What percentage of text-lexis is essential for comprehension?

Batia Laufer
University of Haifa

1989

D.2

25 What Percentage of Text-Lexis is Essential for Comprehension?

Batia Laufer
University of Haifa

INTRODUCTION

Lexis, a neglected aspect of second language acquisition (Levenston, 1979; Meara, 1980) of the sixties and seventies, has been increasingly gaining attention in the eighties. Researchers, material designers and instructors have become more interested than before in vocabulary acquisition in general (see for example Nation, 1983; Ludwig, 1984; Ringbom, 1985; Giacobbe and Cammarota, 1986), in the place of lexis in language production and comprehension (Deville et al., 1985; Kelly, 1985; Ostyn and Godin, 1985) and in vocabulary testing (Read and Nation, 1986; Takala, 1984, 1985).

In the field of Reading for Specific Purposes, it has been realized that the most serious difficulties that foreign learners experience are lexical, particularly at intermediate and advanced levels of reading. This is so because text interpretation was found to be dependent mainly on the lexical and conceptual cues (Ulijn, 1984; Laufer and Sim, 1985a, 1985b) [1]. Syntactic analysis "would normally be superficial except when the lexicon and the conceptual system do not provide enough cues to comprehend the text" (Ulijn, 1984). Moreover, foreign readers sometimes distort the syntax of the text so that it can fit their interpretation of words (Laufer and Sim, 1985).

One of the questions related to lexis and reading that has provoked interest is how many words [2] should a person know in a foreign language to be able to read an authentic text. Though differences in estimates can be found (which are apparently due to the different definitions of the concept of 'word'), recent studies suggest that the receptive vocabulary necessary for comprehension is about 5000 lexical items (Hindmarsh, 1980; Deville et al., 1985; Ostyn and Godin, 1985). The justification for this number of words is indirect. Hindmarsh claims that the above amount is necessary to pass the reading component of the Cambridge First Certificate exam; Deville et al. and Ostyn and Godin claim that a lexicon of

5000 words would give a coverage of 90-98% of the lexis of an authentic text and that the above coverage is necessary for comprehension. It is not clear, however, whether it was actually shown empirically that lower, lexical coverage will indeed result in inadequate comprehension in a significantly large number of cases. In other words, it would be useful to find out whether lexical coverage of 95% (or maybe a lower or higher one) is a determing factor in comprehension, rather than simply an enhancing one.

THE STUDY

Purpose

The present study attempted to measure the relationship between the number of words understood by a reader in an academic text and the quality of comprehension of the text. Specifically, it addressed the following question: What percentage of word tokens must be understood to ensure 'reasonable' [3] reading comprehension of a text? An answer to this question would also provide an indication as to what number of words should be aimed at in teaching if comprehension of authentic material is the aim of the learner. If 95% of lexical coverage is the absolute minimum for reasonable comprehension, then 5000 words should be learnt since this is the number that will give such coverage. If 90% is the necessary coverage then about 3300 words would probably suffice (Deville et al., 1985).

Subjects

The subjects in the experiment were 100 first year University students (from various departments), native speakers of Hebrew and Arabic, who were taking a course in English for Academic Purposes. The purpose of the course was to improve the students' comprehension of academic literature in English.

Procedure

The subjects were given a reading comprehension test (hence RC) which consisted of one of two texts of a general academic nature: one standardized test (a text from the reading component of "Examen Hoger Algemeen Vortgezet Onderwijs" with multiple choice questions; the other - self made with open-ended questions [4]. The learners were assigned a double task: answering comprehension questions; underlining the words they could not understand in the text [5]. The papers which included the answers and the text with the underlined unknown words were then collected.

In the second stage of the experiment, the subjects were given a lexical coverage test: This test consisted of clean copies of the same text as before and a list of 40 words from each text. The subjects were asked to translate the words or paraphrase their meaning in text-context. The lists with the translations were then collected.

For each testee, two scores were calculated: a) a reading comprehension score (RC) on the basis of the answers to comprehension questions; b) a lexical coverage (LC) score, on the basis of the information provided by the underlining of the unknown words and the translations of the vocabulary lists. The latter was done as follows. For each testee, a comparison was made between the underlined words and the list of the translations. A word could be correctly translated and not underlined as unfamiliar, i. e. the testee considered the word as known and was right in his judgement. A word could be left not translated at all and underlined in the text, i. e. reported as unfamiliar. A word could also be mistranslated or not translated at all and yet not underlined in the text, i. e. reported as familiar, but mistakenly so. The number of such discrepancies (between the testees' reports and their actual knowledge) was added up for each subject, and 'awareness index' was calculated. This was 40 (the number of words on the word-translation list) minus the number of words mistakenly judged as familiar out of the 40 words, converted into percentage. The number of unknown words in the text was then calculated by dividing the reported number of unknown words in the text by the 'awareness index', i. e. adjusting the testee's self-assessment by means of increasing the number of reported words by the percentage of words he mistakenly thought he knew. The number of unknown words was then subtracted from the total number of text words. The result, the total number of familiar words, was converted into percentage. The percentage of words known by a testee in a text was taken to the lexical coverage (LC) score.

