A Corpus-based Contrastive Analysis of Spoken and Written Learner Corpora: The Case of Japanese-speaking Learners of English

Mariko Abe (Sophia University)

1. Introduction

The purpose of this research is to investigate the variability of interlanguage that has been claimed in previous study by means of a corpus-based quantitative analysis. It aims to observe the style shifting of various grammatical features and word formation errors by tagging errors in 297 learners' data. Various studies have been undertaken to describe and explain the process of second language (L2) acquisition. Tarone (1983), for example, claims that interlanguage capability continuum of learners diverges with respect to the degree of attention to language form, so-called 'careful' style to 'vernacular' style. Additionally Tarone (1985) showed that learner's performance varies depending on the dissimilarity of task, whether a written grammar test, oral interview and oral narrative. Ellis (1987) has confirmed the style shifting in the L2 learners' use of the past tense. The variability of accuracy in past tense morphemes was observed in his study, when different amounts of planning time were set for a single narrative discourse task. Through his examination Ellis (1987) concluded that "so-called 'natural' order may not be a stable phenomenon" (p.1). In this study I compared the features of variability in L2 written and spoken corpus data based on the hypothesis that processing mode of learners affects their performance in L2. This research focussed on the analysis of the same task that used different production mode. The similarities and differences of learners' errors in each mode were examined mainly from the perspective of grammatical and word formation features. In addition, English proficiency level of learners is another factor added to this study. Although this addition was only possible for the spoken corpus data, we can still observe the performance of learners at different proficiency levels.

2. Corpus selection

The spoken data were extracted from the Standard Speaking Test (SST) Corpus (Tono et al. 2002). This test has 9 different levels to assess the speaking proficiency of learning English. Spoken data came from 100 examinees belong to SST level 2 to 9. Although there are 5 stages in this speaking test, only one of the stages, single picture description stage, was used. A single picture is chosen from 5 different pictures, and the examinees were asked to describe it in 2 or 3 minutes. The written data were all collected by the author using a similar type of picture description task. The 197 Examinees were all university students who have been studying English for 6 years. In addition, 31 out of 197 examinees have also taken the simple version of SST, and these results were used to create a L2 spoken subcorpus.

3. Data processing

All of the hand-written manuscripts for the written corpus were transcribed on a word processor by the researcher, and two sets of data were error-tagged according to The TAO Speech Corpus of Japanese Learner English Error Tagging Manual Ver.1.0. (Isahara, Saiga, and Izumi 2002). This error-tag set is divided into three main levels. The first level consists of three criteria: Word Formation Errors (WF), Grammatical Errors (G), Lexico-Grammatical Errors (LG), Lexical Errors (LXC), and Others (O). The second level is divided into part of speech and the other categories. The final level is the category for the errors as follows: inflection (inf), number (num), Japanese English (je), genitive (gen), agreement (ag), form (f), tense (tns), voice (vo), finite/infinite (fin), negation (ng), question (qst), modal (mo), quantifier (qnt), inflection (inf), position (pst), countability (cnt), complement (cmp), dependent preposition (dprp), word redundancy (rdd), omission (oms), misordering (odr), ambiguity (amb), and unnaturalness (unl). Spelling errors and word division errors were not normalized but manually tagged and included in the section of word formation <wf_o_o> error. Since the total corpus size of written and spoken data was almost the same therefore no normalization has been done in this research.

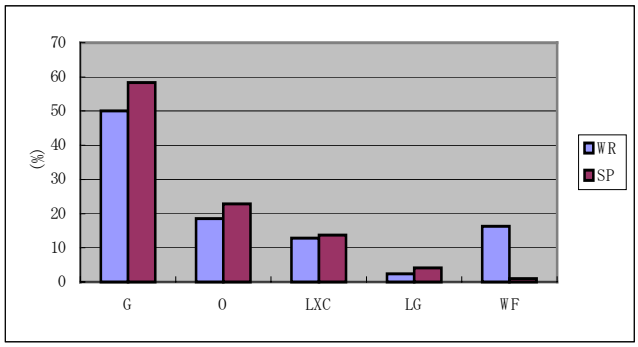
4. Data analysis: general error type

In this section, spoken and written errors are sorted by general error types. The following Table1 shows that grammar is the category with the most errors for Japanese learners of English. Consequently, it might be meaningful to inspect the effect of mode of production on grammatical errors. According to Ellis (1987), previous studies have not mainly investigated the interlanguage variability from the aspect of grammatical structure.

