countries/regions in terms of the two types of essay tasks. Although the range of the lowest and the highest token averages is about 30 words and not very large, Philippine and Singapore students tend to write longer than the other groups. English native speakers again keep the maximum text length faithfully.

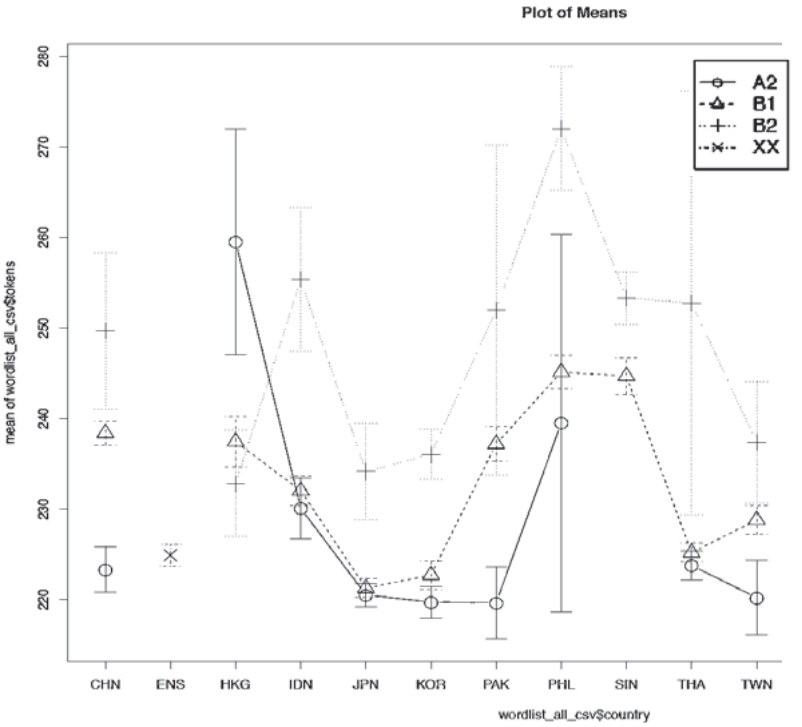

Fig. 4 Plot of average tokens across countries by CEFR levels

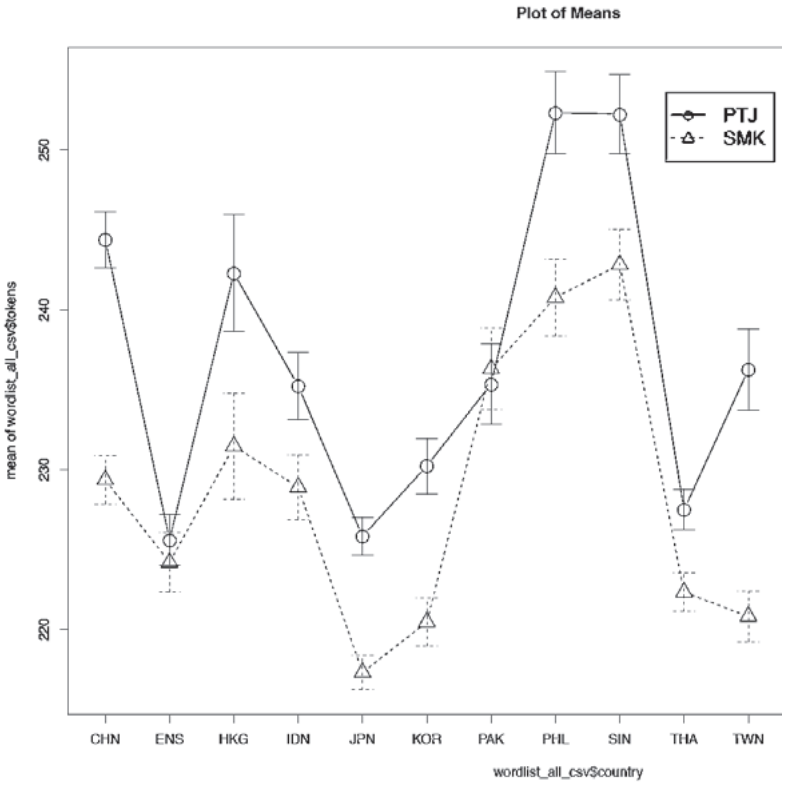

Fig. 5 The average tokens across countries by topics (PTJ vs. SMK)

3.2 Other lexical profile measures (Mean Word/Sentence Length, Sentence No. etc.)

In order to examine the characteristics of the texts across groups, basic lexical profile statistics such as mean word length (MWL), mean sentence length (MSL), the number of sentences in a text (SENTNO) were obtained, using WordSmith Tools 6.0. Figure 6 shows the correlation matrix among the three variables together with tokens, types and TTRs. The circle dots indicate PTJ, and the triangular dots indicate SMK. The original plots indicate the SMK dots in red and the PTJ dots in black.