RESULTS

The two sets of scores - the LC score and the RC score - were analyzed; the LC score was taken to be the independent variable, the RC score - the dependent one.

A 2 test showed that the group that scored 95% and above on the LC test had a significantly higher number of 'readers' (scores of 55% and above) than 'non-readers' (54 and below) in the RC test by comparison with the group that scored lower than 95% on the LC test. (2 df1 23.28 > 10.82 p = 0.001). The difference was not significant, however, between the groups: 90% and above on LC tests and 89% and below. (2 df1 = 1.7 < 3.84). When the top group (95% and above) was compared with the 90-94% LC group, the difference between the number of 'readers' and 'non-readers' was significant again (2 df1 = 13.52 > 10.82 p = 0.001).

A t-test comparing the mean scores on the RC tests showed that there was a significant difference between the 95% and above LC group and the 94% and below group. (t= 8.25 > 3.46 p = 0.001). The difference was not significant for the 90% ad above and 89% and below LC groups. As for the 90-94% group, the mean RC score was significantly lower than that of the 95% and above LC group. (t = 6.5 > 3.74 p = 0.001).

Discussion

In the RC test, the subjects were required to show comprehension of general ideas, to elicit and interpret factual information, supporting material, to draw inferences. In other words, they were performing tasks similar to those any reader of academic texts would be confronted with. As mentioned before, a score of 55% and above would indicate that the testee was a 'reader', i. e. could function in a real-life situation, while a score below that would be characteristic of the 'non-readers'.

The results of the study suggest that reading academic prose is likely to be greatly affected by the lexical knowledge of the text. A chance to become a 'reader' is significantly higher if the lexical coverage of the text is 95% and above, that is if 95% of the text's word tokens are familiar to the reader. This does not mean that a person cannot understand a text when the lexical coverage is lower than that. Other factors, like grammatical clues, text organization, subject-matter familiarity may facilitate comprehension for some readers with more limited lexis. Yet, the study shows that, in the majority of cases, when the lexical coverage was below 95% comprehension was impaired in spite of the other facilitating factors that might have affected reading [6].

Thus, the study supports the 'threshold hypothesis' in reading comprehension (Clarke, 1979, 1980; Cziko, 1980; Cummins, 1979) according to which reading is hampered by limited control of the language, and reading strategies can be applied only when language knowledge is above the threshold. The results of the present study suggest that the lexical component of this threshold level can be defined as 95% of the text-lexis; lower lexical coverage will be associated with unsatisfactory more often than with satisfactory comprehension.

As mentioned before, Deville et al. (1985) and Ostyn and Godin (1985) found that lexical coverage of 95% can be achieved when the learner's lexicon reaches 5000 words. It appears therefore that the lexical threshold necessary for successful reading is 5000 words since this is the number which results in the optimal lexical coverage of an authentic text, which in turn, ensures successful comprehension.

It could be argued that the study did not measure how many of the 95% of the words understood in the text had also been familiar to the reader before he read the text and how many of them were guessed in context. Whatever the proportion might have been, it is still true that 95% of word tokens were necessary to understand the text and that this coverage can be achieved by 5000 words in the learner's lexicon. Whether 95% was known and 0% guessed, or 90% known and 5% guessed, or 80 % known and 15% guessed, the total effect of 95% lexical coverage made the significant difference in comprehension scores. Therefore the argument that lexical guessing takes place cannot invalidate the importance of lexical core of 5000 words to be learnt. It is precisely this solid vocabulary knowledge that will provide the reader with the necessary context for successful guessing. It was shown (Laufer and Bensoussan, 1982) that without such context, it is misinterpretation rather than guessing that often takes place.

Reports are available on the size of vocabulary of students in different countries who embark on University studies and use English for reading academic literature. The vocabulary range reported is from 500 to 3000 (See Laufer, 1987). Is it therefore surprising that very often learners prefer a bad translation of their bibliography to the almost incomprehesible original? A gap of 2000-4000 words between the amount of words they know and they should know turns reading into 'mission impossible'. It is through increasing vocabulary to 5000 and ensuring a lexical coverage of 95% that reading has a good chance of becoming 'mission surmountable'.