Table 1: Frequency of general error types

	WR		*SP*	
	Freq.	*%*	*Freq.*	*%*
Grammatical errors	1,136	50.02	473	58.40
Others	421	18.54	185	22.84
Lexical errors	291	12.81	111	13.70
Lexico-grammatical errors	53	2.33	33	4.07
Word formation errors	370	16.29	8	0.99
Total	2,271		810	

Figure 1: Frequency of general error types



Except for the category of word formation error (WF), there is no difference in frequency rank order between the spoken and written corpora. Almost all the category indicate similar percentage in overall errors, however, interestingly spoken data have higher error frequency than that of written, excluding the category of word formation. Since the criteria of G (Grammatical errors), LG (Lexico-grammatical error) and WF (Word formation) are subcategorised in part of speech, further examination will be provided in following sections.

5. Data analysis: part of speech

When we sort each category by part of speech, another interesting feature can be observed. Articles occupy the highest error rate in both production modes, followed by verbs, nouns, pronouns, adjectives, adverbs, and prepositions. Error frequency rate as well as rank order is almost identical in both modes, whereas spoken data have higher density of error in noun and pronoun. In addition to the low frequency of preposition, the error rate of adjective, and adverb is low. This may not mean that the grammatical rules associated with these parts of speech are mastered, but it also indicates that learners are unconsciously avoiding these rules. Therefore, considering this rank order, learners may underuse nouns and overuse pronouns, since according to Granger & Rayson (1998) the rank order of word category in non-native data is as follows: nouns, verbs, prepositions, articles, adjectives, conjunctions, adverbs, determiners and pronouns.

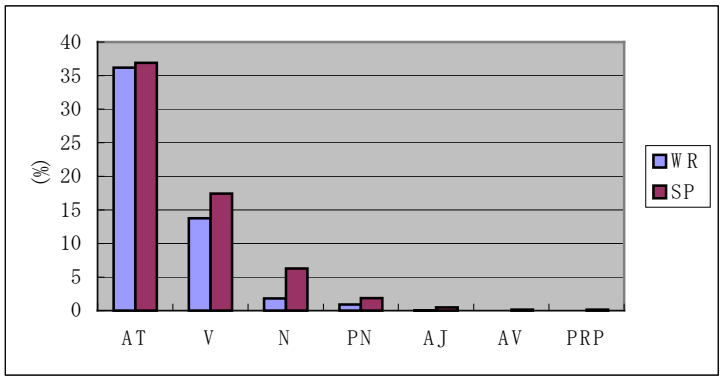
The following Table2 specifies the error distribution pattern in terms of part of speech, but it does not illustrate the accuracy rate. The analysis does not depend on the accuracy rate of learners, but on the frequency rate of the learners' error, so that we cannot generalise the degree of learner's avoidance and acquisition of certain grammar points. However, we can at least compare the error rate between that of written and spoken mode in various grammar categories of each part of speech. From this standpoint, detailed examinations are presented in the subsequent sections through the subcategories of nouns (N) and verbs (V) respectively.

Table 2: Error distribution pattern in part of speech

	WR		*SP*	
	Freq.	*%*	*Freq.*	*%*
AT	822	36.2	299	36.91
V	312	13.74	141	17.41
N	41	1.81	51	6.30
PN	21	0.92	15	1.85
AJ	1	0.04	4	0.49
AV	0	-	1	0.12
PRP	0	-	1	0.12

