The Language Learning Journal, 2016
http://dx.doi.org/10.1080/09571736.2015.1130079

Routledge
Taylor & Francis Group

A comparative study of the variables used to measure syntactic complexity and accuracy in task-based research

Chihiro Inoue*

Centre for Research in English Language Learning and Assessment, University of Bedfordshire, Luton, UK

The constructs of complexity, accuracy and fluency (CAF) have been used extensively to investigate learner performance on second language tasks. However, a serious concern is that the variables used to measure these constructs are sometimes used conventionally without any empirical justification. It is crucial for researchers to understand how results might be different depending on which measurements are used, and accordingly, choose the most appropriate variables for their research aims. The first strand of this article examines the variables conventionally used to measure syntactic complexity in order to identify which may be the best indicators of different proficiency levels, following suggestions by Norris and Ortega. The second strand compares the three variables used to measure accuracy in order to identify which one is most valid. The data analysed were spoken performances by 64 Japanese EFL students on two picture-based narrative tasks, which were rated at Common European Framework of Reference for Languages (CEFR) A2 to B2 according to Rasch-adjusted ratings by seven human judges. The tasks performed were very similar, but had different degrees of what Loschky and Bley-Vroman term 'task-essentialness' for subordinate clauses. It was found that the variables used to measure syntactic complexity yielded results that were not consistent with suggestions by Norris and Ortega. The variable found to be the most valid for measuring accuracy was errors per 100 words. Analysis of transcripts revealed that results were strongly influenced by the differing degrees of task-essentialness for subordination between the two tasks, as well as the spread of errors across different units of analysis. This implies that the characteristics of test tasks need to be carefully scrutinised, followed by careful piloting, in order to ensure greater validity and reliability in task-based research.

Keywords: speaking; task-based research; syntactic complexity; accuracy

Introduction

Since the 1970s, researchers have been seeking reliable indices to measure learners' performance in a second language (L2). The variables of complexity, accuracy and fluency (CAF) have been extensively used, as researchers appear to agree that the triad componential CAF framework best captures the characteristics of different aspects of learner performance (Housen and Kuiken 2009). For example, in task-based research, CAF variables have been used to measure gains in L2 learning, to examine the effects of particular types of feedback in language classrooms or the effects of different task administration conditions, or to identify the characteristics which distinguish among different proficiency levels. A serious

*Email: cherylmarue@gmail.com

concern with using CAF in research is that some of the measurement variables are used conventionally and without empirical justification, and it is this research gap that this article aims to fill. It is crucial for researchers to understand how results may be different if different measures are used, and thus to choose their test tasks and variables adequately to suit the purpose of their study and the characteristics of their data.

In the research on CAF in L2 speaking, the variables used to measure fluency are probably the most well-researched because of their immediately noticeable nature. A number of validity studies on fluency variables have been conducted (e.g. de Jong et al. 2013; Kormos and Dénes 2004), and there seems to be a general consensus that variables such as speech rate and phonation time ratio are good measures of this aspect of learner performance (at least on monologic data). In addition, a number of empirical studies have been published on variables relating to lexical complexity (e.g. Jarvis 2002; McCarthy and Jarvis 2010), thanks to rapid recent developments in digitalised vocabulary lists and tools for quantitative analysis. However, variables used to measure syntactic complexity and accuracy seem to be relatively under-researched, especially in terms of their validity and suitability for use in capturing differing characteristics of L2 development. Accordingly, the research reported in this article focuses on these variables.

Syntactic complexity

Complexity is ‘the extent to which learners produce elaborated language’ (Ellis and Barkhuizen 2005: 139), and syntactic complexity measures the use of structures that are considered more challenging or sophisticated. While there is research evidence that specific measurements of syntactic complexity, such as the raw frequency of target grammatical items, can be more accurate in distinguishing different proficiency levels or gauging the outcome of certain teaching methods (Norris and Pfeiffer 2003), general variables of syntactic complexity are useful because they allow comparisons among different studies (Tonkyn 2013). General variables used to measure syntactic complexity in previous task-based studies have typically been length-based or subordination-based. Examples of length-based variables include the number of words per unit (Bygate 2001 (T-unit); Mehnert 1998 (C-unit); Ortega 1999 (pausally defined unit)) and the number of clauses per chosen unit (Foster and Skehan 1996; Iwashita et al. 2001; Robinson 2001, 2007; Skehan and Foster 1999). Subordination-based variables include the percentage of subordinate clauses in the total number of clauses (Wigglesworth 1997) and the number of subordinate clauses per T-unit (Crookes 1989; Mehnert 1998).

In a recent article offering a comprehensive review of the variables used to measure syntactic complexity, Norris and Ortega (2009) suggest that such traditional general variables may be too crude to capture the multi-dimensional nature of L2 learners’ development. In order to overcome such shortcomings, they recommend using within a single study: (a) length-based variables (such as words per chosen unit) as an overall measure of syntactic complexity, (b) subordination-based variables, (c) variables of phrasal complexity (i.e. clause length) and (d) coordination-based variables (i.e. amount of coordination) as opposed to focussing just on subordination, especially in studies involving learners of lower proficiency. This article will investigate the four types of variables mentioned above in relation to spoken data, as well as the suggestions made by Norris and Ortega that (d) may be most indicative of the beginner level, while (b) reflects intermediate proficiency and (c) advanced.

Accuracy

Accuracy refers to ‘the extent to which the target language is produced according to its rule system’ (Skehan 1996: 23). Examples of general variables for measuring accuracy include AQ2 the percentage of error-free clauses (Foster and Skehan 1996; Skehan and Foster 1999; Yuan and Ellis 2003), the percentage of error-free units (Robinson 2001, 2007 (C-unit)), the number of errors per unit (Bygate 2001 (T-unit)) and the number of errors per 100 words (Mehnert 1998). Interestingly, researchers seem to disagree as to which variables are thought to be most valid. Bygate (2001) suggests that calculating the number of errors per chosen ‘unit’ might be a more sensitive measure of accuracy because it does not obscure the actual occurrences of errors as counting error-free units does. On the other hand, Mehnert (1998) argues that, for relatively lower proficiency speakers, counting errors per 100 words may be more suitable since it does not involve definitions of clauses and units which can be problematic. Identifying which type of variables may be ‘more sensitive’ or ‘suitable’ requires validation (e.g. Kormos and Dénes 2004) – that is, comparing the results from different variables against human judgements of how accurate the performances in question are. Since there exists no previous research validating the variables proposed for measuring accuracy, it is crucial that research is undertaken in this area.

