DOCUMENT RESUME

ED 275 164 FL 016 068

AUTHOR Arnaud, Pierre J. L.
TITLE The Lexical Richness of L2 Written Productions and the Validity of Vocabulary Tests.
PUB DATE 84
NOTE 17p.; In: Practice and Problems in Language Testing. Papers from the International Symposium on Language Testing (7th, Colchester, England, 1984); see FL 016 066.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; *Discourse Analysis; English (Second Language); Foreign Countries; Higher Education; *Language Tests; Second Language Learning; Test Construction; Testing; *Test Validity; *Vocabulary Skills; *Written Language
IDENTIFIERS France

ABSTRACT

This study investigates the validity of second language vocabulary tests, and whether or not test scores accurately reflect the quality of language behavior in real-life situations. It attempts to prove the validity of vocabulary testing by comparing test scores to indices of vocabulary richness in second language production. The student subjects were assigned an essay on a designated topic. Their writing was measured by lexical analysis (lemmatization, text length, and word rareness). The essays of French students of English and of American students studying in France were compared. Results did not show text length to correlate with vocabulary richness. However, results indicated that a low, but significant, correlation existed between discrete-item test scores and lexical richness variables in each homogeneous group. (MSE)

THE LEXICAL RICHNESS OF L2 WRITTEN PRODUCTIONS AND THE VALIDITY OF VOCABULARY TESTS

PIERRE J.L. ARNAUD
(University Lyon 2)

THE LEXICAL RICHNESS OF L2 WRITTEN PRODUCTIONS AND THE VALIDITY OF VOCABULARY TESTS

Pierre J.L. Arnaud (Université Lyon 2)

1 Introduction

Discrete-item tests of separate components of language, although still in use for practical reasons, have tended to fall out of fashion as language testing theory has begun to place more emphasis on validity. It seems, however, that vocabulary tests should be distinguished from the other categories.

In terms of content validity, even keeping in mind the often repeated caveat that a vocabulary is not a "bag of words", it is clear that words are by and large individual entities, and therefore that the lexicon lends itself well to the construction of discrete items. Furthermore, some types of vocabulary tests, such as those in which the testees are required to produce the name of illustrated objects or actions, correspond to the naming or labelling activity which is a frequent part of language behaviour in everyday life, and are immune from the accusation of eliciting only "language-like behaviour".

No reproach can be formulated against vocabulary tests on grounds of concurrent validity, either. In Ingram's (1968) words, the vocabulary test is "the nearest thing we have to a foolproof test, i.e. a constructor-proof test. The statistics, both item analysis and correlation figures, are nearly always good."

However, the ultimate proof of the validity of a test is external: do test scores, obtained in an artificial situation, reflect accurately the quality of language behaviour in real-life circumstances ? The purpose of the present research is to answer this question by comparing vocabulary test scores to indices of vocabulary richness of L2 productions.

2 Previous research

Vocabulary richness has been investigated intensively in lexicometric and stylostatistical research. This can provide useful methodological guidelines, in spite of the fact that the texts examined bear little resemblance to the productions of foreign-language learners.

By vocabulary richness, one usually understands the number of types (V) in a given length of text, i.e. in a given number of occurrences or tokens (N). Muller (1977:115) specifically excludes any other measure, such as the degree of rareness of the vocabulary, which is not a "fact of structure", but a "fact of content". The comparison of texts of different lengths in terms of lexical richness poses a number of problems, as new words occur in decreasing numbers as a text unfolds itself. This can be explained by means of the "urn model":

the lexicon of a writer can be compared to an urn, from which a lexeme/type is drawn each time a token is produced in discourse; the type is then placed back into the urn, from which it can be drawn again. Supposing the lexicon of a writer is not infinite, there remain fewer and fewer lexemes that have not been drawn before. An index such as V/N is therefore unstable. Other indices for the comparison of texts of different lengths have been experimented with, but they have proved disappointing (Ménard 1972:105f.; Muller 1977:117; Dugast 1978). In order to make legitimate comparisons, it is therefore necessary to shorten the longer text, which can be done by actually deleting tokens or, alternatively, by using Muller's method of simulated draws based on the binomial law: knowing the distribution of frequencies for a given length of text, it is possible to calculate the theoretical V' for an inferior length (Muller 1977:101f.,127f.). It should be noted, however, that the confidence interval of V' becomes wider as the original text is more radically shortened.

