Defining a Minimal Receptive Second-Language Vocabulary for Non-native University Students: An Empirical Investigation

SUZANNE HAZENBERG and JAN H. HULSTIJN

Vrije Universiteit Amsterdam

This study aimed to answer the question of how many words of the Dutch language, and which words, an adult non-native speaker needs to know receptively in order to be able to understand first-year university reading materials. In the first part of this study, an assessment was made of the representativeness of a list of 23,550 words (lemmas), taken from a school dictionary, for a 42 million-word token corpus of contemporary written Dutch. It was found that, using frequency as a criterion, text coverage substantially increased with up to 11,123 words (i.e. words occurring more than 100 times in the corpus), but not beyond. In the second part of the study, an assessment was made of the representativeness of the same list of 23,550 words for a relatively small corpus of first-year university reading materials. The percentage of tokens covered in this small academic corpus did not differ substantially from the percentage of tokens covered in the big corpus analysed in the first part. The third part of the study consisted of the development and administration of a 140-item multiple-choice vocabulary test aimed at measuring test takers' receptive knowledge of 18,615 content words of the 23,550 word list. This test was administered to (i) native speakers entering university as freshmen, (ii) non-native graduate students, and (iii) non-native prospective students taking a Dutch language entry examination test battery. Extrapolations of the test scores showed that the average vocabulary size of these three groups of test takers was 18,800, 15,800, and 11,200 respectively. It is concluded that the minimal vocabulary size needed for university studies is 10,000 base words. Earlier Dutch studies, suggesting that knowledge of 3,000 or 5,000 base words would suffice, appear to have underestimated such a minimal vocabulary.

1 INTRODUCTION

Applied linguists, working in the field of language learning and pedagogy, often face practical questions which, for a number of theoretical reasons, cannot be easily answered, or indeed cannot be given a principled answer at all. One such question is: How many words (and which words) does one need to know in order to be able to understand first-year university reading materials? This is a highly relevant question for those who prepare themselves for entry in a university where a language other than their first language is spoken (in another country, another speech community) as well as for educators responsible for

language courses set up to cater for the needs of such non-native prospective students

There are a number of reasons why a principled answer to this practical question is unlikely ever to be attained First of all, there are problems concerning the way in which one defines 'words' in linguistic terms (Nagy and Anderson 1984, Carter 1987, Goulden, Nation, and Read 1990, Vermeer 1992, Beheydt 1993, Bauer and Nation 1994) Furthermore, one needs a conceptualization of 'word knowledge' as a psycholinguistic notion (Anderson and Freebody 1985, Meara 1990, 1992, Nation 1990, De Bot and Schreuder 1993, Read 1993, Verhallen and Schoonen 1993, Scherfer 1994) In addition, it is difficult to develop testing procedures for assessing individuals' vocabulary size in a valid and reliable way (Sims 1929, Melka Teichroew 1982, Meara 1987, Wesche and Paribakht, in press) Finally, it is difficult to give a satisfactory definition of what it means to comprehend a text and the role that vocabulary knowledge plays in this process (Dollerup, Glahn, and Rosenberg Hansen 1989, Bossers 1991, Coady 1993) Small wonder, therefore, that there has been much controversy concerning the question of how many words native speakers in certain age groups know (Just and Carpenter 1987 103ff , D'Anna, Zechmeister, and Hall 1991), and how many they need to know in order to comprehend a textbook used in university freshman courses

Goulden *et al* (1990) and Nation (1993a), however, have demonstrated that, for the practical purpose of assessing how many words someone knows and how many words should be learnt or taught in foreign and second language education, most of the above problems can be satisfactorily solved On the basis of a carefully chosen procedure of sampling headwords from a large contemporary dictionary, Goulden *et al* estimated that 'well-educated adult native speakers of English have a receptive vocabulary of around 17,000 base words' (1990 341) Similar estimates were obtained in studies by D'Anna, Zechmeister, and Hall (1991) and Nusbaum, Pisoni, and Davis (1984) Zechmeister, D'Anna, Hall, Paus, and Smith (1993 203) conclude from these three studies that there is 'converging evidence that the vocabulary size of a university undergraduate is in the range of 14,000–17,000 words', proper names, abbreviations, and compound words not included

On the basis of these figures, one might argue that receptive knowledge of 14,000 headwords (including the 800 words of the University word list compiled by Xue and Nation 1984) should be the desired learning goal for non-native speakers preparing for university entry This leaves unanswered, however, the question of whether non-native speakers could successfully embark on a university study with less than 14,000 words Fourteen thousand, in other words, may be an optimal figure, but is it also a minimal figure? Much lower figures appear in the literature on 'text coverage' (the percentage of word tokens in a running text covered by certain numbers of word types or lexical entries). It has been claimed for various languages that the 5,000 most frequent words (or 3,000 word 'families') yield a coverage of 90 per cent to 95 per cent of the word tokens in an average text This has been claimed, for instance, for

Russian (Steinfeldt 1965), French (Guiraud 1954, Sciarone 1979), English (Palmer 1931, Bongers 1947, Carroll, Davies, and Richman 1971, Johnson 1972, Hirsch and Nation 1992, Nation 1993b), and Dutch (Vannes 1952, Sciarone 1979, Ostyn and Godin 1985, Nieuwborg 1992) [1] Furthermore, it is assumed that in order to reach text comprehension (e g understanding of all main points), readers need to be familiar with 95 per cent of the words in a text (Hirsh and Nation 1992) If these figures could be shown to be valid and reliable, non-native speakers might be sufficiently equipped for university entry with a vocabulary of only 5,000 words [2]