Research questions

Strand 1: syntactic complexity

Strand 1 of this study focused on two research questions:

- RQ1-1. How do the different variables measuring syntactic complexity correlate with one another?
- RQ1-2. How well do the different variables measuring syntactic complexity predict beginner, intermediate and advanced levels, respectively, of L2 speaking proficiency?

As suggested above, four different types of variables measuring syntactic complexity (i.e. length-based variables, variables for coordination, subordination and phrasal complexity) are examined in this article. The two research questions will help us to investigate whether the elicited performances are multi-dimensional in terms of syntactic complexity, and also whether, following Norris and Ortega’s suggestions (2009), different variables discriminate L2 spoken performance better at different proficiency levels.

Strand 2: accuracy

For Strand 2, there are two research questions. Both seek to identify which variable measuring accuracy is most valid, that is, most in line with human ratings of accuracy:

- RQ2-1. How do the variables measuring accuracy correlate with human ratings of accuracy of spoken narrative performances?
- RQ2-2. Do the t-tests on the measures of accuracy between the two tasks reveal the same results as the human ratings? If there are discrepancies, why?

The study

Participants

The participants who took part in this study were 64 students majoring in modern languages at a university in Japan, aged 20.8 years on average (SD = 3.9). All the participants were

Japanese native speakers who were studying English as a first foreign language. Their level of general English proficiency was assessed by a paper-based multiple-choice format Oxford Quick Placement Test (University of Cambridge Local Examinations Syndicate [UCLES] 2001) and ranged from B1 to C1 (B1: n = 15; B2: n = 31; C1: n = 19) on the CEFR.

Each participant performed the two tasks in a one-to-one interview with the author. After signing a consent form, each participant was shown a practice story-narration task which was selected from Hill (1960) and asked to narrate the story with as much detail as possible. Two minutes' planning time was given, just as for the two main tasks (Tasks A and B; see next section for details). The order of presentation of Tasks A and B afterwards was reversed for each participant so as to minimise any order effect. The elicited performances were recorded and later rated and transcribed for analysis.

Tasks

Bearing in mind that the elicitation tasks in this study must allow meaningful comparisons to be made between the tests used to measure syntactic complexity and accuracy, two picture-based spoken narrative tasks (Tasks A and B) were carefully selected with slight modifications from Hill (1960) (see Appendix 1). Quantitative and qualitative investigations into these two tasks are described and discussed in detail in Inoue (2013), and the known characteristics of these two tasks are that they are:

- very similar in storylines and characters involved (i.e. two children playing a trick on their mother in a house), so as to minimise the effects of different degrees of lexical complexity coming into play;[1]
- different in 'task-essentialness' (Loschky and Bley-Vroman 1993) for subordinate clauses (i.e. Task B elicits more subordination due to the constant presence of the mother and the plot of the baby being replaced by a ball), so that the performances on the two tasks will have more varied profiles from different syntactic complexity variables;
- slightly different in task complexity, with Task A being cognitively more complex than Task B[2] (due to the need to explain how the ghost-like figure is constructed and the time gap between Pictures 5 and 6); such that the elicited performances should exhibit different degrees of accuracy.

Rating

Seven raters were trained through the procedures described in Council of Europe (2009), including standardisation using the illustrative performances published by the Centre International d'Études Pédagogiques[3] and benchmarking using the sample performances of the Japanese participants. The native languages of the seven raters included English, Japanese, Arabic and Polish. All raters, except for one English native speaker, had at least two years of experience in TEFL, and four raters had previously been trained as examiners or raters for some tests of spoken English. They gave ratings from below A1 to C1, based on the CEFR Oral Assessment Grid (Council of Europe 2009; see Appendix 2) in the categories of Range, Fluency, Accuracy, Coherence and Sustained Monologue.[4] The raters then decided on a single overall level (i.e. Considered Judgement (CJ)) on the performances on both tasks by the 64 Japanese participants.

The ratings given to each spoken narrative performance were numerically transformed (i.e. 'Below A1' = 1 through to 'C1' = 10) and adjusted using multi-faceted Rasch analysis software, FACETS, which calculates ‘fair average’ ratings taking the seven raters’ differing degrees of harshness or leniency into consideration. All the raters exhibited acceptable infit values between 0.7 and 1.3 (Bond and Fox 2007), which indicated that all of them interpreted and applied the rating scales in a consistent manner. For Strand 1 (syntactic complexity), different groups of proficiency levels were assigned based on the ‘fair average’ CJ ratings, following the procedures by Eckes (2009). These were rounded up or down to give each performance a CEFR level. For Strand 2 (accuracy), the ‘fair average’ accuracy ratings were used for correlation and t-tests.

Variables

All the spoken narrative performances were transcribed and coded for the variables measuring syntactic complexity and accuracy. For the units of analysis, it was decided to employ the AS-unit. An AS unit is defined as ‘a single speaker’s utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either’ (Foster et al. 2000: 365). Unlike other common units such as the T-unit and C-unit, it can take sub-clausal units, into nation and pausing into account, and thus is considered more suitable for spoken data; it offers a consistent classification of false starts, repetitions and self-corrections (Ellis and Barkhuizen 2005). Syntactic complexity was measured by four variables: AS-unit length (words per AS-unit), coordination per AS-unit,[5] subordinate clauses per AS-unit and clausal length (words per clause). For accuracy, errors per AS-unit (cf. Bygate 2001) and errors per 100 words (Mehnert 1998) were examined. Additionally, the percentage of error-free clauses was selected on the grounds that this measurement is considered suitable for an experimental design (Skehan and Foster 1999) and has been used in relevant previous studies (e.g. Foster and Skehan 1996; Skehan and Foster 1999; Yuan and Ellis 2003).

All the errors, AS-units, clauses and subordinate clauses were manually identified by the author. The reliability of coding was ensured by double-coding ten transcripts for each task, and the agreement for all the errors and syntactic units reached 90%. To ensure accuracy when counting errors, the errors in ten transcripts were identified by an English native speaker, who had experience as a TEFL teacher, as well as by the author. The errors were identified based on the broad definition adopted by Skehan and Foster (1997: 195), who regarded language use which is ‘nonexistent in English or indisputably inappropriate’ as errors. The inter-coder reliability reached 95% and was thus considered satisfactory. Subsequently, errors were identified by the author alone.