Ménard (1972,1978) has investigated the relationship between lexical richness, as defined above, and the presence of rare words in a text. After comparing the effects of various empirical norms of rareness, he concludes that too restrictive a definition (such as absence from a dictionary) leads to a Poisson distribution of rare words, thus making statistical treatment difficult. When more liberal norms are applied, highly significant rank correlations appear between lexical richness and the proportion of rare words in V (1).

Less attention has been devoted to lexical richness by applied linguists and the author has found only two publications on the subject: Linnarud (1975) and Mendelsohn (1981). Linnarud, in a pilot study, examines a corpus of 36 essays written in English by Swedish students; in addition, three native speakers, enrolled in the same course, wrote an essay on the same subject. As the author herself points out, the choice of the subject ("Sir, I protest...") was unfortunate as it led to very divergent responses and thus to poor control of vocabulary content. The essays varied in length from 124 to 573 words. Two main variables were taken into account: lexical density and lexical variation. Lexical density is the percentage of lexical words in the text: $LD = (N_{lex} \times 100) / N$. Linnarud mentions difficulties due to the lack of firm criteria for distinguishing lexical words from grammatical ones, and the need to make carefully weighed decisions in some cases. As some essays had a high LD, but contained a poor, repetitious vocabulary, lexical variation was also measured, with the following formula: $LV = (V_{lex} \times 100) / N_{lex}$. Measures were collected once with errors deleted and once with errors retained; only the first series is reported here. The lexical density statistics (minimum: 30.86; mean: 38.51; maximum: 47.15) fall below the values observed by Ure (1971) on L1 texts. The LD of the essays written by native speakers is higher. As the different lengths of the essays made their comparison in terms of lexical variation illegitimate, Linnarud divided them into four groups according to length, and within each group classified the essays into three lexical variation categories: here, too, the native speakers scored better; their essays also contained a higher percentage of hapax. Finally, the essays were classified into three groups according to their global value, and groups were compared in terms of mean length: the "poor" essays are the shortest, but the "average", and not the

"good" ones, are the longest. However, the simple statistical treatments used in this pilot research provide suggestions more than definitive results.

Mendelsohn (1981) studied the oral productions of 42 subjects, native speakers of English and learners. He measured three variables that, according to him, constitute lexical richness. Lexical density was established by means of a formula slightly different from Linnarud's, but which would yield the same rank orders: $LD = (N_{lex} \times 100)/N_{gr}$. Like Linnarud, Mendelsohn mentions the difficulties due to the lexical/grammatical distinction, and recommends keeping notes of decisions made in ambiguous cases. Lexical variation was also measured using a slightly different method: $LV = (V_{lex} \times 100) / N$. A "semantic variation" measure was introduced by Mendelsohn: when students are required to speak on a specific subject, a certain number of "universes of discourse" can be expected to appear in their productions, and the number of types belonging to them can be counted. Mendelsohn found that this was the most discriminating of the three measures of lexical richness. In addition, he measured the percentage of errors. As LD and LV yielded a rank-correlation coefficient of .78 and error percentage and lexical richness one of only .395, the author concludes that lexical richness and errors represent different dimensions of lexical performance, and that both should be taken into account for evaluation purposes. In most situations, however, it seems that errors are the only criterion applied: observation of a group of 35 evaluators working on the corpus showed that subjective judgments were mainly based on them, and that lexical richness had little influence. Indeed, another group, required to classify subjectively a set of transcripts for lexical richness, performed poorly on the task.

Other studies dealing with the assessment of written productions of L2 learners have yielded interesting results concerning the role of vocabulary in the formation of judgments. Perkins (1981), comparing three methods of essay evaluation: global subjective, analytic subjective (i.e. a group of judgments on criteria such as relevance, fluency, grammar, mechanics, vocabulary), and objective (number of errors, number of words per T-unit) found that, among the five subjective criteria, vocabulary had the lowest correlations with the other variables and the subjective global score. On the contrary, Mullen (1980), working on a corpus of 117 texts assessed by four judges globally and analytically along four dimensions, structure, organization of ideas, quantity of text produced and vocabulary, found that vocabulary was best correlated with the global score.

