Investigating the sensitivity of the measures of fluency, accuracy, complexity and idea units with a narrative task

Chihiro Inoue
Lancaster University

Abstract

This study investigates the sensitivity of commonly used performance measures with spoken narrative performance by Japanese learners and native speakers of English. Five English native speakers and 24 Japanese learners at six different levels of the Standard Speaking Test [SST] were required to look at a sequence of pictures and then produce a narrative story in the past tense. The performance measures in this paper include measures of fluency, accuracy, syntactic complexity, lexical complexity, and 'idea units' that quantify how detailed the narrated story is. The 'sensitivity' of the measures in this study is defined as being able to highly correlate with the different levels of proficiency, and to discriminate among them. Statistical tests of Spearman's rho, Kruskal-Wallis, and *post hoc* LSD reveal that the only measure that fully satisfies the two conditions is the speech rate, a temporal fluency measure.

Papers from the Lancaster University Postgraduate Conference in
Linguistics & Language Teaching, Vol. 4: Papers from LAEL PG 2009
Edited by Grace Bota, Helen Hargreaves, Lai Chia-Chun & Rong Rong

Introduction

In the field of task-based research, it is a common practice to use various performance measures to quantify the aspects of spoken language (e.g. accuracy, fluency, complexity) so that comparisons among different performance can be made. This is often done by using narrative tasks with picture sequences. Despite of the popularity of performance measures, few studies have so far justified their use with the support from empirical evidence. This casts doubt on the validity of not only the measures but also of the research findings. This study therefore applies various performance measures to narrative performance by Japanese learners and native speakers of English so as to find out which of the commonly used measures are actually 'sensitive,' i.e. correlate highly with the levels of speakers as well as discriminate among them.

In this paper, various performance measures of spoken language are reviewed and discussed. The data source is discussed in detail, the research questions and the procedures are described. The results are presented and discussed and suggestions are provided for future research.

Performance measures

Fluency, Accuracy, Complexity

The current mainstream of task-based research deals with the effects on learner spoken performance of changing the conditions of task administration. This is based on the frameworks proposed by Skehan (1996, 1998) and Robinson (1995, 2001), both of whom attempt to explain language processing. Such studies include: manipulating the planning time, whether having the sequences in front or not, whether missing some pictures in a sequence or not, to name a few (Foster & Skehan, 1996; Mehnert, 1998; Norris, Brown, Hudson, & Bonk, 2002; Ortega, 1999; Robinson, 1995; Skehan, 1996, 1998; Skehan & Foster, 1999; Wigglesworth, 1997). The most common measures used to capture the differences in the quality of performance under the different conditions are those of fluency, accuracy and complexity.

Fluency

As Lennon (1990: 403) suggests, fluency measures can be classified into two aspects: *temporal measures* which deal with the speed of delivery, and *hesitation markers* that represent disfluency phenomena such as repetition and false starts. A number of researchers have attempted to identify appropriate measures of fluency. Kormos and Dénes (2004) offer the most recent credible results with the largest number of participants and the use of computer technology to identify length of pauses, leading to an empirical justification for using certain measures over the others. They conducted, by means of various measures of fluency, a validation study in which they correlated human ratings of how fluent the speech was with quantified results. Among the temporal measures that were validated in their study, the speech rate and the mean length of runs correlated the most with fluency ratings (Kormos and Dénes, 2004: 148). So, in the current study it is decided to include these two temporal measures to see if they correlate highly with the levels of the speakers and also discriminate among them.

The study by Kormos and Dénes (2004) proves that none of the hesitation markers are in accordance with the human ratings of fluency. However, their data did not involve native speaker performance and they did not investigate whether hesitation markers discriminated among different levels of speakers. So, in the current study, hesitation markers as fluency measures are included for the purposes of a more thorough analysis.

Complexity

Complexity is 'the extent to which learners produce elaborated language' (Ellis & Barkhuizen, 2005: 139), and is often concerned with syntactic and lexical aspects of narrative performance. Measures for syntactic complexity in previous studies include: the number of subordinate clauses per clause (Wigglesworth, 1997); the number of words per T-unit (Bygate, 2001; Daller, van Hout, & Treffers-Daller, 2003); the number of clauses per C-unit (Skehan and Foster, 1999; Foster and Skehan, 1996; Robinson, 2001) and the number of subordinate clauses per T-unit (Mehnert, 1998). The number of words per unit and the amount of subordination appear to be the two syntactic complexity measures that are most commonly used, and therefore will be examined in this study.

Some researchers use T-units as the unit for analysis, however, Ellis and Barkhuizen (2005) recommend using C-units or AS-units because they can take sub-clausal units into account. In addition, Foster, Tonkyn, and Wigglesworth (2000) argue that AS-units are more reliable than C-units. This is because AS-units can clearly distinguish among false starts, repetitions, and self-corrections (pp.362-363). Therefore, in this study AS-units are employed where units are necessary in the measures (the number of words per AS-unit and the average number of subordinate clauses per AS-unit).

For lexical complexity, the following measures have been employed in previous studies: type-token ratio (Robinson, 2001); mean segmental type-token ratio (Yuan & Ellis, 2003); D value (Kormos and Dénes, 2004); Guiraud index (Daller, et al., 2003). The use of type-token ratio [TTR] has been criticised for being greatly affected by the text length (Jarvis, 2002). After testing several measures of lexical complexity against curve-fitting statistical models, Jarvis (2002) justifies the use of D value over TTR, mean segmental TTR, and Guiraud index. So, D value will be used in this study.

D value and other lexical indices mentioned above mainly deal with the 'variety' sense of lexical complexity, and it does not suggest anything concerning how difficult or sophisticated the words are. For this aspect, three measures involving different word lists are selected: (1) Lexical Frequency Profile [LFP] (Laufer & Nation, 1995), (2) JACET8000 (JACET, 2003), and (3) the word lists from English textbooks that are used in junior and senior high schools (aged 12-18) in Japan. These measures aim to calculate the percentage of the words produced that belongs to the lists.

LFP and JACET8000 contain frequency-based word lists. LFP utilises General Service List (i.e. the list of the most frequent 1000 word families and the second 1000) plus Academic Word List (550 words that are frequent in academic texts across subjects). JACET8000 is a collection of eight lists of 1000 words derived from corpus-based research on English newspapers, textbooks, examinations and exam preparation books available in Japan. These two measures show the proportion of the English words that are frequently used in the learner's performance. The third measure using lists from Japanese high school English textbooks contains 600 words from the junior high list and an additional 1000 words from the senior high list. This measure shows the proportion of the words used at Japanese junior or high school levels. While LFP is chosen for its widespread use which enables comparisons with other studies involving non-Japanese learners, JACET8000 and the Japanese high school English textbooks vocabulary lists are expected to appropriately reflect the lexical use of Japanese learners.

Accuracy

Accuracy refers to how well the target language is produced according to its rule system (Skehan, 1996:23). The measures include: the percentage of error-free clauses (Skehan and Foster, 1999; Foster and Skehan 1996; Yuan and Ellis, 2003); the percentage of error-free C-units (Robinson, 2001; 2007b); the number of errors per T-unit (Bygate, 2001); Errors per 100 words (Mehnert, 1998); and the percentage of correct use of target features (Wigglesworth, 1997; Crookes, 1989; Skehan and Foster, 1997). In contrast to Kormos and Dénes (2004), none of the above research included a validation study. Ellis and Barkhuizen (2005) suggest that target-like verbal morphology is suitable for syntactic accuracy for focused tasks that are intended to elicit certain grammatical features. This is also the case for the SST narrative tasks (i.e. past tense) which are used in this study (explained later). Target-like verbal morphology is a *specific* measure for accuracy.