AT=article; V=verb; N=noun; PN=pronoun; AJ=adjective; AV=adverb; PRP=preposition

Figure 2: Error distribution pattern in part of speech



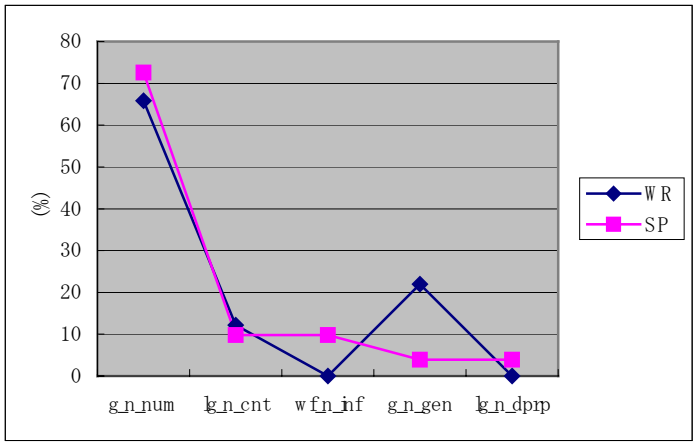
5.1. Noun

Singular common nouns are most frequently used in both spoken and written mode by native speakers of English (Leech et al. 2001), but this rule cannot be applied in this study. There are 5 subcategories in the criteria of noun and its error rate in different production modes is shown in Table3. In this section we will concentrate on the most striking errors, and on the items that have a striking dissimilarity between modes.

Table 3: Error frequency rate of subcategories in noun

	WR		*SP*	
	Freq.	*%*	*Freq.*	*%*
g_n_num	27	65.85	37	72.55
lg_n_cnt	5	12.20	5	9.80
wf_n_inf	0	-	5	9.80
Countability	(32)	(78.05)	(47)	(92.16)
g_n_gen	9	21.95	2	3.92
lg_n_dprp	0	-	2	3.92
Noun total	41		51	

Figure 3: Error frequency rate of subcategories in noun



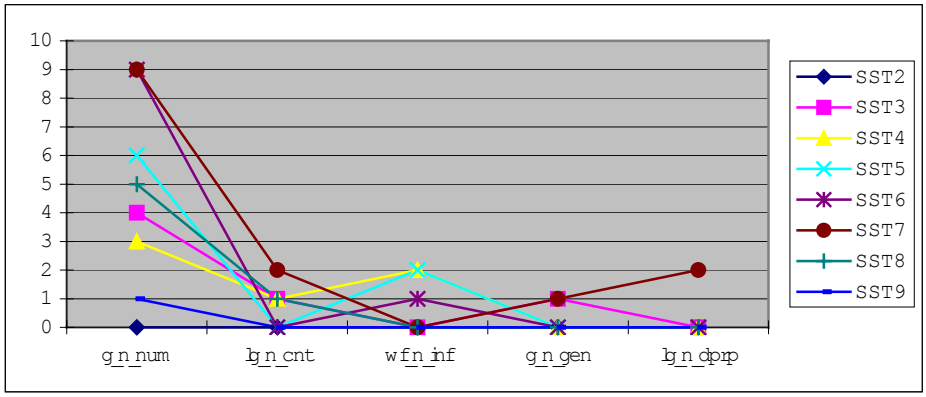
5.1.1. Countability

The following tags, <g_n_num> (many *book), <lg_n_cnt> (listening to a *music) and <wf_n_inf> (*childs) can be included under the category of countability. When Tarone (1985) examined the accuracy rate of plural markers 's', only form shift of morpheme was tested, and no variability was observed due to the different task. In this section, I would like to examine the error rate of the plural marker. Before comparing the error rate of <g_n_num> between written and spoken mode, errors such as "one of the *girl", "I like *dog" should be excluded so that I only count the error rate of plural marker. The result is that the error frequency for writing mode is 14 out of 27 (51.85%) and that of speaking mode is 18 out of 37 (48.65%). There seems to be no significant variability in two different modes as might be expected. However, there is one prominent finding when we sort the error of spoken mode in learners' proficiency level as shown in Table4.