AQ3

Methods of analysis

For RQ1-1 and RQ2-1, Spearman’s r correlation coefficients were used due to the non-normality of the data (i.e. resultant values from the variables under investigation). For RQ1-2, discriminant analysis was employed, following Oh’s (2006) method of analysis. Discriminant analysis is mathematically equivalent to one-way MANOVA, but shifts the focus to present an estimate of the degree to which variables function as predictors for group membership. Here, the predictor variables were the syntactic complexity variables, and the grouping variable was the Rasch-adjusted CEFR levels. For RQ2-2, related sample t-tests were used because the ‘fair average’ accuracy ratings and the values from different accuracy variables were assumed to have ‘a reasonably normal distribution’ (Bachman 2004: 74) with skewness and kurtosis in the range of −2 to +2.

A two-way mixed design ANOVA showed no significant interactions between task order and tasks, and hence, there was no order effect, except for the number of errors per AS-unit (F(1, 32) = 4.916, p = .030). It was decided to include the number of errors per AS-unit in a further analysis, since the actual order effect was considered small, as indicated by the magnitude of power calculated by PASW 17.0, which was .588.

Results and discussion

Strand 1: syntactic complexity

Tables 1 and 2 below, respectively, summarise the descriptive statistics relating to the four variables measuring syntactic complexity and the Spearman's r correlation coefficients between them.

Table 2 shows that on both tasks, AS-unit length, an overall measure of syntactic complexity (Norris and Ortega 2009), correlated moderately highly with subordinate clauses per AS-unit (.58 for both tasks) and clause length (.67 and .69 for Task A and B, respectively). Considering that the definition of AS-unit entails clauses (and subordinate clauses), the longer the clauses (or subordinate clauses) are, the longer AS-units tend to become; therefore, these results are not surprising. Coordination index and AS-unit length, however, showed almost no correlation (.21 and −.12 for Task A and B, respectively). This may be because using coordination requires only a couple of words at the beginning of a spoken phrase, such as 'and', 'but' and 'so', which do not greatly contribute to the length of AS-units.

The correlations between the remaining three variables (i.e. coordination per AS-unit, subordinate clauses per AS-unit and clause length) are either almost none or weak (ranging from −.17 to .30). Therefore, there is not much overlap among these three

Table 1. Descriptive statistics for syntactic complexity variables.

Task	Variable	N	Mean	SD
A	AS-unit length	65	8.67	1.42
	Coord. per AS-unit	65	0.71	0.17
	Subord. per AS-unit	65	0.17	0.14
	Clause length	65	7.37	0.95
B	AS-unit length	65	8.69	1.50
	Coord. per AS-unit	65	0.63	0.18
	Subord. per AS-unit	65	0.24	0.15
	Clause length	65	6.99	0.95

Table 2. Results of correlations using Spearman's rho (RQ1-1).

Task	Variable	AS-unit length	Coord. per AS-unit	Subord. per AS-unit
A	Coord. per AS-unit	0.21		
	Subord. per AS-unit	0.58**	−0.17	
	Clause length	0.67**	0.30*	−0.03
B	Coord. per AS-unit	−0.12		
	Subord. per AS-unit	0.58**	−0.12	
	Clause length	0.69**	0.28*	−0.08

*p < .05.
**p < .01.

variables, which demonstrates that they may be representative of different dimensions of syntactic complexity. Having confirmed that there were only negligible or weak correlations among the three variables, the analysis for RQ1-2 was run. This examined how well the variables predicted different proficiency levels. Based on the multi-faceted Rasch analysis, the 64 participants were classified as 'A2 (i.e. beginner)', 'B1 (i.e. intermediate)' or 'B2 or above (i.e. advanced)'. Tables 3 and 4 show the results of discriminant AQ4 analysis on the amount of coordination on both tasks for these three level groups.

The figures in bold show the percentages of participants where there was an exact match between the CEFR levels assigned to them (i.e. 'Actual Level' in the tables) and the level predicted by the discriminant analysis (i.e. 'Predicted Level' in the tables). Norris and Ortega (2009) suggested that, for L2 writing, the amount of coordination may be the best indicator of accuracy at beginner level. However, for the spoken narrative performances used in this study, this does not seem to hold true; the exact matches of actual-A2 and predicted-A2 account for only 20% and 8% of the participants on Tasks A and B, respectively, and there are a lot more participants 'misplaced' in the wrong levels. Interestingly, exact matches at the levels of B1 and B2 or above outnumber misplacement on both tasks (B1: 58.1% (A) and 50.0% (B); B2 and above: 58.1% (A) and 62.5% (B)), although around 60% exact matches may not be sufficiently high to be considered as good indicators of certain levels.

For subordinate clauses per AS-unit, Tables 5 and 6 summarise the results of discriminant analysis. Norris and Ortega (2009) suggested that the amount of subordination should be a good syntactic complexity indicator of intermediate level L2 writing proficiency, but with the spoken data in this study, the exact matches of actual-B1 and predicted-B1 accounted only for 9.7% and 16.1% of the participants for Tasks A and B, respectively. What is intriguing is the high percentage (40%) of misplacement of actual-A2 participants in predicted-B2 or above category, and this may be due to the higher task-essentialness of

Table 3. Results of discriminant analysis (coordination on task A).

		Predicted level							
		A2		B1		B2 or above		Total	
Task A		%	n	%	n	%	n	%	n
Actual level	A2	20.0	(5)	40.0	(10)	40.0	(10)	100.0	(25)
	B1	12.9	(4)	58.1	(18)	29.0	(9)	100.0	(31)
	B2 or above	0.0	(0)	50.0	(4)	50.0	(4)	100.0	(8)

Note: 42.2% of original grouped cases correctly classified.

Table 4. Results of discriminant analysis (Coordination on Task B).

		Predicted level							
		A2		B1		B2 or above		Total	
Task B		%	n	%	n	%	n	%	n
Actual level	A2	8.0	(2)	60.0	(15)	32.0	(8)	100.0	(25)
	B1	3.2	(1)	58.1	(18)	38.7	(12)	100.0	(31)
	B2 or above	12.5	(1)	25.0	(2)	62.5	(5)	100.0	(8)

Note: 39.1% of original grouped cases correctly classified.