SUMMARY AND CONCLUSION

The study attempted to find out what percentage of text-lexis was essential for a successful reading comprehension in a foreign language. In the experiment, two sets of scores were obtained and analysed: the reading comprehension score and the lexical coverage score. Different methods of analysis led to a similar conclusion: lexical coverage of 95% - the understanding of 95% of word-tokens in a text - can ensure reasonable reading comprehension, i. e. a score of 55% and above. Lower lexical coverage is associated with unsatisfactory more often than with satisfactory comprehension. Since the 95% coverage can be achieved by learning 5000 words, it is suggested that 5000 words seems to be the lexical threshold beneath which other facilitating factors in reading comprehension may not be very effective. Since many students who begin academic studies possess much poorer vocabularies massive vocabulary expansion should be one of the major goals of any course in Language for Academic Purposes.

NOTES

1. In L1 acquisition research, the importance of vocabulary was acknowledged earlier than that. (See for example reports on the relationship between lexical knowledge and reading comprehension in Anderson and Freebody, 1981).
2. A 'word' is taken to be a lexical item, i. e. the smallest unit in the meaning system of language.
3. 'Reasonable' reading comprehension was considered a score of 55% which is the lowest passing grade in our system in Haifa University.
4. Half of the students were given one text; the other half the other. The reliability of the standardized test is .81; the reliability of the self-made one .57.
5. If a word had been unfamiliar to the learner before it was seen in the text but was considered clear in the context, the instructions to the testees was <u>not</u> to mark it as unknown.
6. Since all the testees had a fairly similar educational/ cultural background and a similar training in reading we assumed that there were no great differences in the ways they were using reading strategies and background knowledge.

REFERENCES

Anderson, R.C. and Freebody, P. (1981) Vocabulary knowledge. *Comprehension and Teaching; Research Review*. Newark, Del.: International Reading Association.

Clarke, M. A. (1979) Reading in Spanish and English: Evidence from adult ESL students. *Language Learning*, 29, 121-150.

___ (1980) The short-circuit hypothesis of ESL reading - or

comprehension.

As mentioned before, Deville et al. (1985) and Ostyn and Godin (1985) found that lexical coverage of 95% can be achieved when the learner's lexicon reaches 5000 words. It appears therefore that the lexical threshold necessary for successful reading is 5000 words since this is the number which results in the optimal lexical coverage of an authentic text, which in turn, ensures successful comprehension.

It could be argued that the study did not measure how many of the 95% of the words understood in the text had also been familiar to the reader before he read the text and how many of them were guessed in context. Whatever the proportion might have been, it is still true that 95% of word tokens were necessary to understand the text and that this coverage can be achieved by 5000 words in the learner's lexicon. Whether 95% was known and 0% guessed, or 90% known and 5% guessed, or 80 % known and 15% guessed, the total effect of 95% lexical coverage made the significant difference in comprehension scores. Therefore the argument that lexical guessing takes place cannot invalidate the importance of lexical core of 5000 words to be learnt. It is precisely this solid vocabulary knowledge that will provide the reader with the necessary context for successful guessing. It was shown (Laufer and Bensoussan, 1982) that without such context, it is misinterpretation rather than guessing that often takes place.

Reports are available on the size of vocabulary of students in different countries who embark on University studies and use English for reading academic literature. The vocabulary range reported is from 500 to 3000 (See Laufer, 1987). Is it therefore surprising that very often learners prefer a bad translation of their bibliography to the almost incomprehesible original? A gap of 2000-4000 words between the amount of words they know and they should know turns reading into 'mission impossible'. It is through increasing vocabulary to 5000 and ensuring a lexical coverage of 95% that reading has a good chance of becoming 'mission surmountable'.

SUMMARY AND CONCLUSION

The study attempted to find out what percentage of text-lexis was essential for a successful reading comprehension in a foreign language. In the experiment, two sets of scores were obtained and analysed: the reading comprehension score and the lexical coverage score. Different methods of analysis led to a similar conclusion: lexical coverage of 95% - the understanding of 95% of word-tokens in a text - can ensure reasonable reading comprehension, i. e. a score of 55% and above. Lower lexical coverage is associated with unsatisfactory more often than with satisfactory comprehension. Since the 95% coverage can be achieved by learning 5000 words, it is suggested that 5000 words seems to be the lexical threshold beneath which other facilitating factors in reading comprehension may not be very effective. Since many students who begin academic studies possess much poorer vocabularies massive vocabulary expansion should be one of the major goals of any course in Language for Academic Purposes.