Obviously, the gap between 14,000 and 5,000 is not a trivial one It implies great differences in learning load and in provision of educational and instructional facilities This discrepancy led us to adopt, in our study, an approach which combines a vocabulary testing method (as in the studies by Goulden *et al*) with a method of counting word token coverage in texts (as in the studies by Giraud, and others) The difference between our method and most of the Giraud-type studies is that we had word frequency data at our disposal based on a much larger text corpus than the authors of these studies had We were therefore in a position to make more reliable claims concerning the number of words needed to attain 90 per cent or 95 per cent text coverage

We conducted our investigation with Dutch as the language of study and with native and (prospective) non-native students at Dutch universities as our subjects [3] Although no two languages are alike, although the reading requirements for students vary between countries and educational settings, and although, therefore, our findings cannot be extrapolated to other languages or educational settings, we believe that both the method and the findings of our research should be of interest to researchers of similar (non-agglutinative) languages (e g English) and similar academic contexts In summary, our research aimed to answer the following question How many base words (a term to be defined below) of the Dutch language, and which words, does an adult non-native speaker need to know receptively in order to be able to understand first-year university reading materials? After having selected 23,550 headwords from an existing dictionary, deemed useful for secondary school students, as our basic material, we adopted the following three-pronged approach to answer our research question Firstly, we assessed the representativeness of these 23,550 words by comparing them with a word frequency count conducted on a 42-million-word token corpus of contemporary written Dutch Secondly, we also assessed their representativeness calculating their lemma coverage of a sample of first-year university texts Thirdly, we developed a 140-item multiple-choice vocabulary test aimed at measuring test takers' receptive knowledge of 18,615 content words of our 23,550 word list, divided into four frequency classes This test was administered to three groups of test takers native speakers entering university as freshmen, non-native graduate students, and non-native prospective students taking the Dutch language entry examination test battery For the third group of test takers, we compared their scores on our vocabulary test with their scores on the reading comprehension

subtest of the entry examination This allowed us to distinguish between vocabulary scores of test takers who had failed and of test takers who had passed the reading comprehension test [4]

In the following three sections of this article, we deal with these three main parts of our investigation in succession Each part ends with a discussion and a (preliminary) conclusion In the final section, we attempt to answer our original question and discuss our findings in the light of the discrepancy found in the above-mentioned literature

2 ASSESSING THE TOKEN COVERAGE OF A 23,550 WORD LIST

The aim of part one of our study was to assess the representativeness of a list of 23,550 base words, selected for educational purposes, by determining the extent to which these words, and subsets of these words, 'covered' the 42-million-word tokens in a corpus of contemporary written Dutch

2 1 *Preliminary selection of a basic word list*

We started our investigation with an existing monolingual Dutch dictionary, the *Basiswoordenboek Nederlands* (Basic Dictionary of Dutch, Huijgen and Verburg 1987) This dictionary consists of about 24,400 entries, selected from the large, unabridged Van Dale dictionary of contemporary Dutch, it is intended for use in primary and secondary schools Not included in this base dictionary are (a) words which are specific for professional domains which secondary school students are not likely to encounter during their school readings, (b) transparent derivations and transparent compounds (e g *huisdeur*, 'house door', since its meaning can be derived from *huis*, 'house', and *deur*, 'door'), (c) infrequent foreign words, (d) archaic words, (e) dialect words, and (f) vulgar words The authors included many infrequent words deemed useful for the comprehension of reading materials in a wide variety of school subjects Since transparent derivations and compounds are excluded from this dictionary, its lemmas come close to what Bauer and Nation (1994) have called 'word families' According to these authors, 'a word family consists of a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately' (Bauer and Nation 1994 355) Nation (1993b) and Laufer (1992b) have pointed out that it is important to define a necessary L2 vocabulary in terms of true 'base' words, or 'word families' Thus, we started our investigation using a 24,400 base word list selected, by others, on educational principles, commensurate with the purposes of our study

2 2 *Frequency and coverage calculations*

After the deletion of some thousand abbreviations, cross references, and words with double spellings, the validity of the remaining 23,450 base words as potential learning objectives for non-native prospective university students was assessed in three steps The first step involved determining for each base word its frequency of occurrence in contemporary written Dutch The frequency

count was conducted by the Centre of Lexical Information (CELEX) in Nijmegen For every word in Van Dale's *Basiswoordenboek*, CELEX provided us with the number of times this word was found in the so-called INL corpus, a 42-million-word-token corpus of contemporary written Dutch (930 books on a wide variety of fictional and non-fictional topics, published between 1970 and 1988), compiled by the Institute of Netherlandic Lexicology (INL) at Leyden University We then added 90 words from the INL corpus which happened to be absent from Van Dale's *Basiswoordenboek* [5] This eventually resulted in a list of 23,550 base words, to which we will, from now on, refer as the 'H&H list' (Hazenberg and Hulstijn list) It turned out that the H&H list reached a cumulative coverage of almost 90 per cent of this corpus, as shown in the bottom line of Table 1

Table 1 The 23,550 lemmas from the H&H list in relation to the INL-corpus in terms of frequency and coverage

Frequency		Number of words H&H list	Cumulative corpus coverage	
			Absolute	percentage
>=	1,000,000	4	5,680,893	13 4
>=	100,000	54	19,311,927	45 6
>=	10,000	375	27,793,384	65 6
>=	5000	710	30,151,172	71 1
>=	2500	1359	32,416,028	76 5
>=	1500	2154	33,947,486	80 1
>=	1000	2918	34,879,117	82 3
>=	750	3628	35,497,179	83 8
>=	500	4731	36,175,268	85 4
>=	375	5674	36,584,602	86 3
>=	250	7073	37,015,273	87 3
>=	100	11,123	37,675,077	88 9
>=	75	12,470	37,792,284	89 17
>=	50	14,358	37,910,027	89 45
>=	25	17,224	38,016,418	89 70
>=	10	19,852	38,062,334	89 81
>=	5	21,053	38,071,745	89 83
Total		23,550	38,075,794	89 84