Table 4: Error frequency in different SST level (SP)

SST	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	
g_n_num	0	4	3	6	9	9	5	1	37
Plural marker	(0)	(0)	(0)	(4)	(5)	(3)	(5)	(1)	(18)
Other	(0)	(4)	(3)	(2)	(4)	(6)	(0)	(0)	(19)
lg_n_cnt	0	1	1	0	0	2	1	0	5
wf_n_inf	0	0	2	2	1	0	0	0	5
g_n_gen	0	1	0	0	0	1	0	0	2
lg_n_dprp	0	0	0	0	0	2	0	0	2
Total	0	6	6	8	10	14	6	1	51

Figure 4: Error frequency in different SST level (SP)



As can be seen from the dispersion of the frequency in Table4, a morpheme change of plural marker is an error that often occurs in the intermediate levels but not in the novice level, whereas other grammar rules concerning countability occurs more often in the novice level than the intermediate level. As a result, we can hypothesise that when the necessity of attention to the form and grammar is low, intermediate learners tend to make errors; and when the necessity of attention is high they can avoid making errors. In addition to this, when the necessity of attention is low, because the rule is simple, novice learners make fewer errors; and when the necessity is high, because the rule is complicated, they have a tendency to make errors. By extracting Krashen's monitor model Ellis (1987) explains that easy rules can be monitored consciously causing the differences in style shifting. But this leads to another conclusion that intermediate learners may have an inclination to make much more errors over easily-learned rules than well-acquired rules and novice learners' error rate increases as the load of attention to the rules increases.

5.1.2. Variability in easily learned rules

Regarding the error over gender <g_n_gen> (*woman hair is black) and inflection <wf_n_inf> (*childs), both has variability between the modes. Also both can be categorised as easily learned rules, nevertheless unexpectedly, these errors have the opposite error frequency rate in production modes. While there are only 2 examples in spoken corpus, larger corpus with adequate learners' proficiency level information is needed to understand the cause of these differences.

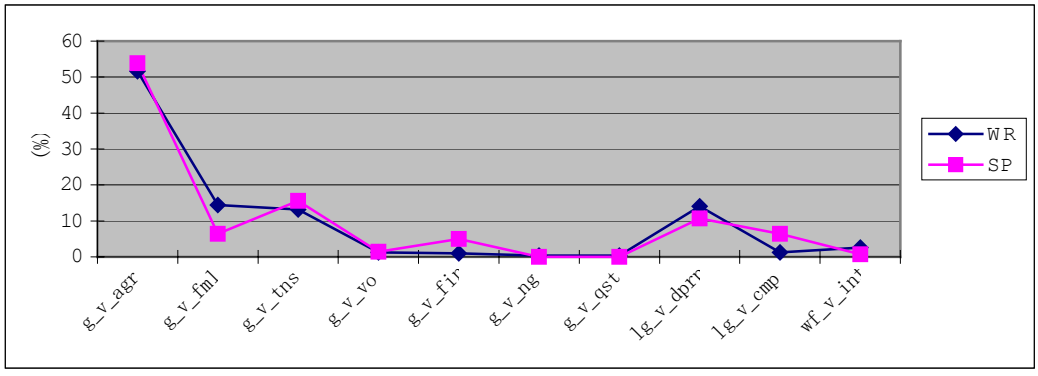
5.2. Verb

The subcategories for verbs and their corresponding error rates in different production modes are shown in Table5. In this section we will mainly focus on the most striking errors, and on the items that have dissimilarity between modes.

Table 5: Error frequency rate of subcategories in verb

	WR		*SP*	
	Freq.	*%*	*Freq.*	*%*
g_v_agr	161	51.60	76	53.90
g_v_fml	45	14.42	9	6.38
g_v_tns	41	13.14	22	15.60
g_v_vo	4	1.28	2	1.42
g_v_fin	3	0.96	7	4.96
g_v_ng	1	0.32	0	-
g_v_qst	1	0.32	0	-
g_v_mo	0	-	0	-
lg_v_dprp	44	14.10	15	10.64
lg_v_cmp	4	1.28	9	6.38
wf_v_inf	8	2.56	1	0.71
Total	312		141	