For *general* measures, the percentage of error-free clauses appears to be frequently selected. However, Bygate (2001) suggests that calculating the number of errors per unit might be more sensitive because it does not obscure the actual occurrences of errors, as is the case with counting error-free units. On the other hand, Mehnert (1998: 86) argues that the amount of errors per 100 words may be suitable for relatively lower proficiency speakers since it does not deal with the definition of clauses and units which is often problematic. As there is no way of knowing which of these measures will be sensitive, it is decided to include all four measures in this study.

Task-specific measure: Idea units

As the focus of this study is on narrative tasks, the organisation of performance should be structured as a narrative. Luoma (2004: 144) describes the requirements for narrative structures as follows: setting the scene; identifying the characters and referring to them consistently; identifying the main events; telling them in a coherent sequence. Luoma's description corresponds with Labov (1972: 360) who defined a minimal requirement for a narrative as 'a sequence of two clauses which are temporally ordered'. For Labov (1972: 363-370), a 'fully-formed narrative' will have the following features: *abstract* (summary of the whole story at the beginning), *orientation* (setting the time, place, characters and situation), *complicating action* (telling all the events in the story), *evaluation* ('indicating the point of the narrative'), *result or resolution* (telling what happened in the end), and *coda* (ending or concluding the narrative).

Appel (1984) employs a more detailed segmentation of events in analysing spoken narrative performance based on a picture sequence. To be more specific, Appel (1984: 188-194) investigates the amount of events or 'idea units' in the story that are covered in the learner's narration. She then compares between the first and second performances by the same learner. Ellis and Barkhuizen (2005: 154) argue that this measure is best suited for use when the performance is based on pre-determined content (e.g. by a picture sequence). The definition of an idea unit is 'a message segment consisting of a topic and comment that is separated from contiguous units syntactically and/or intonationally' (Ellis and Barkhuizen: 154). It is also possible to separate 'main idea units', which are the essential ideas to complete the story, from 'minor idea units' that are not essential but enrich the story (Ellis and Barkhuizen: 154). As this measure for representing narrative structure fits the purpose and data of this study, it is therefore applied here. The idea units of the narrative task used in this study are summarised in Table 3.

Table 1 (see next page) summarises the types of measures that have been discussed and examined so far:

Aspect	Measures	Definition
Fluency (Temporal)	Mean length of runs	Average no. of syllables produced in utterances between pauses of 0.25 seconds and above
	Speech rate	Total no. of syllables produced in a given speech sample divided by the amount of total time required to produce the speech sample (including pause time) expressed in seconds
Fluency (Hesitation)	No. of repetitions	No. of immediate and verbatim repetition f a word or a phrase
	No. of false starts	No. of utterances that are abandoned before completion
	No. of reformulations	No. of phrases or clauses that are repeated with some modification either to syntax, morphology, or word order
	No. or replacements	No. of lexical items that are substituted for another
Syntactic Complexity	No. of words per AS-unit	Average no. of words per AS-unit
	No. of subordinate clauses per AS-unit	Average no. of subordinate clauses per AS-unit
Lexical Complexity	D value	(calculated by CLAN program on CHILDES website at http://childes.psy.cmu.edu/)
	LFP 1, 2, 3	% of words listed in the Lexical Frequency Profile Vocabulary List 1, 2, and 3
	JACET 8000 Vocabulary List Lv. 1-8	% of words listed in the JACET Vocabulary List 1 to 8; these lists are based on British National Corpus as well as the frequent vocabulary found in English textbooks, newspapers, tests, magazines available in Japan.
	Vocabulary Lists for Junior and Senior High School Textbooks in Japan	% or words listed in the two lists for all the words appear in the textbooks that are used in junior high schools and senior high schools in Japan
Accuracy	Percentage of error-free clauses	% of clauses which do not contain any error to the total number of clauses
	No. of errors per AS-unit	No. of errors divided by the total number of AS-units
	Errors per 100 words	No. of errors divided by the total number of words produced divided by 100

	Percentage of target-like use of past tense	% of verbs in the past tense in the obligatory contexts (i.e. where past tense verbs are required)
Narrative Structure	No. of idea units encoded	The total numbers of the "main ideas" that are necessary to complete the story and "minor ideas" that enrich the story

Table 1 Measures Used in This Study

Data

A sensitive measure needs to differentiate between the quality of performance and the aspect that it is supposed to represent. For example, a fluency measure should discriminate the speakers with good fluency from less fluent ones. Also, in general, a speaker's fluency is likely to increase as their proficiency develops, thus the fluency measure may correlate with the learner's proficiency level. Therefore, the measures that are discussed in the previous section should be applied to narrative performance in conjunction with information about the speaking proficiency level. The following section describes the data used in this study that matches this requirement.

Japanese learner data

The Standard Speaking Test

The narrative performance and the task used in this study are derived from a speaking test administered in Japan, the Standard Speaking Test (henceforth, SST). The SST takes the form of a 15-minute structured conversation between an interviewer and a candidate, and includes a single picture description task, a role-play task, and a narrative task. The interview is recorded and rated by two independent raters who listen for certain rating criteria and decide on an overall single level of 1 (Novice Low) to 9 (Advanced)[1]. The raters consider how well the candidate is able to handle or demonstrate control over the following 5 criteria: *global tasks or functions* (asking and answering simple questions, narrating, describing in major time frames); *contexts* (from highly predictable common daily settings to more complex social situations); *content areas or topics* (from personal topics related to the immediate environment to a wide range of general interest topics); *accuracy* (in terms of

[1] Judging from the level descriptors by ACTFL-ALC Press (2000), it is assumed that SST Levels 1 to 9 approximately correspond to Below A1 to B2/C1 levels in the Common European Framework of Reference (CEFR).

grammar, vocabulary, pronunciation, fluency, sociolinguistic appropriateness, and discourse management); and *text type* (i.e. from words and sentences to complex sentences, paragraphs, and extended discourse) (ACTFL-ALC Press, 2000). In deciding an overall SST level, comprehension and interaction with the interviewer are also taken into account.

The number of candidates taking the SST in Japan is relatively small[2], even though it is the data source for the currently largest spoken corpus published in 2004 from Japanese learners of English. This may be because its interviews were recorded on tapes (for rating), making it easy for the ALC Press, the SST administrator, to easily obtain candidates' permission in order to use their data for research purposes (Izumi, Uchimoto, & Isahara, 2004). The corpus was developed in Japan by the National Institute of Information and Communications Technology [NICT] in cooperation with the ALC Press for purposes of research in natural language processing, second language acquisition, and language education (Izumi, et al., 2004). It was named the NICT JLE Corpus (NICT Japanese Learner English Corpus) with about 1.3 million words from the transcripts of 1,281 SST interviews. The Japanese learners' performance on a SST narrative task used in this study is derived from this corpus.

The narrative performance

The NICT JLE Corpus does not contain recordings, however, as this study attempts to examine aspects of fluency, not only the transcripts but also the corresponding recordings were obtained and analysed.

Firstly, 24 transcripts ranging from SST levels 4 to 9[3] were identified in the NICT JLE Corpus. They were selected because the candidates reported the TOEIC scores, a useful external measure of their English proficiency. Descriptive statistics of the number of transcripts at each SST level and their TOEIC scores are shown in Table 2 below. A request to obtain the corresponding recordings was sent to the ALC Press, which was accepted on the condition that the results should be reported back when certain results are obtained. The performance on the narrative task was taken out for analysis from the rest of the interview, in both the transcripts and the corresponding recordings.