From Table 1, it can be seen that the four most frequent base words, that is to say those that have a frequency of more than a million occurrences, cover 13 4 per cent of the corpus, that the 54 most frequent words, that is to say those that have a frequency of over one hundred thousand occurrences, cover 45 6 per cent of the corpus, and so on Table 1 shows that there is a substantial gain in the percentage of text covered up to 11,123 base words, whereas from 11,123 onwards to 23,550 base words the gain is less than 1 per cent

2 3 *Discussion and preliminary conclusion*

The main result of this part of our study is that word frequency yields substantial gains in text coverage up to 11,123 words, but not beyond (Table 1) When interpreting the data of Table 1, one has to bear in mind that 6 per cent of the word tokens in the INL corpus consist of proper names and hapax legomena (nonce words) Thus, it would have been virtually impossible for the H&H list to have reached a coverage greater than 94 per cent Furthermore, as said in the previous section, transparent compound words and transparent derivations are excluded from the H&H list We therefore estimate that L2 learners familiar with the 11,123 most frequent base words of that list would actually know one or two per cent more than 88 9 per cent of the word tokens of an average written text. Should they also be familiar with most of the proper names occurring in a text, then they might come very close to knowing 95 per cent, a percentage generally held desirable for a reasonable level of text comprehension That is, with a 95 per cent word knowledge, readers would be able to understand most of the text and would have to look up only a few words in a dictonary, not losing too much time in reaching their reading goal in comparison to readers familiar with all the words occurring in the text

One might wonder whether 11,123 must be considered a minimal or an optimal figure We acknowledge that ours is a 'safe' interpretation Obviously, a larger vocabulary knowledge is more likely to come close to the desired level of text coverage than a smaller one Yet, our reading of the figures of Table 1 is that this table does not allow the conclusion that knowledge of 3,000 or 5,000 base words would be enough to reach a 95 per cent text coverage The fact that our study yields lower coverage figures at the 3,000 and 5,000 word levels than other empirical Dutch studies (Vannes 1952, Ostyn and Godin 1985) must be explained, we believe, by the differences in the size of the corpora on which frequency and coverage calculations were based Whereas in these two other studies only relatively small text samples were used (mainly due to the fact that the capacity of the electronic hardware and software the authors of these studies had at their disposal was much more limited than the databases and programs used by CELEX for us), in our study, a large corpus was available, boosting the reliability of the frequency and coverage figures considerably [6]

Our preliminary conclusion, then, is that, in the first part of this study, substantial coverage gains were obtained up to the level of 11,123 base words, but not beyond, and that this 11,123 level appears to be sufficient for reaching a familiarity with 95 per cent of the word tokens in an average text

3 A COMPARISON OF THE H&H WORD LIST WITH UNIVERSITY READING MATERIALS

Our aim in the second part of our study was to determine, in a small-scale investigation, how the frequency figures obtained from the first part of our study, yielding coverage figures of a broad range of texts, would relate to the words contained in a sample of more narrowly defined texts, namely texts which

university freshmen must be able to understand This time, we wanted to compute lemma coverage rather than token coverage Since lemmatization of texts can be done only in part by machines and must be done in part by hand, only a small sample corpus was analysed [7]

The composition of the corpus consisted of eight samples, taken from (1) the national high school reading exams of 1989 and 1990 (7,302 tokens, 66 proper names and foreign words, 1,681 remaining lemmas), (2) and (3) the reading exams 1991 and 1992 for non-native speakers who seek entry to Dutch universities (6,719 and 5,733 tokens, 99 and 147 names/foreign words, 2,065 and 1,714 remaining lemmas, respectively), (4) texts about student unions, student fellowships, and student medical services 1990/1991 (1,826 tokens, 7 names/foreign words, 552 remaining lemmas), (5) a first-year anthropology textbook (994 tokens, 60 names/foreign words, 403 remaining lemmas), (6) two first-year economics textbooks (1,014 tokens, 14 names/foreign words, 376 remaining lemmas), (7) three first-year computer science books (1,085 tokens, 19 names/foreign words, 382 remaining lemmas), (8) a first-year medicine textbook (1,104 tokens, 14 names/foreign words, 426 remaining lemmas) [8]

Table 2 Lemma coverage in eight samples of first-year university texts

	Range	Mean
Lemmas*	376–2,065	950
Lemma coverage (%)		
F > 500 (4,731 H&H lemmas)	62–74	68
F > 100 (11,123 H&H lemmas)	71–82	78
F > 25 (17,224 H&H lemmas)	74–84	81
All 23,550 H&H lemmas	77–87	84

* Proper names and foreign words excluded (average 53 per sample)

Table 2 shows how the four most important frequency classes of the H&H list (frequency in terms of the INL corpus, as shown in Table 1) relate to the lemmas contained in this corpus In addition, we calculated token coverages, allowing a comparison with part one of our study Our first-year university corpus consisted of 25,777 tokens The entire H&H list reached an average token coverage of 89 7 per cent of this 25,777-token corpus, virtually the same figure as the one obtained for the 42-million-token INL corpus (89 84 per cent) in part one of our study Lowest coverage figures were always for the medicine text, highest figures for the anthropology sample (at the 4,731 and 11,123 levels) and the computer science sample (at the 17,224 and the entire-list levels)

3 1 *Discussion and conclusion*

The percentage of tokens in this small 'freshman' corpus covered by the entire H&H list did not differ substantially from the percentage of tokens covered in

the big INL corpus (89 7 and 89 8 per cent respectively) We therefore assume that the figures computed for the big INL corpus, given in Table 1, can be taken to offer reliable predictions for first-year university reading materials too Thus, the second part of our study does not suggest that the conclusions drawn from the first part are based on over- or underestimations as far as first-year university texts are concerned