Figure 5: Error frequency rate of subcategories in verb



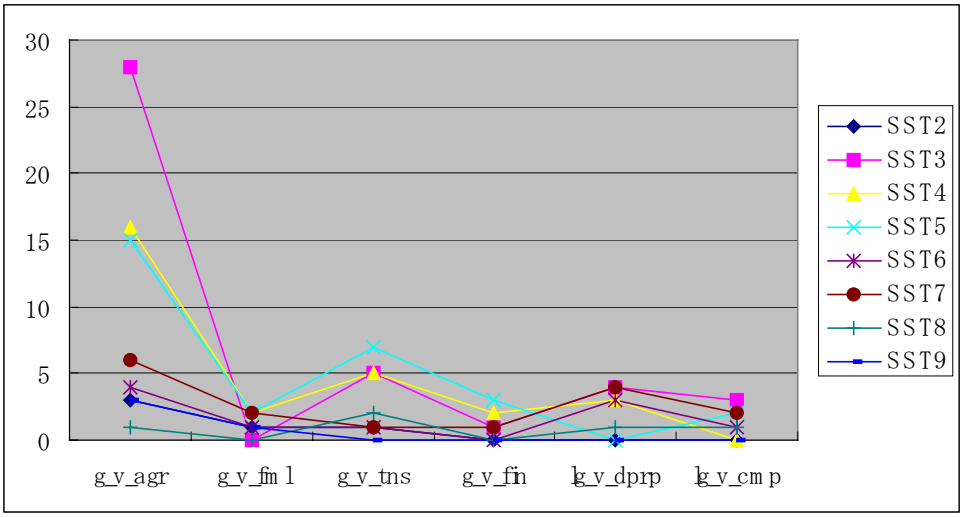
5.2.1. Agreement <g_v_agr> (there *are the lady) (cat *sleep in her bed)

In Tarone (1985) the accuracy level of third singular verb correction was examined, and variability according to tasks was also investigated. In this study the error frequency rate in different production modes does not diverge markedly. However, the error tag-set used in this research involves not only the third singular verb agreement errors but also every type of verb agreement errors. I excluded the agreement error for be-verbs and modal verbs from the data. The result is that the error frequency for writing mode is 95 out of 161 (59.00%) and speaking mode is 48 out of 76 (63.16%). Still a significant difference cannot be seen according to the production mode, but we have to remember the fact that Arabic learners of English were also included in Tarone's study. When we only focus on Japanese learners of English in her study, there is no apparent accuracy rate difference on different tasks. Although variability was found in neither different tasks nor modes, there is a striking result on error frequency in different SST level (see Tabel6).

Table 6: Error frequency in different SST level (SP)

SST	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	
G_v_agr	3	28	16	15	4	6	1	3	76
3rd person sing. verb	(2)	(21)	(10)	(12)	(0)	(3)	(0)	(0)	(48)
Be & modal verb	(1)	(7)	(6)	(3)	(4)	(3)	(1)	(3)	(28)
G_v_fml	1	0	2	2	1	2	0	1	9
G_v_tns	1	5	5	7	1	1	2	0	22
G_v_fin	0	1	2	3	0	1	0	0	7
Lg_v_dprp	0	4	3	0	3	4	1	0	15
Lg_v_cmp	0	3	0	2	1	2	1	0	9
Total	5	41	28	29	10	16	5	4	138

Figure 6: Error frequency in different SST level (SP)



Novice learners overwhelmingly tend to make errors over whole verb agreement, compared with higher-level learners; but when we focus on the third singular verbs and the other verbs separately, another distribution is found. If we follow the hypothesis concluded in section 5.1.1., 'error rate is in inverse proportion to the degree of attention to the rule in intermediate learners, and error rate is in direct proportion to the degree of attention in novice learners,' the agreement rule for third person singular verbs may be much more difficult than for be-verbs and modal verbs for Japanese learners of English.

5.2.2. Form <g_v_fml> (one boy is *listen to music and drinking something)

Interestingly, form error is much more frequently seen in the written mode, when learners can pay more attention to the form than spoken mode. Regarding the distribution of proficiency level in spoken

data, errors can be observed in almost every level in almost same number. When I sorted the errors into present particle errors and the others, there are 42 out of 45 (93.33%) in written mode and 5 out of 9 (55.56%) in spoken mode. In spoken data there is one example in each SST proficiency level 2, 4, 5, 7, and 9. How can we explain the fact that most of the errors are occur with present particle error, which is presumably easy learned, and with written mode? We need more adequate data and precise study to draw a conclusion. However, if learners do not have a high accuracy in simple mechanical verb form changing, we can assume that it is not fully acquired by the learners so that they need ample practice in verb form changing.