Finally, the relationship between essay length and quality has been investigated. Linnarud's (1975) results have already been quoted. Larsen-Freeman and Strom (1977) and Neumann (1977), as quoted by Perkins (1980), found a significant correlation between essay length and overall quality, whereas Perkins himself did not observe such a relationship. Again, Mullen (1980) reports a high correlation between quantity of text and global score, but this quantity itself was estimated subjectively.

3 Method

3.1 Administration

The subjects were first-year specialist students of English at Lyon 2 University. In order to reach a sufficient population, the experiment was carried out during normal teaching hours. The vocabulary test was administered first; it was a 26-item productive test (translation from French into English) with a reliability (K.R./Horst) of .89 (Arnaud 1981). On the following week, an hour was devoted to the essay writing. To ensure motivation, global scores were taken into account for continuous assessment. The subject was chosen so as to give every student something to say:

"What do you suggest to improve secondary education in France ?"

Data treatment possibilities led to the retention of a random sample of 100 test/essay pairs. In addition, four sophomores from Dartmouth College, then at Lyon 2 on a regular programme, wrote an essay on the following, symmetrical subject:

"What do you suggest to improve high school education in the U.S.A. ?"

Although this control group is too small to allow for statistical comparisons, and in spite of the fact that the local students were hardly comparable to those in academic terms, a useful reference is thus provided.

3.2 General considerations

The essays were first reviewed subjectively. An essay can be seen as the conjunction of a number of ideas and their connections on one hand and the linguistic means to express them on the other hand. There exists a much larger degree of independence between the two in the case of L2 learners than in the case of native speakers, and on examining a poor essay it is not always easy to determine what is due to a paucity of ideas or to low language proficiency. Furthermore, when one tries to concentrate on the purely linguistic aspects, clear-cut components seldom appear prominently: for instance, one would hardly ever come across an essay in which the sentence grammar was impeccable and the text grammar disastrous. The following sample is typical of this kind of problem:

"We should find other way for holiday. The summer holidays are too long. During often three months, the pupils forget what they have learnt in the year. It should be better to have more holiday but shorter holiday and to finish the school at 4 o'clock and not at 6 o'clock."

Such repetition might be attributed to lack of vocabulary or ignorance of cohesive devices, but the fragment happens to be strikingly similar to the following (oral, L1) passage discussed by de Beaugrande and Dressler (1981:54):

"There's water through many homes. I would say all of them have water in them. It's just completely under water."

The authors suggest that this discourse is due to lack of planning time and rapid loss of the surface text, two phenomena which obviously play a role when students at this proficiency level have to write English texts in a limited time. An essay including many repetitions like the one quoted from above would result in a very low lexical

variation score, which would however to a great extent be the effect of non-lexical causes. This should serve as a reminder that total isolation of one component is probably illusory and that one should not place excessive trust in objective measures of written text variables.

A number of facts about lexical quality do appear clearly, however. As impressions are produced essentially by lexical words, non-lexical tokens can be left aside, as in Linnarud's (1975) and Mendelsohn's (1981) studies. Some of the essays strike the reader as containing a very limited number of lexical items repeated interminably; this is so widespread a phenomenon, and one that lowers the overall quality of an essay so much that it must be taken into account in any definition of L2 lexical richness. As Mendelsohn (1981) had found that lexical density and lexical variation are highly correlated, it was decided to determine lexical variation through the following formula:

$$LV = \frac{V_{lex}}{N}$$

In other cases, attention is attracted to the vague, general vocabulary used by the writer. For instance, the only way some students were able to express value judgments was literally to use good and not good. What is involved here is the degree of rareness of the vocabulary and it was therefore decided to assess it.