We believe that lemma coverage gives a more realistic idea of how difficult a text is than token coverage (if the text sample is not too small) This can be demonstrated with the following example While the average number of tokens in the eight samples of our academic corpus was 3,222, the average number of lemmas was 1,003 Let us grant the non-native reader of the average sample in the academic corpus with familiarity of the proper names and foreign words, then knowledge of the 4,731 most frequent entries of the H&H list (F > 500) leaves the reader with 304 lemmas unknown (304 is 32 per cent of 950) Note that 4,731 is close to the 5,000 figure often mentioned in the literature as sufficient for reasonable text comprehension (see the introduction of this paper) Even if we assume that some of these 304 lemmas are transparent compounds or derivations of the 4,731 most frequent base words (say 10 per cent), the reader is confronted with 274 (304 − 30) unknown words that is, with 41 different unknown words per page of printed text (Assuming that there are 480 words on an average page, a 3,222-token text needs 6 7 pages Thus, 274 unknown words for the entire text equals an average of 41 unknown words per page) However, if we grant the reader with knowledge not of 4,731 entries but of the 11,123 most frequent entries of the H&H list (F > 100), the number of uncovered lemmas reduces to 209 (22 per cent of 950), and if we allow for 15 per cent reduction due to transparent derivations and compounds (5 per cent more than the 4,731 level because at the 11,123 level the reader knows more base words and hence can understand more transparent derivations and compounds), the reader has to deal with only 178 (209 − 31) unknown words in a 6 7-page text, that is with 27 unknown words per page [9] We believe that 27 unknown words per page (roughly equalling a 95 per cent token coverage) may still impede fluent reading and comprehension, although we acknowledge that not all words need to be familiar to the reader for the text to be comprehensible (Particularly in college textbooks, new words, representing new notions, are presented and explained Thus, even native speakers need not be familiar with all the words in such a text) We therefore conclude that the second part of our study does not allow us to reach any other conclusion than the one we reached at the end of the first part knowledge of ten or eleven thousand base words is the bottom line for reading at the college level

4 TESTING VOCABULARY KNOWLEDGE

Finally, in the third part of the project, we composed and administered a multiple-choice vocabulary test A first reason for doing this was to establish how many of the words contained in the H&H list are known by native and non-native university students as well as by non-native prospective students and, for

the third group of subjects, to compare their vocabulary knowledge to their scores on the reading test of the Dutch language university entry examination The second reason was to determine the extent to which word frequency can be used to predict word knowledge One might expect that the most frequent words are known by all students, whereas more infrequent words are known only by a subgroup, the composition of which varies with such things as hobbies, work, and experiences (Anderson and Freebody 1985)

4 1 *Construction and administration of the vocabulary test*

As mentioned in the introduction, the issue of what the optimal method is to assess individuals' vocabulary size is as yet far from settled Reliable vocabulary-size tests must consist of many items, this, in turn, calls for a non-time-consuming administration procedure We chose a multiple-choice format, thus somewhat increasing the chance that test takers would recall the meaning of a word whose meaning they might not have recalled when encountering it in an ordinary text

The test words were selected in the following way All non-native test takers were assumed to be familiar with the 2,000 most frequent words of the language since they had been enrolled in at least intermediate-level Dutch courses We therefore did not select target words from these 2,000 words There were 18,615 nouns, verbs, and adjectives in the remaining 21,550 words (23,550 – 2,000) These 18,615 content words were rank ordered according to their frequency of occurrence Every 132nd word was then selected for the test, 140 words in total (18,615 – 140 = 132)

For each target word, a short carrier sentence was constructed providing no contextual clues as to its meaning Students were asked to choose the meaning of the target word from four options A fifth option was 'I really don't know' The tested target words did not belong to the 2,000 most frequent words, whereas the words used in the carrier sentences and in the four circumscriptions did Thus, the instrument assessed test takers' receptive knowledge of the 23,550 words of the H&H list, except for the 2,000 most frequent of these words, and except for 2,935 words which were *not* nouns, verbs, or adjectives By way of illustration, an English example is given, testing knowledge of the target word 'celebrity' (in reality, however, all materials were in Dutch)

I met a *celebrity* last night

1 someone who operates on the brain
2 someone who never got married
3 someone who is very famous
4 someone who makes clothes
5 I really don't know

As a pre-test, the test was administered to 59 non-native prospective students of the University of Utrecht The test proved to be reliable with regard to the intended population The Kuder Richardson 20 reliability measure was 93 After a few adjustments based on an item analysis, the final version of the test

was administered to 137 non-native prospective university students taking the Dutch language university entry examination at the Free University of Amsterdam and at the University of Amsterdam in May 1992 (Both universities administered the same exam) Our vocabulary test was administered a few weeks before subjects sat the university entry examination For purposes of comparison, our vocabulary test was also administered to 41 non-native graduate students who had successfully finished their first year at the Free University, as well as to 28 Dutch first-year students of the same institution (natives) in the summer of 1992

4 2 *Vocabulary knowledge of native and non-native test takers*

Table 3 shows the performance of the three groups of test takers It can be seen that, as might be expected, non-native prospective students attain the lowest scores on the test, while Dutch students attain the highest scores, with a difference in average score of over 57 points The average score of the non-native graduate students lies in between the scores of the other two groups This seems to indicate that non-native students are capable of considerably extending their L2 vocabulary within a couple of years

Table 3 Correct scores attained in the vocabulary test (maximum score = 140)

	Dutch students	Non-native graduate students	Non-native prospective students
N	28	41	137
Mean	126 4	103 8	69 2
SD	5 1	17 1	17 0

A first point worth noticing is that the Dutch students did not know all the words in the test An item analysis showed that in none of the questions was there a preference for one particular distractor, nor were there any questions that were incorrectly answered by a large number of students Thus, the reason that certain questions were incorrectly answered by the Dutch students had nothing to do with the quality of the test (for instance, with the way in which the distractors were selected), it simply resulted from a lack of knowledge of these words with some of these students

A Kruskal-Wallis analysis of variance (the non-parametric equivalent of the one-way ANOVA, see Siegel and Castellan 1988) revealed a significant overall effect of Group of subjects on test scores ($X^2 = 119{,}13$, $df = 2$, $p < 000$) *Post-hoc* calculations showed that all three group averages differed significantly from each other ($p < 05$).