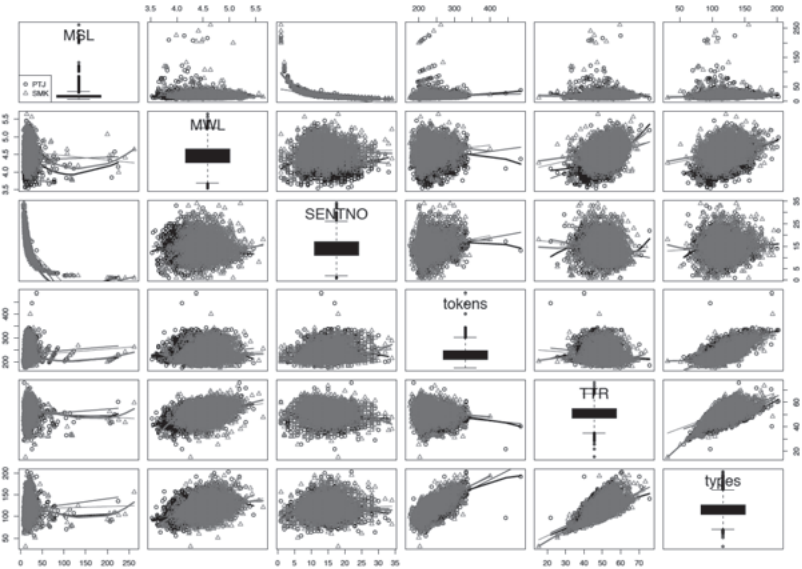

Fig. 6 Correlation matrix among six text characteristic measures by two topics

Figure 6 shows that there is no clear difference between the two types of essay tasks in terms of these measures. In fact, two essay types show very similar plots. The groups classified by the CEFR levels against native speakers, however, show some interesting pictures (see Figure 7). There is a marked tendency that B2 level learners used longer words than the lower levels. There are also positive relationships between mean sentence lengths and tokens, types and TTRs respectively. This means that although the text length is controlled, advanced level learners tend to use longer words, produce longer sentences, compared to lower-level learners. Unfortunately, the number of sentences (fluency measure) cannot show useful information this time, due to the fact that the essay size was strictly controlled. This was clearly seen in the negative correlation between the mean sentence length and the sentence numbers.

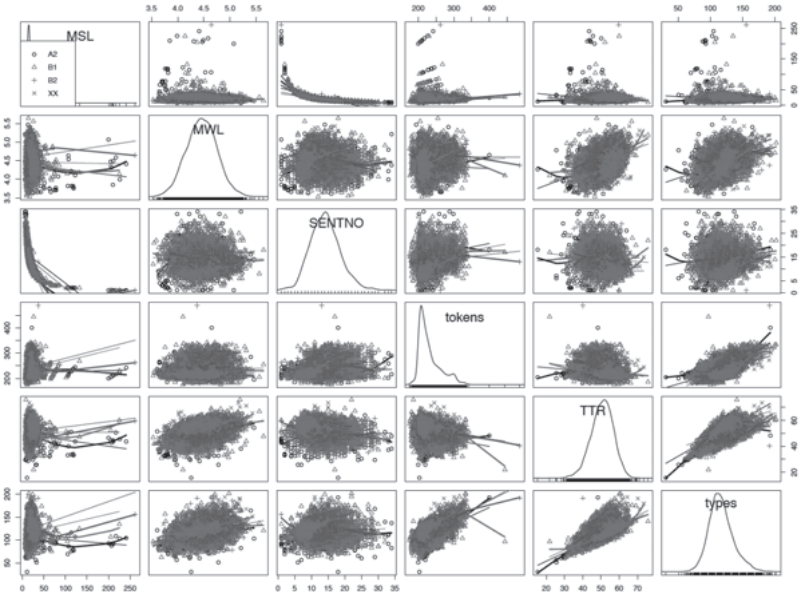

Fig. 7 Correlation matrix among six text characteristic measures by CEFR levels

IV What makes native speakers' performance different?