Finally, the number of lexical errors is sometimes so high as to hinder communication. Naturally, the author is well aware that, as a non-native speaker of English, he may be more intolerant of errors than a native speaker: James (1977) found that it is the non-natives who are less tolerant of lexical errors; Hughes and Lascaratou (1982) have shown that non-natives react to the fundamental nature of the rules that are transgressed, whereas native speakers are sensitive to the way communication is affected. Mendelsohn (1981), as seen above, has brought to light the importance of errors in the subjective evaluation of lexical quality. The present research therefore differs from classical stylistic studies of L1 texts in that it includes the measurement of errors and vocabulary rareness.

3.3 Data treatment

3.3.1 Editing norms

The texts had to be treated on a small computer (2) with programs written by the author, and this required great simplicity in the operations, which made the following editing norms necessary.

Proper nouns (except country names) and numbers were deleted.

Compound words: after their existence as items was checked in a dictionary (Hornby 1974), they were typed hyphenated and thus treated as units by the program.

Verb-particle constructions were typed separated or hyphenated according to the case:

He turned down the street.
He turned-down the offer.

As syntactic errors were not of direct relevance to this research, they were typed verbatim. In most cases, this has no influence on the lexical richness indices:

There is too much pupils in the schools.

but N is sometimes affected:

The philosophy is not properly taught.

This is another proof of the impossibility to isolate lexical aspects of writing entirely.

Morphological errors were corrected when it was clear that ledge of the lexeme was not affected (but see below).

Lexical errors: although we are not concerned here with [illegible] lysis, such regularities occurred that a brief survey is of [illegible] note the prevailing influence of interference.

minor spelling mistakes: personnal, teatcher
major spelling mistakes: schoII
[illegible]vation mistakes: to comparate, he successed
faux-amis (deceptive cognates): They should be prev[illegible] is difficult
interference from another language on the curriculum: [illegible] money
confusion between two lexemes: The teachers learn them [illegible]th

Although the gravity of these types of errors is variable an[illegible] nothing system was applied: any lexically erroneous form was re[illegible] with an error code, and was counted in N but not in V.

Borderline cases: some morphological errors indicate that the corresponding lexeme is not stored correctly: errors on irregular verbs were treated as lexical. Similarly, the preposition in the following example was treated as a lexical mistake, as one cannot consider that a lexeme is known if the accompanying structures are not mastered

They listen at the teacher.

P[illegible] otherwise treated as grammatical items.

Problems due to the subject: the lack of correspondence between the American, British and French educational systems caused a number of problems. When a student used a term from an Anglo-saxon context to designate a French reality, this was not considered as a mistake, but the word was ignored. Furthermore, it was often difficult to determine whether a particular use was legitimate or constituted an interference error, as the context was ambiguous: for instance, it was seldom clear whether minister stood for ministère or ministre, and if class was intended to mean class or classroom (both classe in French). The benefit of the doubt was systematically accorded, [illegible] this introduced a further lack of precision in the measurement.

3.3.2 Lexical vs. grammatical words

For automatic treatment purposes, it was out of the question to have to make a decision every time an ambiguous form turned up; tokens like have or do were therefore systematically treated as grammatical. As a filter for non-lexical occurrences, the program included a dictionary of grammatical forms to which each token was compared in turn. As treatment time grew exponentially with the size of this dictionary, it had to be limited to approximately 100 items. These are the first grammatical items in the rank-order list of the Hofland and Johansson (1982) frequency count. The classical closed class/open class distinction was therefore not respected, as many closed-class members were left outside the dictionary, but this was not really a weakness, for few of the grammatical items present in the corpus passed through the filter, given the proficiency level of the subjects; furthermore, knowledge of a rare grammatical word cannot be considered as independent from lexical knowledge.

3.3.3 Lemmatization

Lemmatization was carried out manually on the print-outs. Plurals were grouped together with the singulars, conjugated forms with the infinitive, comparatives and superlatives with the positive degree. Only one lemma was counted for forms belonging to two categories (book's, booked).

3.3.4 Lengths and indices

The distribution of essay lengths is represented in Figure 1. Variation is considerable.