If we proportionally convert the average number of correct items to the 18,615 words from which the test words had been sampled, adding 2,000 for

assumed knowledge of the 2,000 most frequent words, we can make a rough estimate of the number of words in the H&H list known by the subjects (see Table 4) [10] In view of the enormous individual differences in word knowledge, in particular with the non-native subjects (evidenced by the high standard deviations in Table 3), the numbers in Table 4 should be interpreted with caution Apart from variation in word knowledge, we also have to take into consideration variations in answering strategies In spite of the request to use answer alternative 5, 'I really don't know', if they did not know the word in question, some of the subjects still guessed, seldom using alternative 5 The numbers in Table 4, therefore, give only a rough indication of the size of the subjects' vocabulary

*Table 4 Extrapolation of test scores to size of basic vocabulary**

	Dutch students	Non-native graduate students	Non-native prospective students
X	18,807	15,802	11,201

*The maximum vocabulary size is 20,615 (i e the subset of 18,615 nouns, verbs, and adjectives from which the test words had been sampled, plus the 2,000 most frequent words, with which test takers were assumed to be familiar)

4 3 *The relation between word knowledge and word frequency*

In order to test the assumption by Anderson and Freebody (1985) that the number of words known per frequency class decreases proportionately with a decrease in word frequency, the 140 items of the test were divided into four frequency classes

Class 1 20 items, frequency > = 500
Class 2 46 items, frequency <500 and > = 100
Class 3 44 items, frequency <100 and > =25
Class 4 30 items, frequency <25

Table 5 shows the average *p* value (the percentage of items correctly answered) per group of subjects and per frequency class If word frequency can indeed be used as an indicator of word knowledge, the average percentage of classes 1 through 4 ought to decrease Table 5 shows that this is indeed true for the non-native prospective students, but with the other groups of subjects the percentage in class 3 exceeds that of class 4 For each group of subjects, a Kruskal-Wallis analysis of variance was carried out, with the *p* values per group of subjects as dependent variable and the division into four frequency classes as factor The overall effect of the frequency classes turned out to be significant ($p < 001$) with

Table 5 Average p *values (percentage of items correctly answered) and standard deviations (in parentheses) per group of subjects and per frequency class (1, 2, 3, 4)*

Frequency class	Dutch students	Non-native graduate students	Non-native prospective students
1	94 4	92 4	70 8
	(10 5)	(10 3)	(17 8)
2	93 2	82 3	53 4
	(14 9)	(19 2)	(22 3)
3	84 9	62 4	42 1
	(21 8)	(26 9)	(21 7)
4	90 8	65 8	39 9
	(13 2)	(27 5)	(21 8)

both groups of non-natives, but not with the Dutch students Our next move was to investigate whether the four frequency classes differed significantly among each other With the Dutch students, this turned out not to be the case With the non-native graduate students, however, the classes differed significantly from each other ($p < .05$), with two exceptions. frequency classes 1 and 2, and frequency classes 3 and 4 With the non-native prospective students, frequency class 1 turned out to differ significantly ($p < 05$) from classes 2, 3, and 4, while the other differences proved not to be significant

From Table 5 it can also be seen that the standard deviation increases when the average p value per frequency class decreases Thus, the smaller the percentage of words known, the larger the variance in word knowledge between the subjects

In summary, the relationship between word frequency and word knowledge appears to depend on vocabulary size When individuals have a relatively large vocabulary (as in the case of the native speakers), there is no significant relationship, due to a ceiling effect When individuals have a relatively small vocabulary, as in the case of the non-native speakers in comparison to to the native speakers, there is a significant relationship However, if the total size of the vocabulary is relatively very small, as with the non-native prospective students in comparison with the non-native graduate students, the effect of frequency decreases again Note that the difference between the non-native prospective students and the other two groups are apparent in all four frequency classes, even in class 1, whereas the disadvantage of the non-native graduate students in comparison to the native speakers does not become apparent until frequency class 2 This underlines the importance of the 5,000 most frequent words as a learning

objective for any non-native student, irrespective of his or her field of study (see also section 5 below)

4 4 *The relationship between vocabulary knowledge and reading comprehension*

As stated in the introduction, the non-native prospective students also took the Dutch language entry examination for non-native speakers seeking entry to a Dutch university This exam consisted of separate tests for listening, reading, speaking, and writing The reading part consisted of a test of 'close' reading (15 multiple-choice questions on three texts on general academic topics) and a test of extensive reading (11 skimming and scanning items) The first part lasted 75 minutes, the second part lasted 40 minutes Dictionary use was permitted [11] The maximum score was 47 points, the pass/fail cut-off score was 33 The reliability of this reading test, as assessed on the basis of the performance of a sample of 723 candidates (α), was 86 (Evers 1992) Of the 137 test takers who took our vocabulary test, 97 passed the reading test and 40 failed it Table 6 shows the performance of the two subgroups Vocabulary scores of those who passed were significantly higher than of those who failed the reading test ($p < 000$) This was true not only for test takers' scores on all 140 items together, but also for their scores on each of the four frequency classes separately All test takers with a vocabulary score of 77 or higher had passed the reading test The Pearson correlation between the reading comprehension and vocabulary scores was 63 ($p < 001$)

Table 6 Reading comprehension and vocabulary scores of test takers who passed and failed the reading comprehension test

	Reading comprehension Max = 47	Vocabulary score Max = 140
Passed ($N = 97$)	39 1	73 8
Failed ($N = 40$)	27 0	58 0