So far, most lexical profile information does not show a clear difference between native speakers and the other groups. Since the task was very strictly controlled, most fluency measures, such as total text length or sentence number failed to show the difference. In order to identify the area where native speakers' performance does differ from that of the other groups, further investigation was made to analyse what kind of vocabulary was used in each essay. To this end, all the texts were processed for semantic information, using Wmatrix (Rayson 2008). Tables 3 and 4 are the results of Wmatrix keyword analyses, comparing A2-level Japanese learners against native speakers and vice versa.

The first column shows the USAS tags, a semantic tagging system developed at Lancaster University, followed by raw and relative frequencies (%) in A2-level learners (A2) against native speakers (NS), log-likelihood ratio (LL) and category names. As I compared the list of semantic tag categories, interesting differences were found. First, the key semantic notions expressed in A2-level essays were a series of negative words, shown in the categories with asterisks, such as "Evaluation: Bad", "Negative", "Dislike", "Uninterested/bored", and "Damaging and destroying".

Table 3 The most frequent semantic tags that appeared in A2-level texts (JPN)

item	A2	%	NS	%	LL (+)	Semantic categories
F3	2434	7.55	936	4.4	210.87	Smoking and non-medical drugs
H5	161	0.5	7	0.03	117.78	Furniture
S2	822	2.55	307	1.44	77.92	People
A13.2	228	0.71	44	0.21	71.37	Degree: Maximizers
X3.5	155	0.48	30	0.14	48.36	Sensory: Smell
F1	1178	3.65	552	2.6	45.48	Food
O1.3	345	1.07	117	0.55	42.52	Substances and materials: Gas
A5.1-	220	0.68	67	0.32	34.57	**Evaluation: Bad***
Z6	838	2.6	398	1.87	30.07	**Negative***
N3.6	53	0.16	6	0.03	25.96	Measurement: Area
E2-	75	0.23	14	0.07	24.34	**Dislike***
Z4	292	0.91	114	0.54	24.09	Discourse Bin
X3.+	39	0.12	3	0.01	23.42	Tasty
Z99	136	0.42	40	0.19	22.92	Unmatched
M7	324	1	133	0.63	22.45	Places
X5.2-	19	0.06	0	0	19.24	**Uninterested/bored***
A1.1.2	161	0.5	56	0.26	18.61	**Damaging and destroying***
N4	250	0.78	102	0.48	17.71	Linear order
E2+	169	0.52	64	0.3	15.36	Like
X3.1	73	0.23	21	0.1	12.84	Sensory: Taste

Table 4 The most frequent semantic tags that appeared in NS texts (NS)

item	NS	%	A2	%	LL(+)	Semantic categories
Z2	222	1.04	156	0.48	55.36	**Geographical names**
S7.4+	119	0.56	60	0.19	52.1	Allowed
T1.1.1	58	0.27	14	0.04	50.31	**Time: Past**
Z8	2285	10.75	2902	9	39.96	Pronouns
X2.2+	85	0.4	47	0.15	32.61	Knowledgeable
X4.2	82	0.39	48	0.15	28.76	Mental object: Means, method
A13.3	280	1.32	271	0.84	27.63	Degree: Boosters
T2++	51	0.24	22	0.07	27.07	**Time: Beginning**
A13.5	20	0.09	2	0.01	25.54	Degree: Compromisers
A10+	48	0.23	24	0.07	21.25	Open; Finding; Showing
T1.1.3	140	0.66	121	0.38	20.54	**Time: Future**
N3.2+++	11	0.05	0	0	20.31	Size: Big
S8+	63	0.3	40	0.12	19.2	Helping
X3.4	44	0.21	23	0.07	18.33	Sensory: Sight
T1.3	57	0.27	36	0.11	17.54	**Time: Period**
L1+	30	0.14	12	0.04	17.28	Alive
I3.1	54	0.25	34	0.11	16.71	Work and employment: Generally
T3+	49	0.23	30	0.09	15.93	**Time: Old; grown-up**
A14	111	0.52	97	0.3	15.73	Exclusivizers/ particularizers
X6+	31	0.15	14	0.04	15.6	Decided