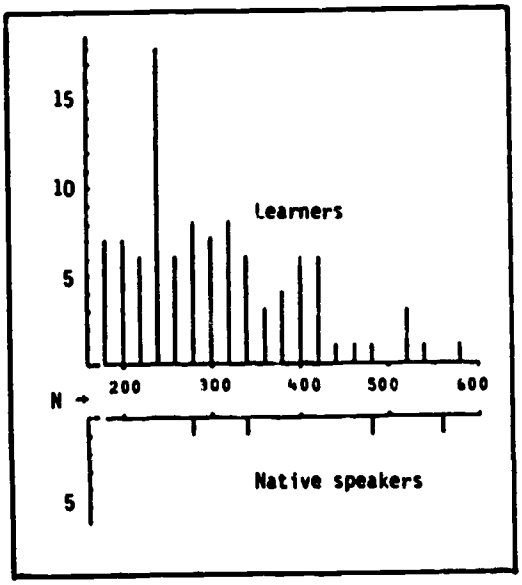


Figure 1

distribution of N (essay lengths)

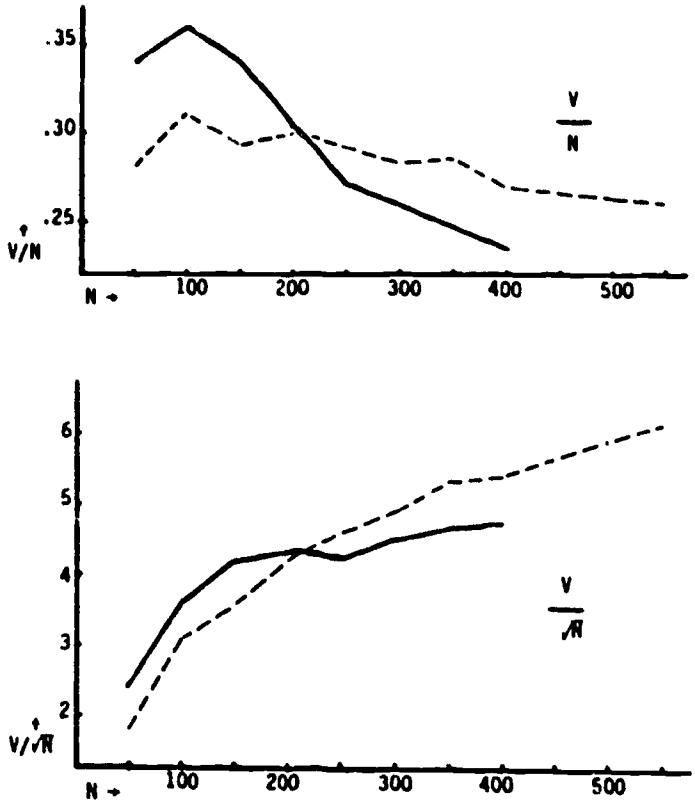
M = 313

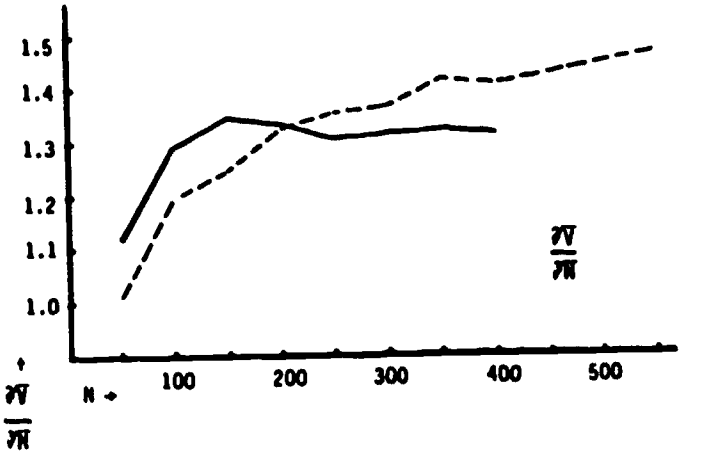
SV = 91

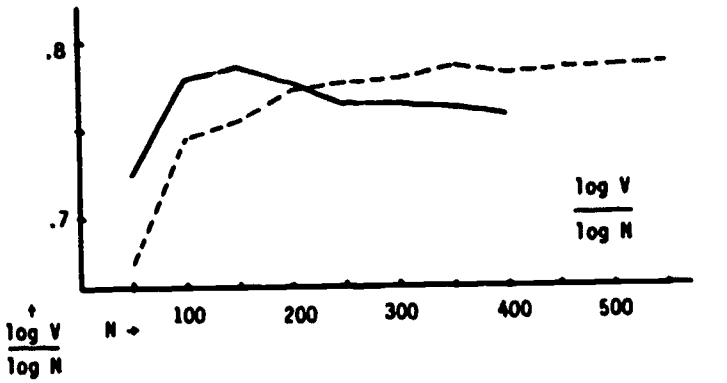
(N_{Ss}: 100 + 4)

Figure 2

Evolution of V/N and other indices: successive 50-word fragments of two essays







The graphs of Figure 2 prove empirically that V_{lex}/N and indices derived from it are unsuitable for the establishment of lexical variation scores on L2 texts in this length range. It was therefore necessary to bring down all the essays to the length of the shortest one, N = 180. Use of Muller's (1977) method of simulated draws would have been exceedingly time-consuming, so token deletion was chosen instead. It was impossible to retain, for instance, the first 180 tokens of each essay as this would not have ensured adequate sampling. A series of instructions was therefore added to the program for the random selection of 180 tokens. The lexical variation score hence-

forth corresponds simply to V_{lex} and the error score (E) to the number of lexical errors in the sample.

3.3.5 Norm of rareness

It would have been theoretically possible to calculate the average frequency of the types in each essay, using a frequency list, but such a task would have proved unsurmountable practically. An empirical solution like Ménard's (1978) was therefore retained, and the norm was chosen as wide as possible. The official French Ministry of Education vocabulary list for the classe de troisième (Elaboration 1976) was used. It contains 1 522 lexical items that are supposed to be known of all students by the time they reach the lycée level. A type was therefore considered as rare if it was not on the list. The following formula was used for establishing the rareness scores:

$$R = \frac{V_{rare}}{V_{lex}}$$

3.3.6 Reliabilities

As text shortening introduces a margin of uncertainty, the reliabilities of V, R and E were checked. The program was run four times over a sample of 20 essays, each time extracting a different set of 180 tokens. Product-moment coefficients of correlation were computed between the values yielded by runs I and II, and III and IV respectively. The results of the first and last two runs were then averaged, and the correlations between the two sets of means calculated. The figures therefore correspond to reliability coefficients for values obtained after two runs of the program, which was the case for the remaining 80 (+4) essays (Table 1).

Table 1: reliability coefficients of lexical richness variables after text shortening; bottom line: r's for mean values after two runs of the program (20 texts)

V				R				E			
I	II	III	IV	I	II	III	IV	I	II	III	IV
.88		.73		.86		.90		.78		.97	
	.87				.95				.92		

3.3.7 Distributions

The distributions of V, R and E are reproduced in Figures 3, 4 and 5. Interestingly, in terms of lexical variation and rareness of vocabulary, some French students performed as well as their American counterparts. The error scores of two of the latter are due to spelling mistakes.