[2] The grand total number of the candidates who took the SST from its inception (Jan. 1997) is about 30,000 (ALC Press, 2007).

[3] Judging from the level descriptors by ACTFL-ALC Press (2000), it is assumed that SST Levels 4 to 9 approximately correspond to CEFR Level A2 to B2/C1.

SST Lv.	N	Mean	SD	Min	Max
4	4	661.3	149.8	450	795
5	5	726.0	110.8	580	860
6	4	816.3	60.2	735	880
7	5	848.0	117.0	640	920
8	4	837.5	55.6	800	920
9	2	962.5	10.6	955	970

Table 2 Descriptive Statistics of the TOEIC Scores of Japanese learner data

English native speaker data

Although the NICT JLE Corpus contains a small 'native corpus' in which several English native speakers performed the SST tasks, there was insufficient native performance of the narrative task for this study. Therefore, 5 native speakers of English studying at Lancaster University participated in this study: 2 linguists and 3 non-linguists (one of whom was a former English teacher). They were met one by one in a quiet room, asked to look at the task and then narrate a story. There were no temporal limitations for the preparation or the story-telling. Their narration and responses were recorded and transcribed for the purposes of analyses.

Task

The narrative task in this study consists of a sequence of 6 pictures 'with a conflict' (ACTFL-ALC Press, 2000: 26) and is given to the SST candidates at an estimated Intermediate or Advanced level in order to elicit a narrative in past tense. The topic is an argument between two people following a car accident.

The narrative performance was prompted by the SST interviewer presenting the picture sequence and asking the test taker / candidate to narrate a story based on it in the past tense, starting with 'One day last week' (ACTFL-ALC Press, 2000: 25). After some planning time, usually less than 30 seconds[4], the candidate was asked to narrate. Since the task was still in use in the SST, neither the actual picture sequence nor the transcripts could be presented. However, the idea units identified in this study are summarised in Table 3 to give an idea of the task and the content of the narration. The idea units are identified from the native speaker performance collected in the study as recommended by Ellis and Barkhuizen (2005).

[4] ACTFL-ALC Press (2000: 19) states that each SST task stage should take 2-3 minutes in total, including explaining, presenting, planning, performing, and answering follow-up questions.

1	A guy was driving a car
2	Which … [car's description (e.g. *he recently bought*)]
3	He wanted to go to … [stating purpose]
4	He was in a hurry … [his state]
5	Another guy was riding a scooter
6	He was talking on the cell phone with a girl
7	He was not concentrating on the road
8	At a corner, they hit each other
9	Rider's cell phone was broken
10	It hit the wing mirror of the car
11	The car was okay
12	They got off their vehicles
13	They got angry
14	Rider complained about the broken scooter (tail light)
15	Rider complained about the broken cell phone
16	Rider requested compensation
17	Driver insisted that it was the rider's fault
18	Because the rider wasn't careful enough
19	The police was called
20	Because they could not resolve the argument
21	Driver explained what happened and insisted the rider was talking on the phone
22	Rider also insisted / gave up
23	Policeman took notes
24	Policeman understood / took the side of the driver
25	Policeman went back to report
26	They were asked to go to the police station
27	Driver drove off or left
28	Rider called the repairman
29	Rider's scooter was taken away by a truck
30	The repair cost would be dealt with by … [whoever]

Note. Shaded cells indicate the main idea units. The others are the minor idea units.

Table 3 Idea Units of the Narrative Task

Research Questions

Following the rationale presented at the beginning of Section 3, two research questions (RQs) are set in this study:

1. For SST levels 4-9 and native speaker level, which measures correlate highly with the levels, and discriminate among them?

2. If the measures do not correlate highly or discriminate among the levels, how can this be explained?

Procedures

The 24 transcripts from Japanese learners with TOEIC scores were extracted from the NICT JLE Corpus, checked for precision with recordings and modified where necessary. Five native speaker recordings were also transcribed. Then, two versions of the transcripts were produced. One was the full transcripts without non-words such as fillers (e.g. *erm*) and uncompleted single words that would not be recognised by the programs for lexical complexity measures. This version was used for fluency and lexical measures. The other version was removed of non-words and also was segmented into AS-units, which were without repetitions, fillers, and self-corrections, for measures of accuracy and narrative features. The measures were identified manually by the author, except for the lexical complexity measures which were calculated by existing programs.

For RQ1, all measures were correlated (Spearman's rho) with the SST levels (4-9) with the native speakers (treated as 'level 10' to enable statistical analyses). Also, to examine which measures discriminate among the SST levels, the Kruskal-Wallis Test and later a *post hoc* least significant difference [LSD] test were run.

For RQ2, the measures that did not correlate highly or discriminate among the levels were considered. The patterns were examined as to how the measures varied across the levels.

Results and Discussions

Descriptive statistics

Table 4 below summarises the descriptive statistics of the measures across the levels.