A proportional conversion of the mean correct scores on the vocabulary test (maximum is again 20,615) yields an estimated mean vocabulary size of 11,813 entries of the test takers who passed the reading part of the entry exam and of 9,712 entries of those who failed the reading test The vocabulary score of 77, beyond which nobody failed the reading test, equals a vocabulary size of 12,282 entries

4 5 *Discussion and conclusion*

The mean vocabulary size of all those who failed the reading test was actually surprisingly high (9,712 entries) This reminds us of the fact that there is much more to reading comprehension than just vocabulary knowledge (Vollmer and

Sang 1983, Bossers 1991, Lumley 1993) Correlations between L2 reading comprehension and L2 vocabulary tests are generally moderate (e g Oller 1983, Upshur and Homburg 1983, Vollmer and Sang 1983), and the figure of 63 which we obtained is no exception Yet, nobody would deny that vocabulary knowledge is essential for reading comprehension The mean vocabulary size of the test takers in our study who passed the reading test is 11,813 entries of the H&H list As in the discussion of the results of part one of our study, we may ask whether the mean vocabulary size of test takers who passed the reading test ought to be seen as an optimal or as a minimal vocabulary needed to enter university Acknowledging that many test takers passed the reading test with vocabularies smaller than the mean of 11,813, we suggest taking a vocabulary score one standard deviation (16 5) below the mean (73 7), i e a vocabulary score of 57, as a lower boundary This score equals a vocabulary of 9,579 entries We therefore conclude that the data of this third part of our investigation suggest that individuals with a vocabulary of fewer than ten thousand base words run a serious risk of not attaining the reading comprehension level required for entering university studies

5 CONCLUSIONS

In this study, we aimed to answer the question of how many base words of the Dutch language (head words as they appear in a school dictionary, *not* including transparent derivations and compounds) and which base words, adult non-native speakers need to know receptively in order to be able to understand first-year reading materials Given the fact that, for reasons stated at the outset of this paper, a truly principled answer to this question is not possible, interpreting the results of our study, as we have done at the end of the three previous sections, is not a simple, straightforward affair We adopted a three-pronged approach, combining dictionary-based methods of vocabulary testing with text-based methods of calculating both token and lemma coverage figures The evidence obtained in all three parts of the study, taken together, points in the direction of a required word knowledge of at least ten thousand base words, rather than three or five thousand, as suggested by some authors (Vannes 1952, Ostyn and Godin 1985)

So far, we have only addressed the question of *how many* words are needed for entry into university Our research question, however, also asked *which* words are needed A quick look at the figures of Table 1 would lead one to conclude that the words needed should be selected on the basis of frequency The results of the third part of our study, however, showed that even the non-native graduate students, i e the non-native students who had demonstrated themselves to be successful, did not know all of the 11,123 most frequent words (Table 5) Indeed, as was shown in section 4 3, the relationship between word frequency and word knowledge is not a straightforward one, it depends on proficiency level and frequency range It would therefore be premature to conclude that L2 learners preparing for university studies should learn at least the ten thousand most frequent base words of the language

We believe that vocabulary selection for L2 instruction is, and should remain, ultimately, a subjective affair However, subjective decisions should be made on the basis of the best evidence available If reliable frequency data are available, syllabus designers (especially those who design computer software for vocabulary learning) are well advised, for every one of the ten thousand most frequent base words of the language, to consider whether it should be included or not We have little doubt that the decision will almost invariably be positive in the case of the 5,000 most frequent base words, but not always so in the case of the next 5,000

The decision should thus not be automatic Syllabus designers might first submit the list to a small group of judges whom they consider, for one reason or another, to be 'experts' Only if a majority of these judges are in favour of including, should a word be included Thus, the selection process should start with words obtained through the use of *objective* criteria such as frequency (preferably computed from a representative, 'valid', corpus), range, availability, familiarity Subsequently, the words thus preselected should be filtered through experts' *intersubjectivity*, taking into account language needs and learning burden (Richards 1970, Nation 1990 21) Our point is that it makes more sense to request experts to pass their judgement on words that have been preselected on the basis of objective and reliable frequency data, than just on all the entries of an entire dictionary (e g a 'modest' dictionary of 50,000 entries)

If vocabulary learning materials are made available via computer-aided programs it would be desirable to incorporate a device into the program which enables the learner to select from a large list of words a subset to be learnt and rehearsed Thus, the decision of whether a word should be learnt or not might well be left to the learner Now that the field of second and foreign language teaching and learning is entering an age of learner autonomy (Brown 1994 80), such decisions could perhaps be better made by (advanced) learners themselves rather than by syllabus designers or teachers However, in order to make wise decisions, learners need to gain some insight in matters such as (a) how many words (function words, content words, derivations, compounds, proper names, etc) there are in a language,[12] (b) what role can be played by criteria such as word frequency, range, familiarity, and experts' judgements (Nation 1990 21), and (c) how many words they need to know receptively for fluent reading at university level We trust that our study has provided some evidence for the claim that the answer to the last question should be that 10,000 constitutes a lower boundary

(Revised version received June 1995)

ACKNOWLEDGEMENTS

We would like to thank Carolien Schouten-van Parreren, Anne-Mieke Janssen-van Dieten, Bart Bossers, Batia Laufer, and Paul Nation for their thoughtful comments on earlier versions of this paper Correspondence concerning this article should be addressed to the second author Applied Linguistics Department, Vrije Universiteit, 1105 De Boelelaan, 1081 HV Amsterdam, The Netherlands E-mail hulstyn@let vu nl

NOTES

[1] Hazenberg (1993) reports on a literature search on 16 text coverage studies She discovered that most authors who claim that 3,000 to 5,000 words cover 90 per cent to 95 per cent of an average text, refer to studies whose empirical status is somewhat doubtful in that their authors often do not clearly show how they arrived at their figures

[2] Laufer (1987, 1992a and 1992b) has argued that 5,000 words is the 'bottom line' for reading English at an academic level