Table 5. Results of discriminant analysis (Subordination on Task A).

		Predicted level							
		A2		B1		B2 or above		Total	
Task A		%	n	%	n	%	n	%	n
Actual level	A2	68.0	(17)	12.0	(3)	20.0	(5)	100.0	(25)
	B1	48.4	(15)	9.7	(3)	41.9	(13)	100.0	(31)
	B2 or above	25.0	(2)	37.5	(3)	37.5	(3)	100.0	(8)

Note: 35.9% of original grouped cases correctly classified.

Table 6. Results of discriminant analysis (subordination on task B).

		Predicted level							
		A2		B1		B2 or above		Total	
Task A		%	n	%	n	%	n	%	n
Actual level	A2	56.0	(14)	4.0	(1)	40.0	(10)	100.0	(25)
	B1	45.2	(14)	16.1	(5)	38.7	(12)	100.0	(31)
	B2 or above	37.5	(3)	12.5	(1)	50.0	(4)	100.0	(8)

Note: 35.9% of original grouped cases correctly classified.

subordinate clauses on Task B than on Task A. To demonstrate this, the transcripts of an actual-A2 participant from both tasks are discussed below. Forward slashes indicate AS-unit boundaries and words in brackets indicate where subordination occurs.

> Task A Performed by Participant 36 (Total of 1 subordinate clause)
>
> a woman is washing clothes / a cat gaze her / she dries big coat / [after] she going to home a man comes to her house / he sells balloons / her children go home / and they buy a balloon / a girl painted a human face on the balloon / and a boy open coat drying / after a while she shocked to something strange looking at window / a big big something strange / she very shocked / and cat also very shock / and run away /
>
> Task B Performed by Participant 36 (Total of 4 subordinate clauses)
>
> a mother was reading book / nearby basket her baby was sleeping / she read book / but [after] finish reading she slept too / then her children a boy and girl come to room / they want to treat / first a girl try to hide baby / next a boy put strange ball in the basket [which] baby slept / [after] their mother woke up she very shocked [because] in the basket there is no baby but very strange ball /

As can be seen, Task B elicited more subordination from an actual-A2 participant, and this tendency was found across participants at different levels of proficiency. Examination of the transcripts suggests that the higher levels of subordination on Task B are partly due to the mother character's constant presence in the pictures, which requires occasional mentioning of the mother's state using while and after. It is also attributable to the plot, in which the baby is replaced by a ball in the basket, which might have necessitated the use of relative pronouns (such as that, which) and adverbs (where).

It has been empirically demonstrated that Task A is cognitively more difficult than Task B (Inoue 2013); however, it was Task B that elicited more subordination from participants across different levels of proficiency. This suggests the strong possibility that subordination may be elicited regardless of how 'cognitively difficult' a task is thought to be, which in turn raises serious doubt about measuring syntactic complexity by means of the amount of subordination. This may threaten the fundamental assumption of the arguments about task complexity, that is, that the more cognitively complex tasks are (in terms of, for example, information organisation (Skehan 2009) or absence of pictures (Robinson 1995)), the more syntactically complex the elicited performance will be. In this regard, task complexity might not be related at all to any complexity in the linguistic performance. Researchers thus need to consider seriously the task-essentialness (Loschky and Bley-Vroman 1993) of subordinate clauses when deciding on the tasks to use for research; selected tasks need to be piloted carefully.

AQ5

Finally, turning to the third syntactic complexity variable, Tables 7 and 8 summarise the results of discriminant analysis for clause length. This variable represents phrasal complexity, and Norris and Ortega (2009) have argued that it is indicative of advanced level proficiency in L2 writing. Interestingly, on both tasks, this variable best predicts A2 level (64% and 68% on Tasks A and B, respectively). The means of clause length on both tasks for each of the levels are given in Table 9, and they show a steady increase as the levels go up. The standard deviations indicate slightly wider variations at higher proficiency levels; so it may be the reason why the shorter lengths of clauses at A2 level have been able to predict the level most accurately. It is also worth noting that half of the B2 or above group has been misplaced as A2, which suggests that higher proficiency learners may not always complexify their speech in ways which influence clause length, such as through pre- or post-modifications using adjectives, adverbs, prepositional phrases or non-finite clauses (Norris and Ortega 2009).

Strand 2: accuracy

For RQ2-1, ‘How do the variables of accuracy correlate with the human ratings of accuracy of the spoken narrative performances?’, the correlation coefficients between the three

Table 7. Results of discriminant analysis (Clause length on Task A).

		Predicted level							
		A2		B1		B2 or above		Total	
Task A		%	n	%	n	%	n	%	n
Actual level	A2	64.0	(16)	12.0	(3)	24.0	(6)	100.0	(25)
	B1	41.9	(13)	16.1	(5)	41.9	(13)	100.0	(31)
	B2 or above	25.0	(2)	25.0	(2)	50.0	(4)	100.0	(8)

Note: 39.1% of original grouped cases correctly classified.

Table 8. Results of discriminant analysis (Clause length on Task B).

		Predicted level							
		A2		B1		B2 or above		Total	
Task A		%	n	%	n	%	n	%	n
Actual level	A2	68.0	(17)	0.0	(0)	32.0	(8)	100.0	(25)
	B1	58.1	(18)	9.7	(3)	32.3	(10)	100.0	(31)
	B2 or above	50.0	(4)	0.0	(0)	50.0	(4)	100.0	(8)

Note: 37.5% of original grouped cases correctly classified.

Table 9. Descriptive statistics of clausal length.

Level	Task A	Task B
A2	7.08 (SD = .69)	6.83 (SD =.81)
B1	7.48 (SD = .99)	7.12 (SD = 1.03)
B2 or above	8.07 (SD = .1.06)	7.20 (SD = 1.00)

general accuracy variables and the Rasch-adjusted accuracy ratings are summarised in Table 10.