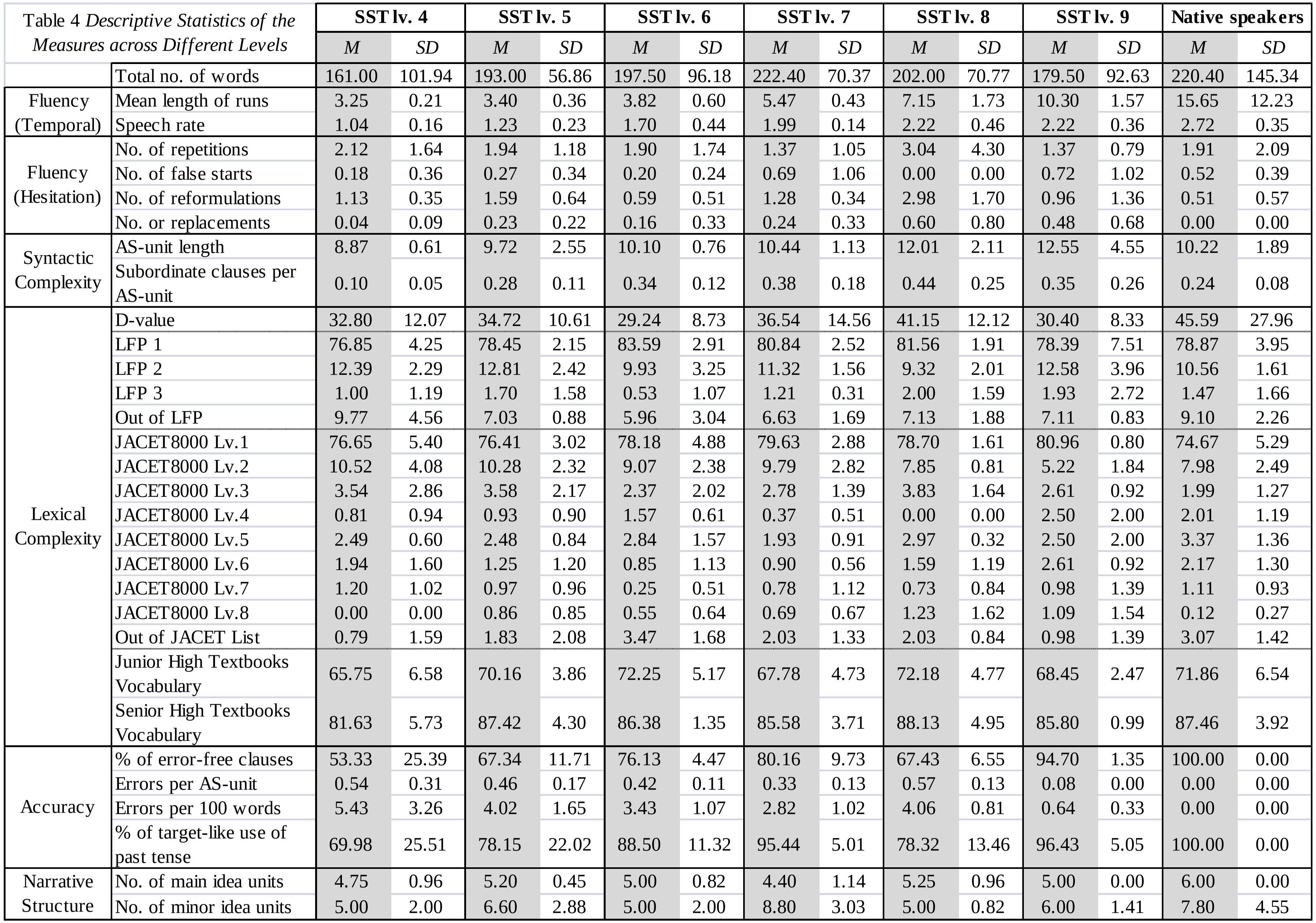


Table 4 *Descriptive Statistics of the Measures across Different Levels*		**SST lv. 4**		**SST lv. 5**		**SST lv. 6**		**SST lv. 7**		**SST lv. 8**		**SST lv. 9**		**Native speakers**	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
	Total no. of words	161.00	101.94	193.00	56.86	197.50	96.18	222.40	70.37	202.00	70.77	179.50	92.63	220.40	145.34
Fluency (Temporal)	Mean length of runs	3.25	0.21	3.40	0.36	3.82	0.60	5.47	0.43	7.15	1.73	10.30	1.57	15.65	12.23
	Speech rate	1.04	0.16	1.23	0.23	1.70	0.44	1.99	0.14	2.22	0.46	2.22	0.36	2.72	0.35
Fluency (Hesitation)	No. of repetitions	2.12	1.64	1.94	1.18	1.90	1.74	1.37	1.05	3.04	4.30	1.37	0.79	1.91	2.09
	No. of false starts	0.18	0.36	0.27	0.34	0.20	0.24	0.69	1.06	0.00	0.00	0.72	1.02	0.52	0.39
	No. of reformulations	1.13	0.35	1.59	0.64	0.59	0.51	1.28	0.34	2.98	1.70	0.96	1.36	0.51	0.57
	No. or replacements	0.04	0.09	0.23	0.22	0.16	0.33	0.24	0.33	0.60	0.80	0.48	0.68	0.00	0.00
Syntactic Complexity	AS-unit length	8.87	0.61	9.72	2.55	10.10	0.76	10.44	1.13	12.01	2.11	12.55	4.55	10.22	1.89
	Subordinate clauses per AS-unit	0.10	0.05	0.28	0.11	0.34	0.12	0.38	0.18	0.44	0.25	0.35	0.26	0.24	0.08
Lexical Complexity	D-value	32.80	12.07	34.72	10.61	29.24	8.73	36.54	14.56	41.15	12.12	30.40	8.33	45.59	27.96
	LFP 1	76.85	4.25	78.45	2.15	83.59	2.91	80.84	2.52	81.56	1.91	78.39	7.51	78.87	3.95
	LFP 2	12.39	2.29	12.81	2.42	9.93	3.25	11.32	1.56	9.32	2.01	12.58	3.96	10.56	1.61
	LFP 3	1.00	1.19	1.70	1.58	0.53	1.07	1.21	0.31	2.00	1.59	1.93	2.72	1.47	1.66
	Out of LFP	9.77	4.56	7.03	0.88	5.96	3.04	6.63	1.69	7.13	1.88	7.11	0.83	9.10	2.26
	JACET8000 Lv.1	76.65	5.40	76.41	3.02	78.18	4.88	79.63	2.88	78.70	1.61	80.96	0.80	74.67	5.29
	JACET8000 Lv.2	10.52	4.08	10.28	2.32	9.07	2.38	9.79	2.82	7.85	0.81	5.22	1.84	7.98	2.49
	JACET8000 Lv.3	3.54	2.86	3.58	2.17	2.37	2.02	2.78	1.39	3.83	1.64	2.61	0.92	1.99	1.27
	JACET8000 Lv.4	0.81	0.94	0.93	0.90	1.57	0.61	0.37	0.51	0.00	0.00	2.50	2.00	2.01	1.19
	JACET8000 Lv.5	2.49	0.60	2.48	0.84	2.84	1.57	1.93	0.91	2.97	0.32	2.50	2.00	3.37	1.36
	JACET8000 Lv.6	1.94	1.60	1.25	1.20	0.85	1.13	0.90	0.56	1.59	1.19	2.61	0.92	2.17	1.30
	JACET8000 Lv.7	1.20	1.02	0.97	0.96	0.25	0.51	0.78	1.12	0.73	0.84	0.98	1.39	1.11	0.93
	JACET8000 Lv.8	0.00	0.00	0.86	0.85	0.55	0.64	0.69	0.67	1.23	1.62	1.09	1.54	0.12	0.27
	Out of JACET List	0.79	1.59	1.83	2.08	3.47	1.68	2.03	1.33	2.03	0.84	0.98	1.39	3.07	1.42
	Junior High Textbooks Vocabulary	65.75	6.58	70.16	3.86	72.25	5.17	67.78	4.73	72.18	4.77	68.45	2.47	71.86	6.54
	Senior High Textbooks Vocabulary	81.63	5.73	87.42	4.30	86.38	1.35	85.58	3.71	88.13	4.95	85.80	0.99	87.46	3.92
Accuracy	% of error-free clauses	53.33	25.39	67.34	11.71	76.13	4.47	80.16	9.73	67.43	6.55	94.70	1.35	100.00	0.00
	Errors per AS-unit	0.54	0.31	0.46	0.17	0.42	0.11	0.33	0.13	0.57	0.13	0.08	0.00	0.00	0.00
	Errors per 100 words	5.43	3.26	4.02	1.65	3.43	1.07	2.82	1.02	4.06	0.81	0.64	0.33	0.00	0.00
	% of target-like use of past tense	69.98	25.51	78.15	22.02	88.50	11.32	95.44	5.01	78.32	13.46	96.43	5.05	100.00	0.00
Narrative Structure	No. of main idea units	4.75	0.96	5.20	0.45	5.00	0.82	4.40	1.14	5.25	0.96	5.00	0.00	6.00	0.00
	No. of minor idea units	5.00	2.00	6.60	2.88	5.00	2.00	8.80	3.03	5.00	0.82	6.00	1.41	7.80	4.55

Table 4 Descriptive Statistics of the Measure across Different Levels

'Sensitive' measures that correlated highly with and discriminated among the levels

Table 5 (see next page) summarises the results from Spearman's rho tests, the Kruskal-Wallis tests, and *post hoc* LSD tests. With *post hoc* LSD tests, the levels that showed a significant difference in their respective means are listed.