NOTES

1. In L1 acquisition research, the importance of vocabulary was acknowledged earlier than that. (See for example reports on the relationship between lexical knowledge and reading comprehension in Anderson and Freebody, 1981).
2. A 'word' is taken to be a lexical item, i. e. the smallest unit in the meaning system of language.
3. 'Reasonable' reading comprehension was considered a score of 55% which is the lowest passing grade in our system in Haifa University.
4. Half of the students were given one text; the other half the other. The reliability of the standardized test is .81; the reliability of the self-made one .57.
5. If a word had been unfamiliar to the learner before it was seen in the text but was considered clear in the context, the instructions to the testees was <u>not</u> to mark it as unknown.
6. Since all the testees had a fairly similar educational/ cultural background and a similar training in reading we assumed that there were no great differences in the ways they were using reading strategies and background knowledge.

REFERENCES

Anderson, R.C. and Freebody, P. (1981) Vocabulary knowledge. *Comprehension and Teaching; Research Review*. Newark, Del.: International Reading Association.

Clarke, M. A. (1979) Reading in Spanish and English: Evidence from adult ESL students. *Language Learning*, 29, 121-150.

___ (1980) The short-circuit hypothesis of ESL reading - or when language competence interfers with reading performance. *The Modern Language Journal*, 64, 203-209.

Cummins, J. (1979) Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, 19, 107-205.

Cziko, G. A. (1980) Language competence and reading strategies: a comparison of first-and-second language oral reading errors. *Language Learning*, 30, 101-114.

Deville, G., Vandecasteele, M., Ostyn, P. and Kelly, P. (1985) Measuring the F. L. learner's lexical needs. Paper presented at the 5th European LSP Symposium. Leuven: Belgium.

Giacobbe, J. and Cammarota, M.A. (1986) Learner's hypothesis for the acquisition of lexis. *Studies in Second Language Acquisition*, 8, 327-342.

Hindmarch, R. (1980) *Cambridge English Lexicon*. Cambridge: Cambridge University Press.

Kelly, P. (1985) A Dual Approach to FL Vocabulary Learning: The Conjoining of Listening Comprehension and Mnemonic Practices. Ph.D. Thesis. LA-Neuve: Université Catholique de Louvain.

Laufer, B. and Bensoussan, M. (1982) Meaning is in the eye of the beholder. *English Teaching Forum*, 20(2), 10-14.

Laufer, B. and Sim, D. D. (1985a) Taking the easy way out: non use and misuse of contextual clues in EFL reading comprehension. *English Teaching Forum*, 23(2), 7-10, 22.

—— (1985b). Measuring and explaining the threshold needed for English for Academic Purposes texts. *Foreign Language Annals*, 18, 405-413.

Laufer, B. (1987) A case for vocabulary in EAP Reading Comprehension Materials. Cornu, A.-m., Vanparijs, J. Delahaye, M. and Baten, L. (eds) *Beads or Bracelet? How do we approach LSP*. Oxford University Press.

Levenston, E. A. (1979) Second language acquisition: issues and problems. *Interlanguage Studies Bulletin*, 41, 147-160.

Ludwig, J. (1984) Vocabulary acquisition as a function of word characteristics. *The Canadian Modern Language Review*, 40, 552-562.

Meara, P. M. (1980) Vocabulary acquisition: a neglected aspect of language learning. Kinsella, V. (ed) *Language Teaching Surveys I*. Cambridge: Cambridge University Press.

Nation, I. S. P. (1983) Testing and teaching vocabulary. *Guidelines*, 5, 12-25.

Ostyn, P. and Godin, P. (1985) RALEX: An alternative approach to language teaching. *The Modern Language Journal*, 69, 346-353.

Read, J. and Nation, P. (1985) Some issues in the testing of vocabulary knowledge. Paper presented at the ACROLT

conference. Jerusalem.

Ringbom, H. (1985) The influence of Swedish on the English of Finnish learners. Ringbom, H. (ed) *Foreign Language Learning and Bilingualism*. Åbo: Åbo Akademi Press.

Takala, S. (1984) *Evaluation of Students' Knowledge of English Vocabulary in the Finnish Comprehensive School*. Reports from the Institute for Educational Research, University of Jyväskylä.

—— (1985) Estimating students' vocabulary sizes in foreign language teaching. Kohonen, V., von Essen, H. and Klein-Braley (eds) *Practice and Problems in Language Testing*. AFinla Series of Publications No. 40.

Ulijn, J.M. (1984) Reading for professional purposes: psycholinguistic evidence in a cross-linguistic perspective. Pugh, A. K. and Ulijn, J. M. (eds) *Reading for Professional Purposes*. London: Heinemann.