Although all three variables correlated moderately highly with the Rasch-adjusted ratings, errors per 100 words showed slightly stronger correlations on both tasks. This finding will be discussed together with the findings for RQ2-2 below. The issue of error gravity (Hughes and Lascaratou 1982) may explain why the correlations were only moderately high. The straightforward counting of errors does not distinguish between global errors (e.g. incomprehensible lexical choice) and local errors (e.g. article omission or subject-verb agreement which do not impede communication). The difference in the 'seriousness' of errors may distinguish lower level and higher level performers, but the variables for assessing accuracy investigated here may result in, for example, higher level performers with a lot of local, minor errors being ranked lower than performers making fewer, but more serious, global errors. Thus, it is assumed that coefficients in the range |.6| to |.7| might be as high as the correlations can go, if general accuracy variables are to be used; the remaining variance could be explained by the seriousness of each error.

Table 11 shows the descriptive statistics as well as the results of t-tests for RQ2-2, 'Do the t-tests on the measures of accuracy between the two tasks reveal the same results as the human ratings? If there are discrepancies, why?'. For the adjusted ratings of accuracy, a significant difference with a very small effect size (−.12) was found between the performances on Tasks A and B, demonstrating that Task B elicited slightly more accurate performances. For the accuracy variables, significant differences were found for the percentage of error-free clauses (t(64) = −2.81, p = .006) and the number of errors per 100 words (t(64) = −2.59, p = .01). Effect sizes were small on the percentage of error-free clauses (−.29), and a very small effect was found on the number of errors per 100 words (.14). It is striking that the three accuracy variables produced different results: a significant difference with a small effect on the percentage of error-free clauses, no significant difference on the number of errors per AS-unit, and a significant difference with a very small effect on the number of errors per 100 words.

Although the three accuracy variables all correlated highly with the ratings of accuracy, the correlation with the number of errors per 100 words measure was higher than for the other two measures (see Table 10), as well as revealing a significant difference with a very small effect (see Table 11), just as the ratings did. This demonstrates its better suitability to reflect the raters' judgements of accuracy. This supports Mehnert's (1998)

Table 10. Results of correlations between the ratings and variables of accuracy.

Variables	Pearson's Coefficients
(Task A, B)	
% of error-free clauses	.644**, .683**
Errors per AS-unit	−.652**, −.687**
Errors per 100 words	−.723**, −.731**

**p < .01.

Table 11. Results of related sample t-tests on the ratings and variables of accuracy

	M		SD		t	df	p	d
	A	B	A	B				
Adjusted ratings	2.47	2.55	.68	.63	−2.61	64	.02*	−.12
% of error-free clauses	52.37	58.66	21.07	21.64	−2.81	64	.01**	−.29
Errors per AS-unit	.71	.66	.35	.40	1.26	64	.21	
Errors per 100 words	8.56	7.89	4.71	4.91	−2.59	64	.01*	.14

*p < .05.
**p < .01.

claim that this variable may be appropriate for learners with relatively lower proficiency, though for a slightly different reason from the one she gave. Mehnert prefers this variable because having 100 words as a denominator avoids the use of clausal units whose definitions may be problematic. However, as the transcripts in the next section demonstrate, having clausal units as a denominator may, depending on how errors are spread across clauses, produce completely different results. As the spread of errors might also depend on the proficiency levels of participants, using clausal units as denominators could function well with a narrower range of proficiency levels.

Another researcher who has argued for using one particular accuracy variable over others is Bygate (2001). Bygate suggests that calculating the number of errors per unit (i.e. T-unit in his study) might be most appropriate because it does not obscure the actual occurrences of errors in the way that counting error-free units does. However, as the transcripts in the next section will show, segmenting transcripts into clausal units, such as AS-units, can also change the results. In this study, the errors per 100 words measure aligns best with the ratings, and thus can be considered the most valid measure of accuracy for the current data set consisting mostly of learners at A2 and B1 levels. Nevertheless, as mentioned above, having clausal units as denominators could be appropriate and in line with the ratings for a narrower range of proficiency levels, as a narrower ability range may show less variation in the length of clausal units, thus not altering the distribution of the resulting values too much.

As noted earlier, the variable of errors per AS-unit was found to be affected by task order, but was included in further analysis because the size of the order effect was thought to be small. This variable did not show any significant difference between Task A and Task B, whereas errors per 100 words showed a significant difference with a very small size effect (Cohen’s d = .14). A significant difference between Task A and Task B was also found in the percentage of error-free clauses, but with a small effect of −.29. In order to explore how such differences between the three variables might have been caused, we first discuss the errors per AS-unit and errors per 100 words, as they have the same numerator (i.e. number of errors) but different denominators. The ensuing discussion then compares errors per 100 words and the percentage of error-free clauses.

If two of the accuracy variables – errors per AS-unit and errors per 100 words – yielded different results, then the difference must have been due to the difference in the denominators. The task transcripts were examined for this trait, and the transcripts for Participant 64 on the two tasks are shown below in order to illustrate how such differences might occur. AS-units are indicated by forward slash, and errors are marked by underscore (omission), underlining (inflection/mischoice) and strikethroughs (unnecessary parts).

Task A Performed by Participant 64 (0.6 errors per AS-unit; 6.52 errors per 100 words)

there is a housewife washing her clothes and her family_ / and beside her there is a black cat / and then after washing them she starts hanging them outside / then she gets inside / after maybe one hour or two hours the clothes got dry/ and here comes a man with some balloons in his hand
/ and the children from the house come out / and get _ balloon from the man / and they get some bad not so bad but some funny idea / and a girl starts painting some face on the balloon / a boy _
get ~~out~~ the cloth from the rope / and they makes a balloon and a clothes like a man / and from the window they show~~ed~~ this figure / and the housewife is very surprised / and even the black cat is afraid of them / and then run away /

Task B Performed by Participant 64 (0.67 errors per AS-unit; 6.06 errors per 100 words)

there is a woman reading some book on the chair in her room / and beside her or in front of her there is a baby sleeping in ___ small basket / but during reading the mother fe ll asleep / and two children a boy and a girl come in the room / and they try to take the baby from the basket / and instead of the baby they bring a ball painted a face on it / and put it on the basket / and after few minutes the woman wakes up and see it / and surprised because the baby has changed to a ball /

As can be seen from the length of the transcripts, the performance on Task A was longer and contained more words (138 words) than that on Task B (99 words). The number of errors was also greater on Task A (9 errors) than B (6 errors). This produced a larger number of errors per 100 words for Task A (6.52) than B (6.06). However, because the AS-units were slightly longer in Task B, each AS-unit contained a higher number of errors, which resulted in a larger number of errors per AS-unit for Task B (0.66). Therefore, Task A was more erro- neous according to the variable of errors per 100 words, whilst the errors per AS-unit measure showed the opposite. Thus segmenting the transcripts into different clausal units of analysis can produce quite different results, which suggests these two accuracy variables need to be used with due care.