5.2.3. Complement <lg_v_cmp> (this person is teaching *how to them)

While Tarone (1985) found the variability in the aspect of third person singular direct object pronoun (D.O. Pro 'It') in different task, it was difficult to discover the variability in different mode. This is because not all the examples in both modes cannot be categorised as errors in D.O. Pro 'It'. In the written mode 3 out of 4 (75%) examples are of this type: "A woman wears *to it", "I will characterize *of her", and "A girl crosses *with her legs." Whereas in spoken mode only one example out of 9 (11.11%) can be found: "I don't know how to say *in English." Moreover, the error related with <lg_v_cmp> does not seem to be connected to learners' level. Considering the fact that there are not sufficient examples in both written and spoken mode, complement errors over the verb may have deeper connection with the misunderstanding of usage of verbs as far as Japanese learners are concerned. Consequently, analysis must be more focussed on the verbs themselves.

6. Data analysis: others

In previous sections we were able to observe the variability and consistency in some of the categories. In this section, we examine the general tendency of errors that have not been mentioned so far.

Table 7: The size of four corpora

	WR	*SP*	*WR-sub*	*SP-sub*
Picture	1	1-5	1	1
SST level	---	2-9	2-6	2-6
File	197	100	31	28
Tokens	17,863	17,222	3,221	3,908
Types	951	1,314	422	451
Type/Token/Ratio	5.32	7.63	13.1	11.54
Ave. Word Length	4.56	3.37	4.51	3.31
Sentences	1,507	986	265	252
Sent.length	10.5	16.51	10.42	14.71
sd. Sent. Length	5.45	14.7	6.26	13.66

Table 8: Rank frequency of error in each corpus

	WR			*SP*			*WR-sub*			*SP-sub*		
	17,863			17,222			3,221			3,908		
Rank		Freq.	%		Freq.	%		Freq.	%		Freq.	%
1	g_at	822	36.20	g_at	299	36.91	g_at	130	34.30	g_at	95	46.12
2	wf_o_o	362	15.94	lxc	111	13.70	lxc	52	13.72	g_v_agr	29	14.08
3	lxc	291	12.81	o_oms	102	12.59	wf_o_o	55	14.51	o_oms	28	13.59
4	o_oms	278	12.24	g_v_agr	76	9.38	o_oms	44	11.61	lxc	22	10.68
5	g_v_agr	161	7.09	o_rdd	51	6.30	g_v_agr	39	10.29	g_v_tns	5	2.43
6	o_odr	70	3.08	g_n_num	37	4.57	g_v_tns	12	3.17	g_pn	4	1.94
7	g_v_fml	45	1.98	g_v_tns	22	2.72	lg_v_dprp	11	2.90	lg_v_dprp	4	1.94
8	lg_v_dprp	44	1.94	g_pn	15	1.85	g_v_fml	7	1.85	g_n_num	3	1.46
9	o_amb	42	1.85	lg_v_dprp	15	1.85	o_amb	6	1.58	g_v_fin	3	1.46
10	g_v_tns	41	1.81	o_amb	13	1.60	o_odr	6	1.58	o_odr	3	1.46
11	o_rdd	31	1.37	o_odr	10	1.23	g_n_num	4	1.06	lg_v_cmp	2	0.97
12	g_n_num	27	1.19	g_v_fml	9	1.11	g_pn	3	0.79	g_av_pst	1	0.49
13	g_pn	21	0.92	lg_v_cmp	9	1.11	o_rdd	3	0.79	g_v_fml	1	0.49
14	g_n_gen	9	0.40	o_unl	9	1.11	g_n_gen	1	0.26	lg_aj_dprp	1	0.49
15	wf_v_inf	8	0.35	g_v_fin	7	0.86	g_v_ng	1	0.26	o_amb	1	0.49
16	lg_n_cnt	5	0.22	wf_n_inf	5	0.62	g_v_qst	1	0.26	o_unl	1	0.49
17	g_v_vo	4	0.18	lg_n_cnt	5	0.62	g_v_vo	1	0.26	wf_o_o	1	0.49
18	lg_v_cmp	4	0.18	g_av_pst	4	0.49	lg_n_cnt	1	0.26	wf_v_inf	1	0.49
19	g_v_fin	3	0.13	g_n_gen	2	0.25	lg_v_cmp	1	0.26	lg_n_cnt	1	0.49
20	g_v_ng	1	0.04	g_v_vo	2	0.25	wf_v_inf	1	0.26			
21	g_v_qst	1	0.04	lg_n_dprp	2	0.25						
22	g_aj_qnt	1	0.04	wf_v_inf	1	0.12						
23				wf_o_je	1	0.12						
24				wf_o_o	1	0.12						
25				lg_aj_dprp	1	0.12						
26				lg_prp_cmp	1	0.12						
	Total error	2271			810			379			206	