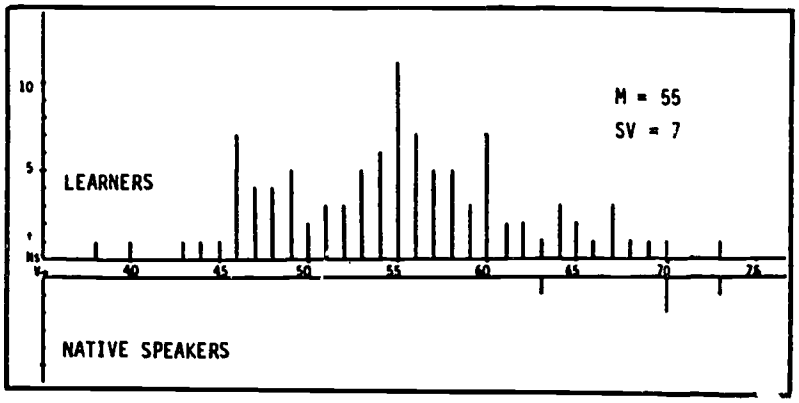


Figure 3: distribution of V

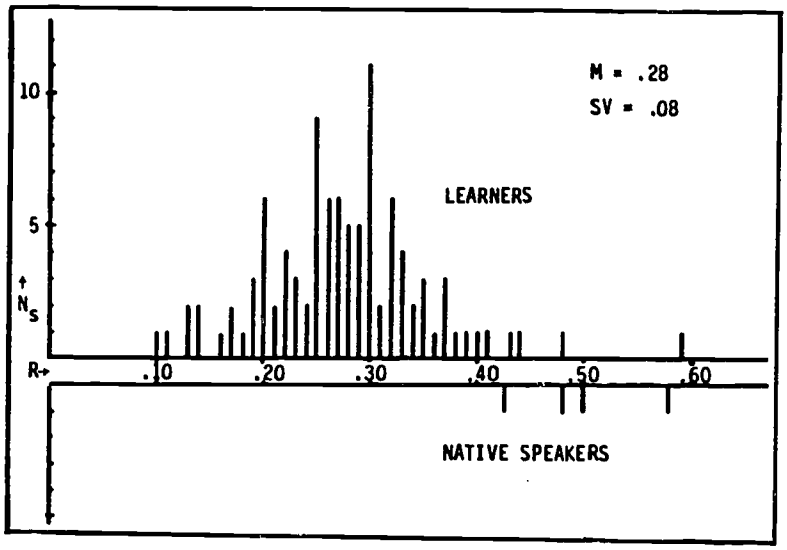


Figure 4: distribution of R

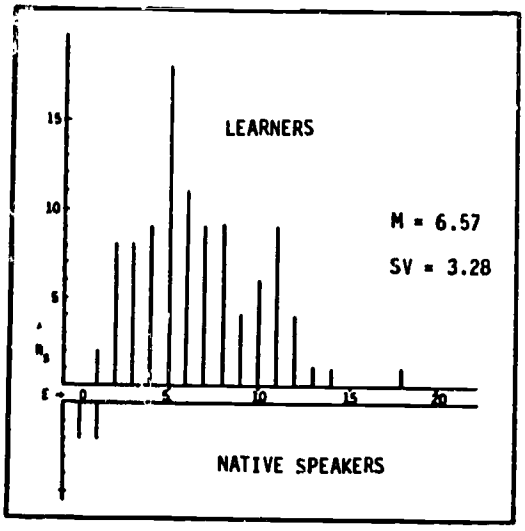


Figure 5: distribution of E

4 Results and discussion

It is not unreasonable to think that the more proficient students are able to write longer texts in a limited time. Subjective examination, however, had shown that some of the longest essays were among the most repetitive and unconnected. This impression is confirmed, as far as vocabulary richness is concerned, by the correlation coefficients in Table 2, none of which is significant. In view of the contradictory results of previous research (see section 2), it seems that no generalizations should be made in this respect.

Table 2: correlations between essay length and the three lexical richness variables as established on 180-token samples

r_{NV}	=	.06
r_{NR}	=	.10
r_{NE}	=	.05

Product-moment coefficients of correlation between test scores (T) and the three lexical richness variables are reproduced in Table 3. There obtain significant, but low correlations between V and R and V and E (3); R appears to be uncorrelated to T and E. However, it would be unwise to conclude that R measures a different dimension of lexical richness before replication studies can be carried out.

Table 3: correlations of test scores (T) and lexical richness variables (N_s = 100); * : p<.025; ** : p<.01

	T	V	R	E
T		.36**	.09	-.21*
V			.27**	-.24**
R				.09

At this stage, a few points should be remembered. First, a test score is the result of a concentrated measurement situation, which is not the case of an essay. Secondly, a substantial amount of uncertainty was introduced successively by editing norms, decisions as to errors and text shortening; absolute isolation of the lexical component was furthermore impossible. Finally, the students, all native speakers of French with fresh baccalauréats, constituted a very homogeneous group and this obviously lowered statistics (see Raatz and Klein-Braley 1981). In view of these considerations, the low, but significant correlations between T and V and T and E make it possible to conclude that discrete-item vocabulary tests are valid.