We turn now to the third accuracy variable, the percentage of error-free clauses, in com- parison with the number of errors per 100 words. These two accuracy variables showed sig- nificant differences for RQ2-1 (comparison between Task A and Task B), both indicating that performances on Task B were more accurate. However, the effect size, as measured by Cohen’s d, was different in each case; it was very small for errors per 100 words (.14) but small for percentage of error-free clauses (−.29). With a very small size effect, the significant difference between Task A and Task B for errors per 100 words can be ignored. In contrast, the difference between the two tasks appeared to have a slightly stron- ger effect on the percentage of error-free clauses. This raises the question as to how the two accuracy variables could reveal such different results, and which variable better reflects the actual differences (as judged by rating) in accuracy of the performances.

To answer the first question, the raw data and transcripts were revisited. It was found that the cause of the difference in size effect between the two accuracy variables is likely to have been, again, the difference in denominators (i.e. 100 words or clauses), as the denominators must have affected the extent to which each error mattered when calculating values. Even if two participants made about the same number of errors per 100 words, the percentage of error-free clauses might be quite different. Let us take two narrative perform- ances, by Participants 15 and 16 on Task A as examples. Clauses are indicated by forward slash, and errors are marked by underscore (omission), underlining (inflection/mischoice) and strikethroughs (unnecessary parts).

Participant 15 (5.60 errors per 100 words; 66.67% of error-free clauses)

in a room a mother is washing clothes on the table / and under the table a cat is looking at her / number two she is hanging washed clothes to dry in the yard / number three a man is walking along a fence / and he sells balloons / and the two children_running up / number four the children buy a balloon from him / number five a girl is drawing a face on the balloon / and a boy is standing on the box / and he is taking a clothes to do something / number six the first lady is surprised by the children's ____ / a mother and a cat is surprised by_monster-like man / he was made by the children / his face is the balloon / and his body is ___ washed clothes /

Participant 16 (5.42 errors per 100 words; 72.73% of error-free clauses)

there was a lady / who was washing her clothes in the kitchen / and in the kitchen there was a black cat too / after washing the clothes / she went out of the house / and she put them to the rop~~es~~ / and she went back to the room / and then the man came ~~to~~ near the house / and in his hand he had balloons to sell / and now the boy and the girl bought a balloon from the man / and then an idea came up to them / the girl painted a face / it looks like a Humpty Dumpty on the face of the balloon / and the boy picked up a shirt / or maybe it's a ~~sheets~~ from the rope / and then they acted like a big man / and they stood in front of the window from the outside / and then because the lady was in the kitchen / and looking ___ the big man from the window / she was so astonished / she was surprised / and the black cat run away /

Participants 15 and 16 had about the same number of errors per 100 words (5.60 and 5.42, respectively), but the percentages of error-free clauses were different (66.67% and 72.73%, respectively) depending on how the errors were spread across clauses. Whilst most of Participant 15's errors were spread out as one error per clause, those of Participant 16 were clustered together to produce a count of two errors per clause, which resulted in a higher ratio of error-free clauses. Therefore, with the number of errors per 100 words, each error is taken into account when calculating a value, while the percentage of error-free clauses can have clustered errors in certain clauses, leaving others error-free, and hence produce quite a different value.

Thus far, it has been demonstrated how a difference in the spread of errors can result in quite different values from different accuracy variables. As shown in the previous section, correlations were highest between the numbers of errors per 100 words and the adjusted global accuracy ratings. In addition, a t-test on averaged global accuracy ratings for Task A ($M = 2.47$, $SD = .68$) and Task B ($M = 2.55$, $SD = .63$) revealed a significant difference ($t(64) = -2.614$, $p < .05$) with a very small effect of $-.12$. This is the same as the result for errors per 100 words. Given these two findings, it can be concluded that the errors per 100 words is the most valid in this study.

Conclusion

This article has examined two strands of research: Strand 1 focused on the measurement of syntactic complexity, Strand 2 on accuracy. Despite their different foci, both strands were aimed at examining the variables that are conventionally used in task-based research and to highlight precautions that researchers could take when conducting research on monologic speaking tasks.

It was found that, on both tasks, the syntactic complexity variables did not correlate highly with one another with the exception of AS-unit length and subordinate clauses per AS-unit and clause length (RQ1-1). Discriminant analysis showed results which were not consistent with the suggestions made by Norris and Ortega (2009) (RQ1-2). Considering that their suggestions were for L2 writing rather than speaking, such inconsistent results may not be so surprising. Rather, it is worth considering how complex the researchers expect the spoken performances to be and to pilot tasks carefully. It was found in this study that the results were influenced by the differing degrees of task-essentialness for subordination of the two tasks, which implies that careful piloting prior to any manipulation in task-based research is essential. The discrepancies between the findings in this study and the suggestions by Norris and Ortega (2009) may also be due to the 'ceiling effect' of the tasks, possibly not pushing the participants with higher proficiency enough to demonstrate the full range of syntactically complex structures that they have at their disposal. Knowing beforehand the ‘ceiling’ of performances that can be elicited by the tasks in question is vitally important, and this can ideally be done by using native speakers’ data as baseline (as they are free from the constraints of L2 processing) as Skehan (2009) strongly recommends.

For accuracy, all the variables correlated moderately highly with the ratings on both tasks (RQ2-1), but the errors per 100 words measure was most in line with the ratings in capturing the differences of the performances on the two tasks (RQ2-2). Analysis of transcripts revealed that these results were strongly influenced by the denominators of the variables and how errors were distributed in the performances. This also emphasises the importance of checking beforehand the profiles of participant groups under investigation, such as the range of their proficiency levels and the degrees of mastery of relevant grammatical items (such as subordination), etc.