For RQ1, the 'sensitive' measures, with high correlation with and discrimination among the levels, will have one or two asterisks in the columns for Spearman's and Kruskal-Wallis tests (which means that the values were statistically significant), and have pairs of levels listed in the last column for LSD (i.e. the pairs of levels that were discriminated between) in Table 5 that given previously. The measures that satisfied these two conditions were: temporal fluency measures (i.e. the mean length of runs and the speech rate) and accuracy measures (i.e. the percentage of error-free clauses, errors per AS-unit, errors per 100 words, and the percentage of target-like use of past tense). Both of the temporal fluency measures showed very high correlation (r=.894 and .913), and the accuracy measures had moderately high correlation ($|r|$=.528 to .660).

However, when the LSD columns were closely examined, it became clear that most of these measures were only able to discriminate between a limited number of levels. Mean length of runs, one of the temporal fluency measures, only discriminated between SST levels and the native speakers. It could not differentiate among Japanese learners of English. The same applied to the percentage of target-like use of past tense, which discriminated among even less levels (SST levels 4, 5, 7 and the native speakers). If they can discriminate only between distant levels, such as the lower-level learners and the native speakers, these measures may only capture poorly the differences in learner performance.

The other three accuracy measures discriminated more pairs of levels; each had 10 pairs listed out of 20 possible pairs. Still, they seldom succeeded in differentiating adjacent learner levels (i.e. SST levels 4-5, 5-6, 6-7, etc.), especially at lower levels.

Compared to the rest of the measures discussed above, the speech rate, the other fluency measure, more often discriminated between levels that were closer to each other. Together with its high correlation (r=.894), the speech rate may be considered as the 'most sensitive' measure in this study.

Aspect	Measures	Spearman's		Kruskal-Wallis		*post hoc* LSD
		r	*p*	χ^2(6, 29)	*p*	Discriminated between[3]
Fluency (Temporal)	Mean length of runs	.913**	.000	24.35*	.000	4-NS, 5-NS, 6-NS, 7-NS, 8-NS
	Speech rate	.894**	.000	22.67*	.001	4-6, 4-7, 4-8, 4-9, 4-NS, 5-6, 5-7, 5-8, 5-9, 5-NS, 6-8, 6-NS, 7-NS, 8-NS
Fluency (Hesitation)	No. of repetitions	-.129	.504	0.83	.991	
	No. of false starts	.191	.321	6.33	.387	
	No. of reformulations	-.154	.425	13.97*	.030	4-8, 5-8, 5-NS, 6-8, 7-8, 8-9, 8-NS
	No. or replacements	-.052	.788	5.32	.503	
Syntactic Complexity	AS-unit length	.464*	.011	8.96	.176	
	Subordinate clauses per AS-unit	.264	.166	12.68*	.048	4-6, 4-7, 4-8
Lexical Complexity	D-value	.156	.418	2.63	.853	
	LFP 1	.105	.588	9.518	.588	
	LFP 2	-.271	.154	7.153	.154	
	LFP 3	.122	.528	3.043	.528	
	Out of LFP	.086	.656	7.630	.656	
	JACET8000 List Lv.1	.015	.938	5.64	.464	
	JACET8000 List Lv.2	-.384*	.040	8.31	.216	
	JACET8000 List Lv.3	-.216	.260	4.72	.580	
	JACET8000 List Lv.4	.202	.293	13.89*	.031	4-9, 5-9, 6-8, 7-9, 7-NS, 8-NS
	JACET8000 List Lv.5	.235	.220	5.33	.502	
	JACET8000 List Lv.6	.222	.247	6.98	.322	
	JACET8000 List Lv.7	.007	.969	3.29	.771	
	JACET8000 List Lv.8	.038	.843	6.28	.392	
	Out of JACET List	.244	.203	7.24	.299	
	Junior High Textbooks Vocabulary	.184	.340	4.85	.564	
	Senior High Textbooks Vocabulary	.245	.201	4.50	.609	
Accuracy	% of error-free clauses	.660**	.000	19.78**	.003	4-6, 4-7, 4-9, 4-NS, 5-9, 5-NS, 6-NS, 7-NS, 8-9, 8-NS
	Errors per AS-unit	-.553**	.002	17.75**	.007	4-9, 4-NS, 5-9, 5-NS, 6-9, 6-NS, 7-8, 7-NS, 8-9, 8-NS
	Errors per 100 words	-.638**	.000	17.72**	.007	4-7, 4-9, 4-NS, 5-9, 5-NS, 6-9, 6-NS, 7-NS, 8-9, 8-NS
	% of target-like use of past tense	.528**	.003	14.40*	.025	4-7, 4-NS, 5-NS
Narrative Structure	No. of main idea units	.368*	.049	9.91	.129	
	No. of minor idea units	.155	.422	6.04	.419	

Note. *Correlation is significant at 0.05 level.
**Correlation is significant at 0.01 level.

[3] Numbers indicate the SST levels. NS=native speakers.

Table 5 Results of Correlation and Discrimination with the Levels and Measures

It is clear now that most of the measures that more or less satisfied the two conditions of 'sensitivity' actually failed to demonstrate good discriminating power, the patterns that they display across the levels should be examined according to RQ2: why did they not discriminate well?

The patterns shown are based on the methods that were introduced in Table 4 for each measure. Firstly, Figure 1 displays the patterns of the temporal fluency measures.

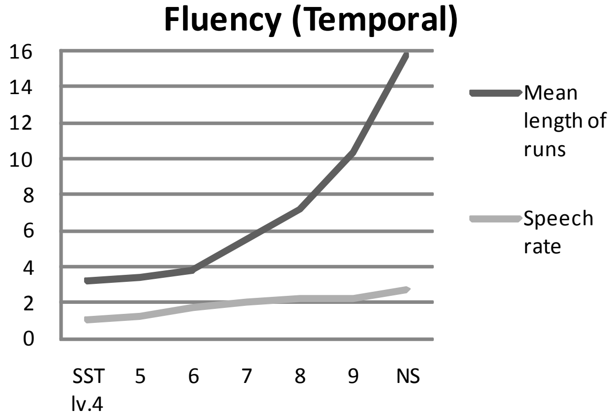


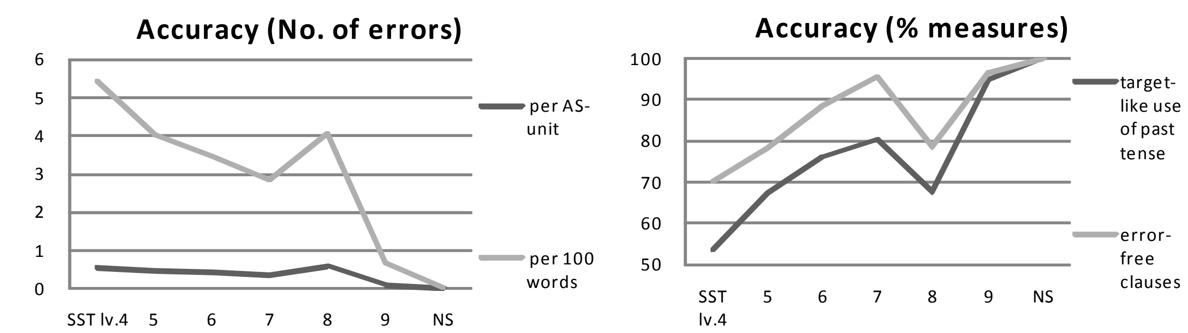
Figure 1 Patterns of Temporal Fluency Measure

While the speech rate increased steadily from SST level 4 to the native speaker [NS] level, the mean length of runs showed a drastic increase between SST level 9 and NS. It is probable that this caused the relatively small differences among other levels to be non-significant. This may suggest that, even if a measure correlates very highly, it does not necessarily guarantee its 'sensitivity' of distinguishing the different levels of performance by the learners.