Errors due to the omission of one or more necessary words <o_oms> have relatively high frequency, but there is no significant difference in the error rate for the two modes. Regarding errors over word order <o_odr>, only the written corpus has a high frequency, but again there is not so much difference in error rate. This high error frequency in written corpus can be explained by novice low-level learners' consistent errors patterns. It is caused by the total misunderstanding of English structure such as "*Open door" ("door is open"), and this was counted as an error related word order. The ranking of error concerning ambiguity <o_amb> is almost equal but except for the subcorpus of speaking data. Errors that were difficult to categorise in any of the criteria were included in this group. Another criterion is errors related to redundancy <o_rdd>, and the error rate is dissimilar in each corpus. The most striking finding is that there is no error of this type in spoken subcorpus, while error rate in spoken corpus is fairy high. One possible explanation for this difference is that the proficiency level of learners in the spoken corpus is higher than that of the sub-corpus, consisting of learners from SST level 2 to 6. Therefore, we can presume that higher-level learners have a tendency to make redundancy errors. Lastly, the study did not show variability in the category of "other", but the comparison between different proficiency level corpora will be useful in further studies.

7. Conclusion

Through the detailed analysis on subcategories of nouns and verbs, we can observe the error rate difference in error over noun gender and verb form. Another finding that is noteworthy is that the error category, which has a high error rate, also has a large distribution among the learners' proficiency level, as can be seen from the example of errors related to countability and agreement. Also we were

able to acknowledge that the error rate is in inverse proportion to the degree of attention to rules for intermediate learners, and error rate is in direct proportion to the degree of attention to rules for novice learners. Granger and Rayson (1998) have shown in their research the resemblances of written and spoken production of learners, and they conclude that communicative approach is one of the factors that have an influence on "speech-like nature of learner writing" (p.130). Since the rise of this ELT methodology we may come to emphasise fluency but not accuracy, however, this study suggests that it is also necessary to take notice on learners' errors through instruction and feedback in the classroom.

Since not all the examinees of written data were able to take the SST test in this study, it was unfortunately impossible to investigate the correlation of written and spoken modes in terms of learner's proficiency level. More detailed data on learner's proficiency level and much larger corpora will be needed in future studies. Another drawback was that all the analysis comprised of the error rate, but not of the accuracy rate. Much more impartial examination could be done, if it were possible to determine whether learners are avoiding the certain usage or not. The last point is that subcategorised tag-sets that accord with learners' error tendency will be necessary for further study. Tag-sets for relative pronoun and conjunction, for example, were eliminated in this study. By analysing the similarities and differences between the two modes of learner corpora, I have arrived to identify the features of interlanguage variability in a more objective way, which will shed some light on the nature of the interlanguage development and possible implications for EFL pedagogy.

References

Ellis R 1987 Interlanguage variability in narrative discourse: Style-shifting in the use of the past tense. *Studies in Second Language Acquisition* 9: 1-20.

Granger S, Rayson P 1998 Automatic profiling of learner texts. In Granger S (ed), *Learner English on computer*. London, Longman, pp119-131.

Isahara H, Saiga T, and Izumi E 2002 *The TAO Speech Corpus of Japanese Learner English Error Tagging Manual Ver.1.0.*

Leech G, Rayson P, Wilson A 2001 *Word Frequencies in Written and Spoken Language*. London, Longman.

Tarone E 1983 On the Variability of Interlanguage Systems. *Applied linguistics* 4(2):143-163.

Tarone E 1985 Variability in interlanguage use: a study of style-shifting in morphology and syntax. *Language learning 35*: 373-404.

Tono Y, Kaneko T, Isahara H, Saiga T, Izumi E 2002 The Standard Speaking Test Corpus. *Studies in Lexicography* 11(2): 7-18.339