The limitations of this study lie in the relatively small size of the data, especially at B2 or higher level. The majority of participants in this study were at A2 or B1 levels, and more data at higher levels of proficiency may offer more concrete and reliable findings. Further, examining variables using a wider range of task types (rather than just narrative tasks) should make a significant contribution towards the wider field of task-based research. Finally, the use of CEFR Assessment Grid might not have been the best choice for giving ratings to the elicited performances, although it was decided to use this grid in this study because the grid and accompanying sample performances are publicly available and used extensively in L2 and language assessment research. Although the raters in this study understood and applied the scales in a reliable manner as indicated by multi-faceted Rasch analysis, the CEFR Assessment Grid contains some descriptors that are vague and difficult to interpret; so ideally, the rating scales should be tailored for the types of tasks under investigation.

Notes

1. Hill’s book (1960) was intended for preparing students for the oral and written compositions of the Cambridge Lower Certificate examination (i.e. Cambridge FCE today), and the vocabulary intended to be used is all in the General Service List of English Words (Hill, 1960). The author further restricted the range of vocabulary by selecting tasks which were very similar in storylines and characters.
2. The difficulty of two tasks, which was calculated by multi-faceted Rasch analysis based on the ratings of the elicited performances, was −0.14 logits for Task A and −0.54 logits for Task B. The difference was statistically significant (χ^2 (1, 455) = 24.7, p < .01). See Inoue (2013) for details.
3. Sample performances can be seen on the CIEP website: http://www.ciep.fr/en/publi_evalcert/dvd-productions-orales-cecrl/index.php. A booklet which explains the rationales for the sample performances (ffrench, n.d.) is also available for downloading.
4. Since this article does not handle interactive tasks, the column for Interaction in the original CEFR Oral Assessment Grid was replaced by Sustained Monologue (Council of Europe 2001: 58–9). Revising the grid is supported by the Council of Europe (2009) in order better to suit the rating of performances in the samples.
5. Norris and Ortega (2009) recommended using the Coordination Index (Bardovi-Harlig 1992), which uses the numbers of sentences, clauses and coordination devices. Since the spoken narrative data in this article does not use sentences for analysis, it was decided to calculate the amount of coordination per AS-unit, as opposed to the amount of subordination per AS-unit.

References

Bachman, L.F. 2004. Statistical Analyses for Language Assessment. Cambridge: Cambridge University Press.

Bardovi-Harlig, K. 1992. A second look at T-unit analysis: reconsidering the sentence. TESOL Quarterly 26, no. 2: 390–395.

Bond, T.G. and C.M. Fox. 2007. Applying the Rasch Model: Fundamental Measurement in the Human Sciences (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Bygate, M. 2001. Effects of task repetition on the structure and control of oral language. In Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing, ed. M. Bygate, P. Skehan and M. Swain, 23–48. Essex: Pearson Education.

Council of Europe. 2001. Common European Framework of Reference for Languages: learning, teaching, assessment. Cambridge: Cambridge University Press.

Council of Europe. 2009. Manual for relating language examinations to the Common European Framework of Reference for Languages (CEFR). http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp (accessed 24 November, 2009).

Crookes, G. 1989. Planning and interlanguage variability. Studies in Second Language Acquisition 11: 367–383.

Eckes, T. 2009. Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment; Section H: Many-Facet Rasch measurement. http://www.coe.int/t/dg4/linguistic/Source/CEF-refSupp-SectionH.pdf (accessed 23 February, 2010).

Ellis, R. and G. Barkhuizen. 2005. Analysing Learner Language. Oxford: Oxford University Press.

ffrench, A. n.d. Spoken performances illustrating the 6 levels of the Common European Framework of Reference for Languages: comments on the assigned levels in English. http://www.ciep.fr/en/publi_evalcert/dvd-productions-orales-cecrl/docs/comments_en.pdf (accessed 3 November, 2009).

Foster, P. and P. Skehan. 1996. The influence of planning and task type on second language performance. Studies in Second Language Acquisition 18: 299–323.

Foster, P., A. Tonkyn and G. Wigglesworth. 2000. Measuring spoken language: a unit for all reasons. Applied Linguistics 21: 354–374.

Hill, L. A. 1960. Picture Composition Book. London: Longman.

Housen, A. and F. Kuiken. 2009. Complexity, accuracy and fluency in second language acquisition. Applied Linguistics 30, no. 4: 461–473.

Hughes, A. and C. Lascaratou. 1982. Competing criteria for error gravity. ELT Journal 36, no. 3: 175–182.

Inoue, C. 2013. Task Equivalence in Speaking Tests. Bern: Peter Lang.

Iwashita, N., T. McNamara and C. Elder. 2001. Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. Language Learning 51, no. 3: 401–436.

Jarvis, S. 2002. Short texts, best-fitting curves and new measures of lexical diversity. Language Testing 19, no. 1: 57–84.

de Jong, N.H., R. Groenhout, R. Schoonen and J.H. Hulstijn. 2013. Second language fluency: speaking style or proficiency? Correcting measures of second language fluency for first language behavior. Applied Psycholinguistics 34: 1–21.

Kormos, J. and M. Dénes. 2004. Exploring measures and perceptions of fluency in the speech of second language learners. System 32: 145–164. doi:10.1016/j.system.2004.01.001

Loschky, L. and R. Bley-Vroman. 1993. Grammar and task-based methodology. In Tasks and Language Learning: Integrating Theory and Practice, ed. G. Crookes and S. Gass, 123–167. Clevedon: Multilingual Matters.

McCarthy, P.M. and S. Jarvis. 2010. MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods 42, no. 2: 381–392. doi:10.3758/BRM.42.2.381

Mehnert, U. 1998. The effects of different lengths of time for planning on second language performance. Studies in Second Language Acquisition 20: 83–108. doi:10.1017/S0272263198001041

Norris, J.M. and L. Ortega. 2009. Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. Applied Linguistics 30, no. 4: 555–578. doi:10.1093/applin/amp044

Norris, J.M. and P. Pfeiffer. 2003. Exploring the use and usefulness of ACTFL guidelines oral proficiency ratings in college foreign language departments. Foreign Language Annals 36: 572–581.

Oh, S. 2006. Investigating the relationship between fluency measures and second language writing placement test decisions. Unpublished Master’s diss. University of Hawai’i.

Ortega, L. 1999. Planning and focus on form in L2 oral performance. Studies in Second Language Acquisition 21, no. 1: 109–148.

Robinson, P. 1995. Task complexity and second language narrative discourse. Language Learning 45: 99–140.