Accuracy measures had moderately high correlation, but they did not discriminate among SST levels, either. Figures 2(a) and 2(b) present the patterns.

(a)

(b)



Figures 2(a), 2(b) Patterns of Accuracy Measures

The patterns were largely consistent across all four measures. There was a steady decrease in errors from SST level 4 to 7, however, at level 8, there was an increase in errors. This is surprising, as one might assume that the higher the candidate's level is the less errors they will make in their performance. One explanation is that, judging from the larger means in syntactic complexity measures at SST level 8 in Table 4, the SST level 8 candidates might have attempted to use more complex structures than the lower level ones but failed to use them accurately. It is possible that up to SST level 7, candidates may tend to avoid trying new structures or items and prefer speaking with the ones that they are familiar with and confident in using. To explore this hypothesis, we need to scrutinise the structures and error types with a larger sample size.

The rest of the measures

The rest of the measures were not proven 'sensitive' according to the operationalisation in this study. Some measures satisfied only one of the two conditions for being 'sensitive', and others did not satisfy either. The patterns are examined as to why they could not satisfy the conditions in the section.

Fluency Measures (Hesitation Phenomena)

The fluency measures concerned with hesitation phenomena met neither of the conditions, except for the number of reformulations which discriminated between some levels. Figure 3 demonstrates the patterns below.

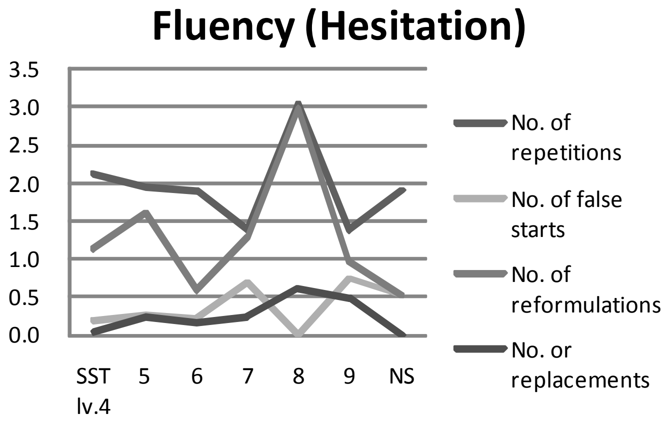


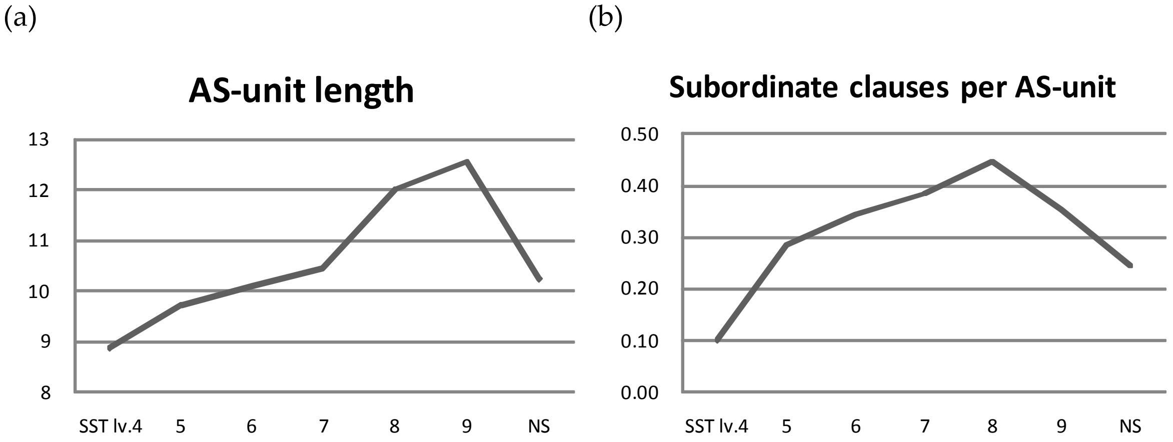
Figure 3 Patterns of Fluency Measures (Hesitation Phenomena)

The second line from the top in the graph shows the number of reformulation. The line peaked at SST level 8, and the LSD test showed that there were statistically significant differences between level 8 and others: 4-8, 5-8, 6-8, 7-8, 8-9, and 8-NS. The largest amount of hesitation at this level might be related to the low accuracy as presented in the previous section.

Reformulations did not correlate with the levels because there was not a linear pattern of increase or decrease as the levels went up. Similarly, non-linear patterns were observed with the other three measures. What is more, the NS apparently produced more repetitions than Japanese learners at SST levels 7 and 9. Taking these observations into account, the fluency measures of hesitation phenomena do not appear to be very credible in representing the proficiency levels.

Syntactic complexity

As for syntactic complexity, AS-unit length correlated significantly (*r*=.464) and the subordinate clauses per AS-unit discriminated between some non-adjacent levels (i.e. 4-6, 4-7, 4-8). Figures 4(a) and 4(b) present the patterns.



Figures 4(a), 4(b) Patterns of Syntactic Complexity Measures

AS-unit length displayed a steady increase among Japanese learners, but not with the NS. Subordinate clauses per AS-unit showed a very similar pattern, except that it starts to decline at SST level 9.

It is quite interesting that the NS performed lower than the candidates of higher proficiency (i.e. SST levels 7-9) according to this measure. One possible explanation for this is the differences in the conditions that the task was given. Compared to the SST candidates who were under pressure to prove their language proficiency within limited time, the NS performed the task with no limits in planning time or time for presentation. This suggests that the conditions of task administration should be controlled for all candidates in future research.

Alternatively, the less complex performance by the NS could be attributed to the task requiring narration, which might not encourage individuals to use complex language. Given this possibility, a review of previous literature may be warranted. Seeking to identify what makes a good narrative, other than the features (i.e. idea units) used in this study clearly deserves attention in future research.

Another explanation for this phenomenon might be that, contrary to our intuitive expectations, NS do not usually produce syntactically more complex speech than high level candidates whether they are given the task in the same situation or not. NS may be more prone to being 'economical' with language, with little intention to produce complex speech in most situations. This issue deserves further investigation, as it would be a significant finding for the fields of language testing and task-based research.

Lexical complexity

Although Jarvis (2002) justifies using D value as the best lexical complexity measure, it did not satisfy either of the conditions for being 'sensitive', and showed no consistent pattern (see Figure 5).

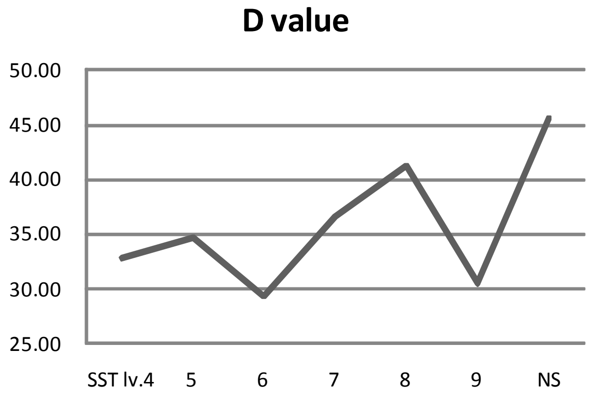


Figure 5 Pattern of D value

Two reasons might explain this result. Firstly, D value is meant for measuring lexical variety. Thus, it may not be suitable for applying to spoken performance on a narrative task because the content is largely pre-determined and the vocabulary range cannot be expected to vary as much as with tasks with more freedom to produce a wider variety of content. Secondly, the time limit of the interview could have influenced some SST candidates. Since the narrative task is given at the last stage of an SST, there can be different degrees of urging by the interviewer depending on how much time is left. If, for example, the SST level 6 and 9 candidates (who scored low on D value) had to finish telling the story quickly, then they might not have been able to demonstrate fully the vocabulary range that they possessed. The NS, who told a story to the author with unlimited time, might have been able to demonstrate more fully their vocabulary range. This issue needs to be examined with new sets of data, obtained under the same conditions and with no time limit for narration.