[3] Most reading materials for first-year university students in The Netherlands are in Dutch and virtually all freshmen courses are taught in Dutch Hence, for non-native students it is essential to have a large vocabulary

[4] In the present article, we can only give the main results of this study We refer the interested reader to Hazenberg (1994) for detailed analyses

[5] These were moderately frequent words with frequences of between 500 and 1,753 occurrences in the INL corpus We consider these 90 items as borderline cases for which it is difficult to assess whether they should have been included or excluded in Van Dale's *Basiswoordenboek* according to the authors' criteria mentioned in section 2 1 To be on the safe side, we decided to include them in our list

[6] We can only speculate why studies on English appear to show higher coverage percentages than our Dutch study for the same number of base words One reason might be that in English most compound nouns are spelt as different words, i e with a space in between (e g *North Pole*, *night club*), whereas in Dutch most compound nouns are spelled without a space (e g *noordpool*, *nachtclub*) Thus, a computer program comparing base words with lemmatized text tokens, is more likely to overlook (opaque) compound words and therefore to overestimate text coverage in the case of English than in the case of Dutch Furthermore, it might be that English has more (frequent, regular) affixes than Dutch and thus that the number of base words in English is smaller (with a higher text coverage) than the number of Dutch base words We emphasize the speculative nature of these points, however

[7] We would like to stress that we did not use these texts to find additional words for our vocabulary list, but to test the representativeness of the H&H list for these texts

[8] The first three samples mainly contained popular science texts from magazines and newspapers Anthropology, economics, computer science, and medicine are the most preferred academic studies among non-native students at our university

[9] This observation nicely illustrates the tendency of 'diminishing returns', as an anonymous reviewer of this article rightly points out If learners double their vocabulary from 5,000 to 11,000, they only gain from 41 to 27 unknown words (lemmas) per page rather than dropping to a half of 41 This illustrates that the progression in text coverage slows down with the growth of vocabulary knowledge

[10] Let it be clear that the test measured knowledge of words in the H&H list The numbers of Table 4 should therefore not be seen as representative of test takers' *total* receptive vocabulary

[11] For our study, which aimed at comparing test takers' vocabulary knowledge with their reading proficiency, it might have been better if dictionary use had not been allowed Research, however, has shown that availability of dictionaries during a reading comprehension test does not significantly affect students' test scores (Nesi and Meara 1991)

[12] The study of Zechmeister *et al* (1993) has shown that many people, even professionals, have poor conceptions of these matters

REFERENCES

Anderson, R. C. and **P Freebody.** 1985 'Vocabulary knowledge' in H Singer and R B Ruddel (eds) 1985 *Theoretical Models and Processes of Reading* (3rd edition) Newark, Delaware International Reading Association

Bauer, L. and **P Nation.** 1994 'Word families' *International Journal of Lexicography* 6/4 253–79

Beheydt, L 1993 'Lexical memory A linguist's point of view' in J Chapelle and M Th Claes (eds) 1993 *Proceedings of the 1st International Congress on Memory and Memorization in Acquiring and Learning Languages* Louvain-la-Neuve, Belgium Centre de Langues a Louvain-la-Neuve et en-Woluwe

Bongers, H 1947 *The History and Principles of Vocabulary Control as it Affects the Teaching of Foreign Languages in General and English in Particular* PhD dissertation Utrecht University, Woerden, Netherlands Wocopi

Bossers, B 1991 'On thresholds, ceilings and short-circuits The relation between L1 reading, L2 reading and L2 knowledge' in J H Hulstijn and M F Matter (eds) 1991 *Reading in Two Languages, AILA Review* 8 45–60

Brown, H D. 1994 *Teaching by Principles An Interactive Approach to Language Pedagogy* Englewood Cliffs, NJ Prentice Hall

Carroll, J B, P Davies, and **B. Richman.** 1971 *The American Heritage Word Frequency Book* Boston Houghton Mifflin

Carter, R 1987 *Vocabulary Applied Linguistic Perspectives* London Allen & Unwin

Coady, J. 1993 'Research on ESL/EFL vocabulary acquisition Putting it in context' in Th Huckin, M Haynes, and J Coady (eds) 1993 *Second Language Reading and Vocabulary Learning* Norwood, NJ Ablex

D'Anna, C. L., E B Zechmeister, and **J. W. Hall** 1991 'Toward a meaningful definition of vocabulary size' *Journal of Reading Behaviour* 23 109–22

De Bot, K. and **R. Schreuder** 1993 'Word production and the bilingual lexicon' in R Schreuder and B Weltens (eds) 1993 *The Bilingual Lexicon* Amsterdam Benjamins

Dollerup, C., E Glahn, and **C. Rosenberg Hansen.** 1989 'Vocabularies in the reading process' in P Nation and R Carter (eds) 1989 *Vocabulary Acquisition, AILA Review* 6 21–33

Evers, R. 1992 *Verslag evaluatie examens Nederlands 1992* [Evaluation report of the 1992 Dutch exams | Technical report University of Nijmegen

Goulden, R., P Nation, and **J. Read.** 1990 'How large can a receptive vocabulary be?' *Applied Linguistics* 11/4 341–63

Guiraud, P 1954 *Les Caracteres statistiques du Vocabulaire* Paris Presses universitaires de France

Hazenberg, S. 1993 'Tekstdekking Goochelen met cijfers?' [Text coverage Juggling with figures?] *Levende Talen* 476 10–15

Hazenberg, S. 1994 *Een Keur van Woorden De Wenselijke en Feitelijke Receptieve Woordenschat van Anderstalige Studenten* [A choice of words The desired and factual receptive vocabulary of non-native students | PhD dissertation Amsterdam Vrije Universiteit

Hirsch, D and **P. Nation.** 1992 'What vocabulary size is needed to read unsimplified texts for pleasure?' *Reading in a Foreign Language* 8/2 689–96