Robinson, P. 2001. Task complexity, task difficulty, and task production: exploring interactions in a componential framework. Applied Linguistics 23: 27–57.

Robinson, P. 2007. Criteria for classifying and sequencing pedagogic tasks. In Investigating Tasks in Formal Language Learning, ed. M.d.P. García Mayo, 7–26. Clevedon: Multilingual Matters.

Skehan, P. 1996. A framework for the implementation of task based instruction. Applied Linguistics 17: 38–62.

Skehan, P. 2009. Modelling second language performance: Integrating complexity, accuracy, fluency. Applied Linguistics 30: 510–532.

Skehan, P. and P. Foster. 1997. Task type and task processing conditions as influences on foreign language performance. Language Teaching Research 1: 185–211.

Skehan, P. and P. Foster. 1999. The influence of task structure and processing conditions on narrative retellings. Language Learning 49: 93–120.

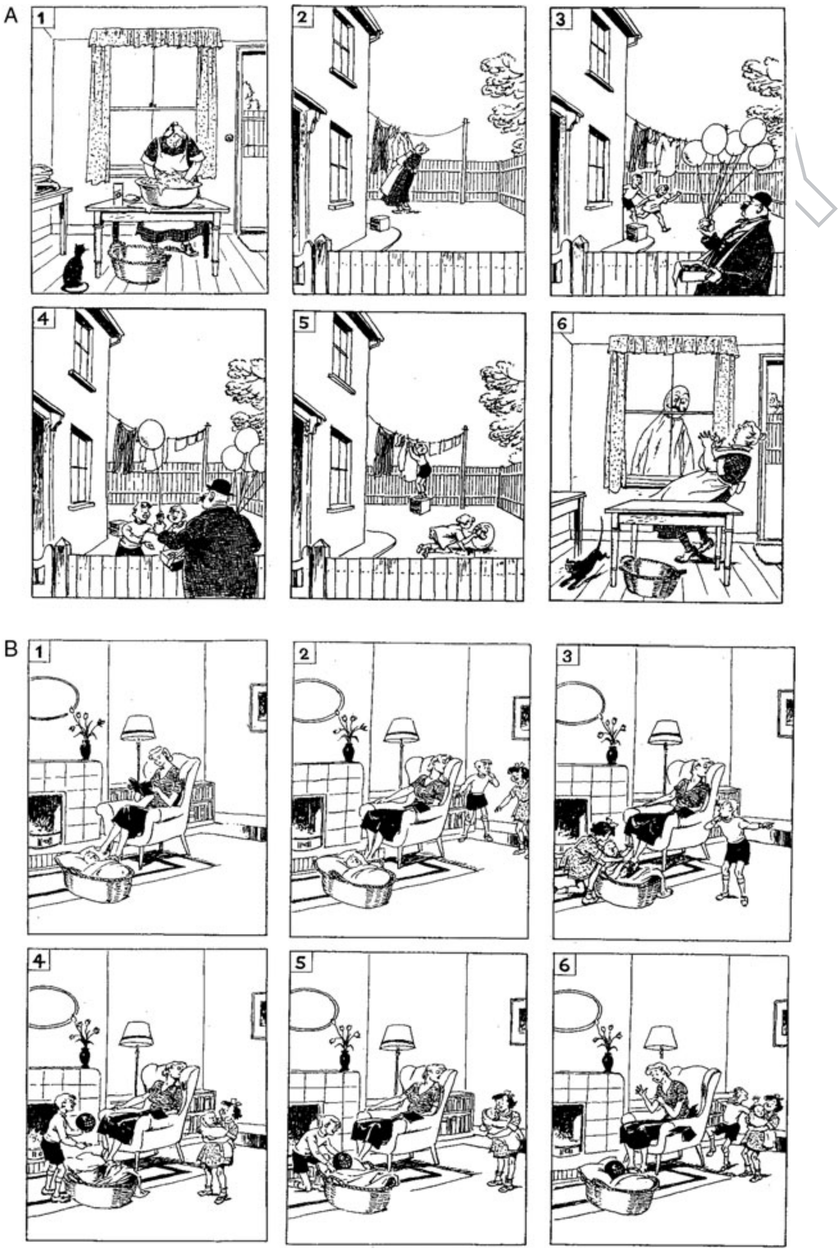
Tonkyn, A. 2013. Measuring and perceiving changes in oral complexity, accuracy and fluency: examining instructed learners' short term gains. In Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA, eds. A. Housen, F. Kuiken and I. Vedder, 221–245. Amsterdam: John Benjamins.

UCLES. 2001. CD ROM User Manual: Quick Placement Test. Oxford: Oxford University Press.

Wigglesworth, G. 1997. An investigation of planning time and proficiency level on oral test discourse. Language Testing 14: 167–197.

Yuan, F. and R. Ellis. 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.

Appendix 1: Tasks A and B (originally by Hill (1960); reprinted with permission.)



Appendix 2 Modified CEFR Oral Assessment Criteria Grid.

	Range	Accuracy	Fluency	Coherence	Sustained monologue
C1	Has a good command of a broad range of language allowing him/her to select a formulation to express him/ herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/ she wants to say	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices	Can give elaborate descriptions and narratives, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion
B2	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/ her mistakes	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some 'jumpiness' in a long contribution	Can give clear, detailed descriptions on a wide range of subjects related to his/her field of interest
B1	Has enough language to get by, with sufficient vocabulary to express him/ herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events	Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points	Can reasonably fluently relate a straightforward narrative or description as a linear sequence of points. Can give detailed accounts of experiences, describing feelings and reactions. Can describe events, real or imagined. Can narrate a story

(Continued)

Appendix 2 Continued.

	Range	Accuracy	Fluency	Coherence	Sustained monologue
A2	Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations	Uses some simple structures correctly, but still systematically makes basic mistakes. 94/146	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident.112	Can link groups of words with simple connectors like 'and', 'but' and 'because'.94/46	Can describe people, places, and possessions in simple terms. Can give a simple description or presentation of people, living or working conditions, daily routines, likes/dislikes, etc. as a short series of simple phrases and sentences linked into a list
A1	Framework Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorised repertoire	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication	Can link words or groups of words with very basic linear connectors like 'and' or 'then'	Can produce simple mainly isolated phrases about people and places