The frequency or 'difficulty'-based measures of lexical complexity presented rather flat patterns across the levels as shown in Figures 6(a) to (d).

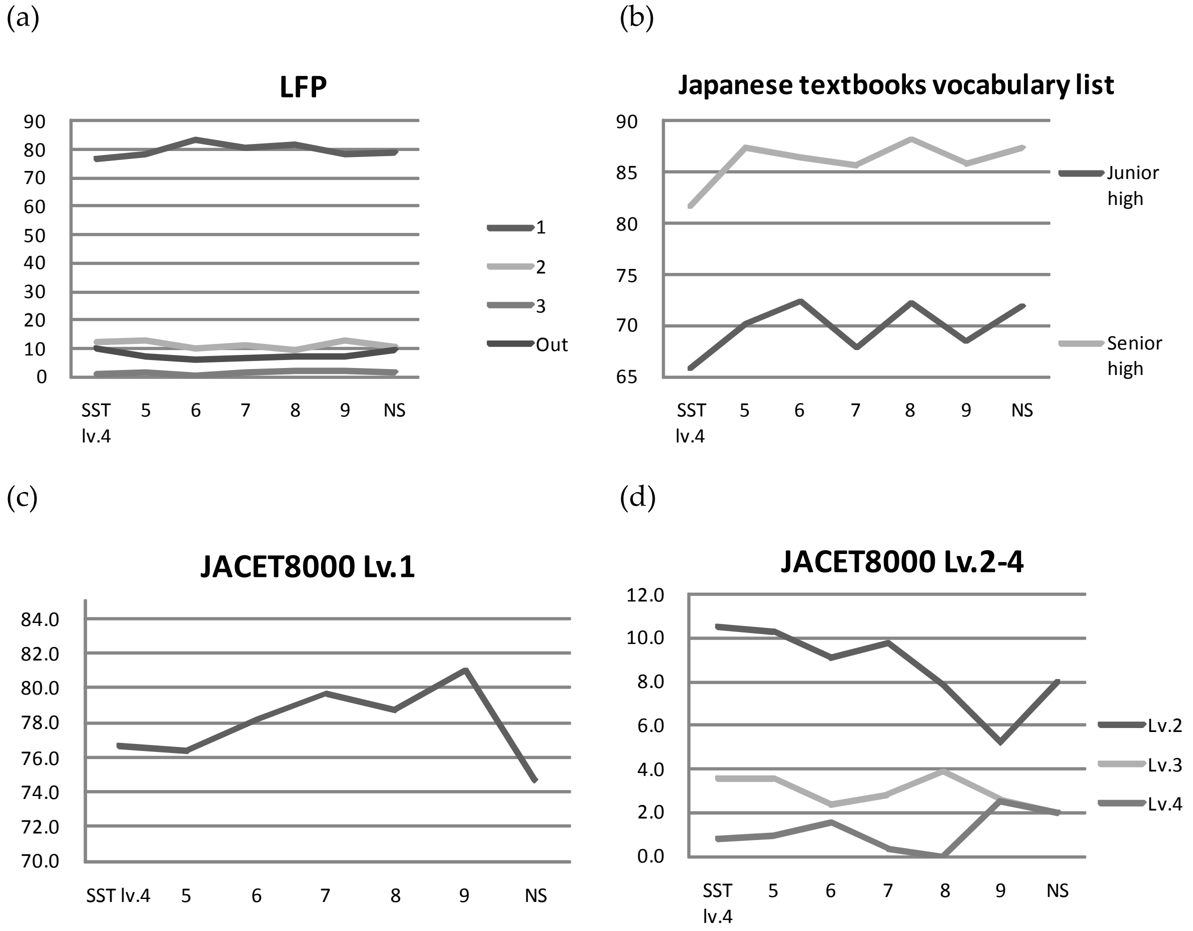


Figure 6(a), 6(b), 6(c), 6(d) Patterns of Lexical Complexity (Vocabulary Lists)

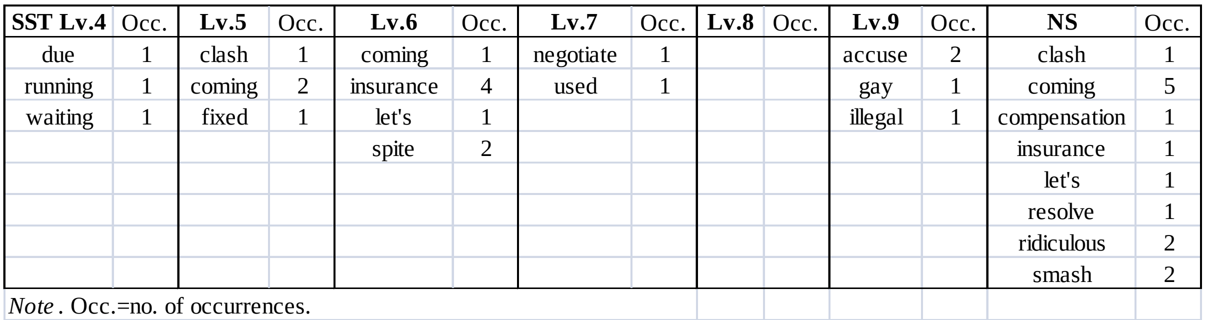
The figures suggest that the Japanese learners and the NS used more or less similar levels of vocabulary according to LFP and Japanese English textbooks vocabulary for junior and senior high schools. This is in line with the discussion made earlier on D value; since the content is pre-specified, the vocabulary range is decided by the task to some extent, thus leading to the use of similar vocabulary across the different levels.

However, JACET8000 drew somewhat different patterns. Its Lv.2 list had a moderate, negative significant correlation, and its Lv.4 list discriminated between some levels. In order to find out why these phenomena were related to a particular level of vocabulary, it was decided to examine the lists of actual words observed with their frequency.

By scrutinising JACET8000 lists, it was shown that lower level SST candidates used the word *policeman* more frequently than higher level candidates who more often used the term *police* or *police officer*. This is an unexpected result because *policeman* is classified as Lv.2, and therefore regarded of lower frequency whereas *police* and *officer* are classified as Lv.1.

Therefore, according to JACET8000 lists, lower level candidates succeeded in using 'less frequent' words, where higher level candidates used 'more frequent' words, leading to a negative correlation. It calls for caution that results can be hugely influenced by such slight differences in the words used.

In regard to the JACET8000 Lv.4 list, all the words in the narrative performance that belong to the Lv.4 list were identified at each level (i.e. SST and NS). Table 6 shows the words and the number of their occurrences in the transcripts at each level.



SST Lv.4	Occ.	**Lv.5**	Occ.	**Lv.6**	Occ.	**Lv.7**	Occ.	**Lv.8**	Occ.	**Lv.9**	Occ.	**NS**	Occ.
due	1	clash	1	coming	1	negotiate	1			accuse	2	clash	1
running	1	coming	2	insurance	4	used	1			gay	1	coming	5
waiting	1	fixed	1	let's	1					illegal	1	compensation	1
				spite	2							insurance	1
												let's	1
												resolve	1
												ridiculous	2
												smash	2

Note . Occ.=no. of occurrences.