5 Suggestions for further research

One of the questions which it has not been possible to answer in this study is that of the concurrent validity of the lexical richness measures. The only way to do this would be to analyse two sets of essays written by the same subjects, which was administratively beyond the author's reach.

For further research, use of the following empirical index is suggested:

$$LR = V_{lex} + V_{rare} - 2E$$

(on text samples of the same length). Rare types and errors are given double weight. This index was found to separate native speakers from learners very clearly.

Notes

1 Ménard, however, does not prove that an index like V_{rare} / V is independent from the size of V.

2 The author wishes to thank his colleagues of the Service Informatique, Université Lyon 2 for providing unlimited access to the computer and indispensable advice.

3 See Note 1 for a possible dependence between R and V. Furthermore, as error words are at the same time counted in E and ignored in the establishment of V, the two measures are not entirely independent.

Bibliography

Arnaud, P.J.L. 1981. Comparaison pratique de cinq types de tests de vocabulaire anglais et recherche sur la nature de la compétence en vocabulaire. ms. (to be presented at the 7th World Congress of Applied Linguistics, Brussels, Aug. 5-10, 1984)

Beaugrande, R.de, and W. Dressler. 1981. Introduction to text linguistics. London: Longman

Culhane, T., Klein-Braley, C., and D.K. Stevenson (eds). 1981. Practice and problems in language testing. University of Essex, Department of Language and Linguistics, Occasional Papers 26

Dugast, D. 1978. "Sur quoi se fonde la notion d'étendue théorique du vocabulaire?". Le Français Moderne 46: 25-32

Elaboration d'un programme lexical pour l'enseignement de l'anglais dans le premier cycle. 1976. Paris: I.N.R.D.P., Recherches Pédagogiques 84

Hofland, K., and S. Johansson. 198 d-frequencies in British and American English. Bergen: The .. gian Computing Centre for the Humanities

Hornby, A.S. 1974. Oxford advanced learner's dictionary of current English. London: O.U.P.

Hughes, A., and C.Lascaratou. 1982. "Competing criteria for error gravity". English Language Teaching Journal 36: 175-181

Ingram, E. 1968. "Attainment and diagnostic testing". In Davies, A. (ed). 1968. Language testing symposium. London: O.U.P. 70-97

James, C. 1977. "Judgments of error gravities". English Language Teaching Journal 31: 116-124

Larsen-Freeman, D., and V. Strom. 1977. "The search for a second-language acquisition index of development". Workpapers in TESL 11: 35-83

Linnarud, M. 1975. Lexis in free production: An analysis of the lexical texture of Swedish students' written work. University of Lund, Department of English: Swedish-English Contrastive Studies, report n°6

Ménard, N. 1972. Mesure de la richesse lexicale, Méthodologie et vérifications expérimentales. Thèse de troisième cycle, Université Strasbourg 2

- . 1978. "Richesse lexicale et mots rares". Le Français Moderne 46: 33-43

Mendelsohn, D.J. 1981. We should assess lexical richness, not only lexical error. Paper presented at the 1981 TESOL convention, Detroit

Mullen, K.A. 1980. "Evaluating Writing Proficiency in ESL". In Oller, J.W., and K. Perkins (eds). 1980. Research in language testing. Rowley: Newbury House. 160-170

Muller, C. 1977. Principes et méthodes de statistique lexicale. Paris: Hachette

Neumann, R. 1977. An attempt to define through error analysis the intermediate level at UCLA. unpublished thesis, UCLA

Perkins, K. 1980 "Using objective methods of attained proficiency to discriminate among holistic evaluations". TESOL Quarterly 14: 61-69

- . 1981. "The test of ability to subordinate: Predictive and concurrent validity for attained ESL composition". In Culhane & al. (eds). 1981. 104-112

Raatz, U., and C.Klein-Braley. 1981. "The C-test - A modification of the cloze procedure". In Culhane & al. (eds). 1981. 113-138

Ure, J. 1971. "Lexical density and register differentiation". In Perren, G.E., and J.L.M. Trim (eds). 1971. Applications of linguistics. Cambridge: C.U.P. 443-452