Huijgen, M. W and **M. E Verburg.** 1987 *Van Dale Basiswoordenboek van de Nederlandse taal* Gorinchem, The Netherlands De Ruiter

Johnson, D. B. 1972 'Computer frequency control of vocabulary in language learning reading materials' *Instructional Science* 1 121–31

Just, M. A. and **P. A. Carpenter.** 1987 *The Psychology of Reading and Language Comprehension* Boston Allyn and Bacon

Laufer, B 1987 'Teaching vocabulary The lexical perspective of reading comprehension' *English Language Teaching Journal* (Israel) 35 (June) 58–67

Laufer, B 1992a 'Reading in a foreign language How does L2 lexical knowledge interact with the reader's general academic ability?' *Journal of Research in Reading* 15/2 95–103

Laufer, B 1992b 'How much lexis is necessary for reading comprehension?' in P Arnaud and H Béjoint (eds) 1992 *Vocabulary and Applied Linguistics* London Macmillan

Lumley, T. 1993 'The notion of subskills in reading comprehension tests An EAP example' *Language Testing* 10/3 211–34

Meara, P 1987 'An alternative to multiple-choice vocabulary tests' *Language Testing* 4/2 142–54

Meara, P 1990 'A note on passive vocabulary' *Second Language Research* 6/2 150–4

Meara, P. 1992 'Network structures and vocabulary acquisition in a foreign language in P J L Arnaud and H Bejoint (eds) 1992 *Vocabulary and Applied Linguistics* London Macmillan

Melka Teichroew, F J. 1982 'Receptive versus productive vocabulary A survey' *Interlanguage Studies Bulletin* 6/2 5–33

and **R C. Anderson.** 1984 'How many words are there in printed school English?' *Reading Research Quarterly* 29 304–30

Nation, I. S. P. 1990 *Teaching and Learning Vocabulary* New York Newbury House Publishers

Nation, P. 1993a 'Using dictionaries to estimate vocabulary size Essential, but rarely followed, procedures *Language Testing* 10/1 27–40

Nation, P 1993b 'Vocabulary size, growth, and use' in R Schreuder and B Weltens (eds) 1993 *The Bilingual Lexicon* Amsterdam Benjamins

Nesi, H and **P. Meara.** 1991 'How using dictionaries affects performance in multiple-choice EFL tests' *Reading in a Foreign Language* 8/1 631–43

Nieuwborg, E 1992 'Tekstdekking en tekstbegrip Een experimenteel onderzoek' [Text coverage and reading comprehension An experimental investigation] in A Halbo (ed) 1992 *Evaluation and Language Teaching Liber Amicorum Frans van Passel* Bern Peter Lang

Nusbaum, H. C., D. B Pisoni, and **C K Davis.** 1984 'Sizing up the Hoosier mental lexicon Measuring the familiarity of 20,000 words' *Research on Speech Perception, Progress Report No 10* Bloomington In Indiana University

Oller Jr, J W 1983 'Evidence for a general language proficiency factor An expectancy grammar' in J W Oller Jr (ed) 1993 *Issues in Language Testing Research* Rowley MA Newbury House

Ostyn, P. and **P. Godin.** 1985 'An alternative approach to language teaching' *The Modern Language Journal* 69 346–55

Palmer, H. E. 1931 *Second Interim Report on Vocabulary Selection* Tokyo The Institute for Research in Teaching

Read, J. 1993 'The development of a new measure of vocabulary knowledge' *Language Testing* 10/3 355–71

Richards, J C 1970 'A psycholinguistic measure of vocabulary selection' *IRAL* 8 87–102

Scherfer, P. 1994 Ueberlegungen zu einer Theorie des Vokabellernens und -lehrens' in W Borner and K Vogel (eds) 1994 *Kognitive Linguistik und Fremdsprachenerwerb* Tubingen Gunter Narr

Sciarone, A G 1979 *Woordjes Leren in het Vreemde-talenonderwijs* [Vocabulary learning in foreign-language instruction] Muiderberg, Netherlands Coutinho

Siegel, S and **N J Castellan, Jr.** 1988 *Nonparametric Statistics for the Behavioral Sciences* (second edn) New York McGraw-Hill Book Company

Sims, V M. 1929 The reliability and validity of four types of vocabulary test' *Journal of Educational Research* 20 91–6

Steinfeldt, E A 1965 *Russian Word Count* translated by V Korotky Moscow Progress Publishers

Upshur, J. A and **T J Homburg** 1983 'Some relations among language tests at successive ability levels' in J W Oller Jr (ed) 1983 *Issues in Language Testing Research* Rowley, MA Newbury House

Vannes, G. 1952 *Vocabulaire de base du Néerlandais* Antwerp, Belgium De Sikkel

Verhallen, M and **Schoonen, R** 1993 'Lexical knowledge of monolingual and bilingual children *Applied Linguistics* 14/4 344–63

Vermeer, A 1992 'Exploring the second language learner lexicon' in L Verhoeven and H A L de Jong (eds) 1992 *The Construct of Language Proficiency* Amsterdam Benjamins

Vollmer, H J. and **F. Sang** 1983 Competing hypotheses about second language ability A plea for caution in J W Oller Jr (ed) 1983 *Issues in Language Testing Research* Rowley, MA Newbury House

Wesche, M and **T S. Paribakht** (forthcoming) 'Assessing L2 vocabulary knowledge Depth versus breadth' in B Harley (ed) (forthcoming) *Studies in Vocabulary Learning*

Xue, G and **Nation, I S P** 1984 'A university word list' *Language and Communication* 3 215–29

Zechmeister, E. B., C D'Anna, J W. Hall, C. H. Paus and **J A Smith** 1993 'Metacognitive and other knowledge about the mental lexicon Do we know how many words we know?' *Applied Linguistics* 14/2 188–206