Table 6 JACET8000 Lv.4 Words Used at Each Level

Table 5 above presented earlier indicates that the JACET8000 Lv.4 list discriminated between SST levels 4-9, 5-9, 6-8, 7-9, 7-NS, and 8-NS. The numbers of occurrences appear different between SST level 7-NS, 6-8, and 8-NS. There are hardly any differences at SST levels 4, 5, and 9. However, as the numbers of transcripts differed (i.e. SST lv.4=4; lv.5=5; lv.9=2), the resultant percentage of JACET8000 Lv.4 words was larger at SST level 9.

This, again, raises questions about using such word lists to identify which levels of words the speakers were able to produce during narration. In addition to the discussion on the pre-determined vocabulary range by the task, there is an issue of selective use of words by the learners. SST level 8 candidates in this study did not use any JACET8000 Lv.4 words, but it does not necessarily imply that they did not have any lexical knowledge of them. The same applies to SST level 7 candidates who did not use many words at JACET8000 Lv.4. In sum, rather than expecting to find meaningful differences in lexical use among different proficiency levels with these measures, it would be more sensible to analyse the narrative performance qualitatively. For example, we might explore if there are any differences in the expressions about the same characters, items, or events in the story at different levels.

Narrative structure

The last measure that is discussed here is narrative structure: the numbers of main and minor idea units. As Figure 7 below plots, most of the SST candidates covered more than 4 main idea units out of 6 (by NS performance), which means that even lower level learners could convey the essential events of the story to an extent. The minor idea units showed more variation.

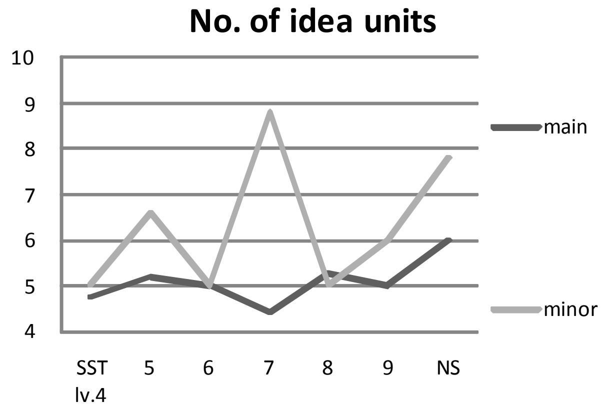


Figure 7 Patterns of the idea units

The numbers of idea units closely relate to how much they talked. Judging from the means of the number of words in Table 4, SST level 8 candidates talked less than level 7, which explains why the level 8 candidates produced less numbers of minor idea units. As the time available for narration may influence the number of words produced, the patterns of this measure emphasise again the importance of task administration to produce the same conditions for every speaker.

Conclusions, Limitations and Suggestions for Future Research

To sum up, this study has found that the only 'sensitive' measure with high correlation with the proficiency levels and high discrimination among many levels was the speech rate. Among the rest of the measures, some either correlated highly with the SST levels but could not discriminate, or could discriminate to some extent but did not correlate highly with the levels. Others satisfied neither of the conditions and were not 'sensitive.'

The major limitation of this study lies in that it used SST levels as a reference measure for the quality of Japanese learners' performance on the narrative task. As it was explained in Section 3.1.1., an SST level is an overall rating for the entire interview with three different types of tasks. The narrative task is only one of them. Although the SST raters decide the provisional level for the performance on each task type, they are averaged out and not revealed with the final SST level. So, it is possible for a candidate to do well (or poorly) on the narrative task but poorly (or well) on the other tasks, and his final SST level does not reflect the quality of performance specifically on the narrative task.

What is more, SST raters use analytic scales (explained briefly in Section 3.1.1.) for different aspects of the performance on each task, but these 'sub-levels' are not revealed either. Therefore, SST levels cannot provide information on how the candidate is profiled in different aspects of their performance. By employing the SST levels as the reference measure for correlation, this study implicitly presupposed that there was a linear increase in complexity or decrease in errors as the levels go up, which is not always the case in second language research (Fulcher, 2003: 103). Thus, it is highly desirable to rate the narrative individually and then to use the ratings, rather than the SST levels which are decided after considering performance on other tasks in the interview.

In addition to having the ratings solely based on the narrative performance, three suggestions should be made for future research. Firstly, the task should be given under the same conditions for every speaker. It may be especially important to allow speakers to talk as much as possible with no test-like pressure or no time limit for narration, so that the measures for syntactic complexity and idea units can be fully explored without the possible interference of pressure and time. The second suggestion is to run more qualitative analyses, especially for lexical complexity, rather than relying on the word lists for meaningful information about the differences in how the story is expressed. Lastly, the design of this study needs to be replicated with a larger sample size in order to verify if the patterns observed in this study can be generalised. Measures for accuracy and syntactic complexity may benefit the most from this suggestion.

Although this study has its limitations, its contribution is very unique in that it systematically and empirically examined various measures for their sensitivity. In the future, a similar study using a larger sample size, ratings based solely on the narrative performance, as well as qualitative analyses, and stricter control in task administration will be able to build on the conclusions drawn from this study, and is likely to have important implications for the field of language testing and task-based research.

Acknowledgement

I would like to thank Mr Hirano at the ALC Press for granting me permission to have access to their recordings data. The time his team spent on retrieving the 24 SST recordings used in this study is greatly appreciated.

References

ACTFL-ALC Press (2000). *The SST Manual*. Tokyo: Author.

ALC Press (2007). Heisei 20 nendo dai 5 ki Keiei Keikaku [The Management Plan for the 5th Term in Fiscal Year 2008] Retrieved January 26, 2008, from http://alc.irbridge.com/ja/custom1/custPage/04/custDownloadFile1/dai4ki%20kessan%20setsumeikai%20shiryou_070720.pdf

Appel, G. (1984). Improving second language production. In H. W. Dechert, D. Möhle & M. Raupach (Eds.), *Second language productions* (pp. 186-210). Tübingen: Gunter Narr.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 23-48). Essex, UK: Pearson Education.

Crookes, G. (1989). Planning and interlanguage variability. *Studies in Second Language Acquisition, 11*, 367-383.

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics, 24*, 197-222.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*, 299-323.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*, 354-375.

Fulcher, G. (2003). *Testing second language speaking*. Essex, UK: Pearson Education.

Izumi, E., Uchimoto, K., & Isahara, H. (2004). *Nihonjin 1200nin no eigo speaking corpus [A spoken English corpus by 1200 Japanese learners]*. Tokyo: ALC Press.

JACET (2003). *JACET8000*. Tokyo: JACET English Vocabulary SIG.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19*(1), 57-84.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*, 145-164.

Labov, W. (1972). *Language in the inner city: Studies in the Black English vernacular*. Oxford: Basil Blackwell.

Laufer, B., & Nation, P. (1995). Vocabulary Size and Use - Lexical Richness in L2 Written Production. *Applied Linguistics, 16*(3), 307-322.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*, 387-417.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition, 20*, 83-108.

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19*(4), 395-418.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition, 20*, 109-148.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning, 45*(1), 99-140.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 23*, 27-57.

Skehan, P. (1996). A framework for the implementation of task based instruction. *Applied Linguistics, 17*, 38-62.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49*, 93-120.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*, 167-197.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity, and accuracy in L2 monologic oral production. *Applied Linguistics, 24*, 1-27.