Since the topic is whether smoking should be banned at the restaurant, it is very likely that you express negative attitude toward smoking or banning of smoking in a public place. Either way, your essays tend to sound negative. The essays written by the A2-level learners naturally contained many negative words. The followings are some excerpts from the A2-level learners' writings:

A5.1 (bad) 192 occurrences

From the beginning , smoking is the **bad** things for health . The man who likes
r body . Moreover , smoke smell very **bad** and makes people who are near a smoke
es people who are near a smoker feel **bad** . In a restaurant , guests would like
people who smoke everyday often get **bad** health . For example , to feel tired
, people around smoking people feel **bad** , and is also get bad effects . The s
ng people feel bad , and is also get **bad** effects . The smells are bad for non
lso get bad effects . The smells are **bad** for non smoking people . They ca n't
y , some smoking people 's manner is **bad** . For example , after smoking , some
wever , the job making someone to be **bad** health is the worst job . Smoking is
ing is dangerous , and is completely **bad** . It is better that buy others than c

E2- (hate) 31 occurrences

gree with the statement . Because I **hate** smoking , I think that it most be co
e best reason of person not smoking **hate** Smokers is that smoke from smokers i
good people . It is not good thing **hate** people because they are smoking . Wh
king smell . I 'm nonsmoker , and I **hate** smoking smoke too . And generally sp
ke than smoker themselves . So they **hate** tobaccos smoke , all the more they a
But , the other hand , some guests **hate** smoking . I am often inquired if the
. Surely , nonsmokers are likely to **hate** smoke because it smells bad and hurt
ing . Even if we do n't smoke (and **hate** smoking) , we are snuffing smoke co
In fact , I do n't smoke at all and **hate** smoking . Then , there are some reas
ke reduce . We , who do not smoke , **hate** smoke which is produced by tobacco ,

On the other hand, native speakers' essays never had any negative words listed as salient semantic tags. Native speakers deliberately avoided direct negative statements about smoking. Instead, there were two interesting strategies emerging from the frequency counts of semantic tags. One is the use of the most salient key semantic tag, Z2, which denotes geographical names. In actual essays, native speakers frequently mentioned the situation of Japan and compared it against theirs.

all , that 's called democracy and **Japan** has a democratic government do n't t
n other parts of the world , but in **Japan** , there are still a lot more smoking
. Otherwise , come election time in **Japan** , the government might discover that
n ban smoking across restaurants in **Japan** . Instead of trying to force and enf
sed that it has taken this long for **Japan** to consider this ban . Smoking relat
ng related deaths and sicknesses in **Japan** are increasing , not decreasing and
o be mothered . I do n't know about **Japan** , but in Australia the government is
sometimes . From what I 've seen in **Japan** so far and the restaurants that I ha
ready . I think that it would be in **Japan** 's best interests to keep up with wh
ening in the world in this regard . **Japan** could easily introduce bans on smoki

Another strategy is shown in the use of semantic tags related to TIME. As shown in the tags, native speakers contrasted the past, the present and the future as they discussed the issue of smoking in a restaurant. The typical temporal markers for this strategy are time adverbials such as "already" (T1.1.1 Time: Past), "still" (T2++ Time: Beginning), and "will/one day/soon" (T1.1.3 Time: Future). By contrasting the past, present and future situations, native speakers approach the issue of banning of smoking from historical perspectives, which seems to be an effective way to avoid too direct a statement about this issue.

V Conclusion

The ICNALE is a unique collection of texts written by English users with various backgrounds. I found a comparison between native speakers and the other varieties especially interesting, since the topics and tasks are nicely controlled, one can approach a certain linguistic problem in a quite sophisticated way. On the other hand, a number of learner variables can be strengths, but also weaknesses as well. The blind use of learner variables could end up looking at a very tiny portion of the corpus, which might mislead users to a wrong conclusion. In such a case, it is always advisable to check whether the number of observations per subcorpora is sufficiently big in order to ensure the power of statistical test.

I do want to congratulate the release of the ICNALE and hope that this corpus will facilitate more rigorous research in this area and relevant fields of corpus linguistics and second language acquisition and foreign language teaching/learning.

References

Granger, S. (Ed.). (1998). *Learner English on computer*. London: Addison Wesley Longman.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, *13*(4), 519-549.

Thewissen, J. (2012). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *Modern Language Journal*, *97*(1), 77-101.