



The Lama Temple, Beijing

This article was first published in *AILA Review: Vocabulary Acquisition (Special Issue)* edited by Paul Nation and Ron Carter # 6 (1989). 21-33.

## VOCABULARIES IN THE READING PROCESS

Cay Dollerup, Ester Glahn, Carsten Rosenberg Hansen, Copenhagen, Denmark

*Using a study of Danish freshman undergraduates' vocabularies as a springboard, the paper explores and discusses a number of current assumptions about vocabularies in the mother tongue and in foreign language teaching. The conclusion is that as far as reading is concerned, a reader's vocabulary is part of the process of reading: it is a function of the texts and its contents, of the reader's reading strategies, and of the reader's more or less stable "word knowledge". In the reading of a specific text there is a constant interplay between these factors which suggest that a vocabulary in reading is "fluid". Pedagogically, this theory implies that there should be a deliberate teaching of reading strategies in addition to other methods.*

### 1. Introduction

The purpose of the present article is to call attention to a number of shortcomings in much thinking about the "size of vocabularies". It proposes that it would be sounder and more in keeping with reality to assume that vocabularies in reading are fluid and depend on the text read, on the reading strategies employed, and on the words the readers feel they know.

Vocabularies may differ in size and composition for a variety of reasons. In the following discussion we look at the effect of three factors on a learner's vocabulary size. These factors are (1) frequency, (2) experience with the language, and (3) the interaction between a reader and a text.

It is taken for granted that among native speakers of English "most people know all the very common words" (Anderson and Freebody, 1981: 101); and as very frequent words make up a large percentage of the running words in text (see Anderson and Freebody, 1981; Nation, 1983), it is no

surprise that frequency and frequency lists are taken into account in language teaching. For example, Thorndike and Lorge (1944) recommended what frequency bands the teachers should concentrate on at different grades for teaching native speakers of English. The same idea has been applied in the teaching of English as a foreign language with frequency based word lists being used in course preparation. ...// 22 ...

The very frequent words constitute a "core" (or "store") of words which all students, native speakers and foreign learners must learn and master. This basic "core" serves as a stepping stone for branching out into more specialised vocabularies concerning our hobbies, interests, and backgrounds.

## 2. The "core assumption"

The "core assumption" is illustrated graphically by Hansen and Stetting (1977):

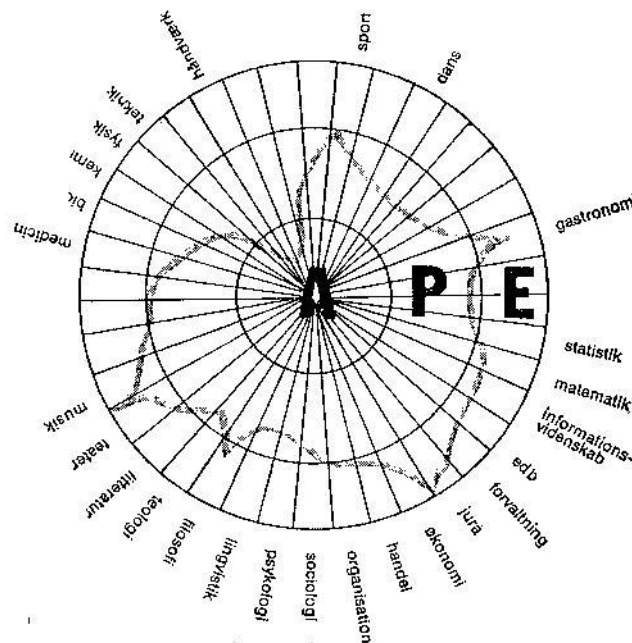


Figure 1.

The figure is an *abstraction*: the outer ring shows the total linguistic capacity in the population; and the shading indicates the vocabulary of a lawyer ("jura") who is a specialist (with an E(xpert) language) on music and fine foods ("gastronomi"). His generalised specialist language (P) reveals his interest in sports and philosophy and his total lack of interest in handicraft ("håndværk"); and, of course, he has mastered the general language (A), i.e. the syntax and the vocabulary of everyday communication, the "core" which we all know.

The "core assumption" is widespread among teachers - probably because their familiarity with a language is better than their students'; it is interesting that Brutten (1981) found that teachers were more inclined than students to pay attention to frequency when they identified the words they thought might be obstacles to comprehension. ...// 23 ...

Using frequency bands as our yardstick we can transfer the "core assumption" into two curves, one showing the vocabulary of a foreigner with a large English vocabulary, and another one with a fairly small vocabulary, as follows:

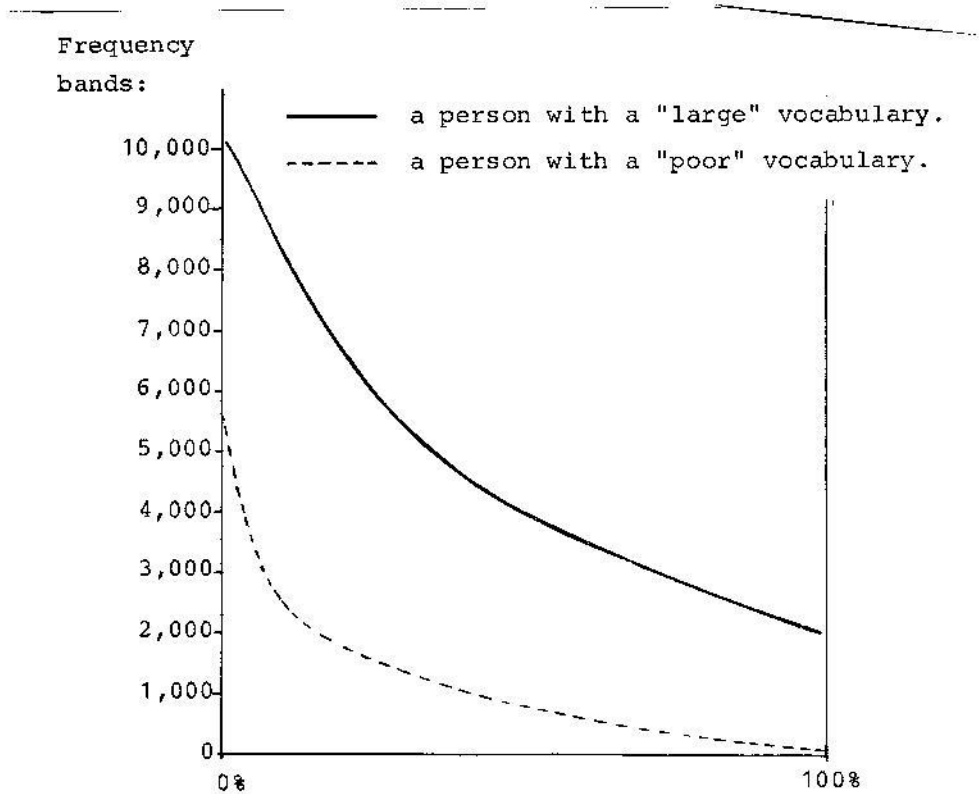


Figure 2: Curves showing the English vocabulary of two foreigners, against frequency bands

Studies have shown a relationship, particularly at the level of high frequency words, between vocabulary knowledge and frequency of occurrence.

When we are dealing with large groups of learners, a high frequency word unknown to some persons with large vocabularies should logically be unfamiliar to more readers with small vocabularies. We must therefore assume that every time persons with large vocabularies do not know a word, even more people with small vocabularies will find the word unfamiliar.

In order to investigate the core aspect of vocabulary knowledge, thirty volunteer freshman undergraduates at the Department of English at the University of Copenhagen participated in the vocabulary study. All participants answered four questions on their backgrounds. In addition, the participants in the vocabulary study were given six tests from the "Sprogtest" programme. ...// 24 ... The questionnaire, the instruction, and six tests (whose order was rotated) were handed out in envelopes.

The instructions requested the undergraduates to underline all the words that they did not know, or did not understand from their immediate experience of the text. We were aware that the instruction was ambiguous and might cover a wide spectrum: on the one hand some readers might underline any word they could not translate precisely in the context. On the other hand, other readers might underline only those words which they considered major stumbling blocks in their comprehension of the texts.

The texts were a newspaper article outlining a Swiss plan for the country's future development, a newspaper article about British miners who worked to keep out water from the coal mines during a strike, a popular science article on the potentialities of geothermal energy, especially in the US; a popular science article on natural disasters (floods and droughts) which were ultimately caused by human exploitation of nature, the opening of Chapter 31 in C.S. Lewis's *Arrowsmith* describing a plague spreading from China to the West Indies; and, the opening of Conan Doyle's short story *The Lost Special* where a man orders a special train to go to London.

No. of words underlined:		No. of readers	Cumulative no. of readers
0-10	**	2	2
11-20	***	3	5
21-30	****	4	9
31-40	*****	6	15
41-50	***	3	18
51-60	****	4	22
61-70	*	1	23
71-80	*	1	24
81-90	**	2	26
91-100	*	1	27
101-110			
111-120			
121-130	*	1	28
131-140	*	1	29
141-150			
151-160			
161-170			
171-180			
181-190	*	1	30
191-200			

\* = 1 reader

*Figure 3* Distribution of readers according to the number of different words they underlined in the whole sample.

...// 25 ...

Using the data in Figure 3, two groups of readers were chosen for analysis, namely the

five 'best' and the five 'poorest'.

When we speak of our "best" readers in the subsequent discussion, it must be understood that this is only an operational definition, a convenient, stylistic short-hand for "readers who have underlined few words in the six texts"; the five "best" readers underlined from 5 to 19 words in the whole sample (of 1156 different words). Conversely, our five "poorest" readers are the five participants who underlined the highest number of words - namely, from 87 to 183 words.

The only means for validating the underlining is a comparison between the backgrounds of the five "best" readers and those of the five "poorest" readers. The duration of their stays in the English-speaking world, their educational backgrounds, and the number of books they had read seem to bear on the readers' familiarity with words: the two best readers - who underlined 5 and 10 words respectively - had both read more than 50 English books and spent more than a year in the English-speaking world. The importance of prolonged stays in the English-speaking world was also uncovered in Johansson's study (1973) of Swedish undergraduates, so by and large there seems to be reason to assume that the greater the participants' familiarity with English, the lesser their inclination to underline words in the vocabulary study. There is, therefore, reason to believe that the underlining are not completely relativistic although this claim cannot be substantiated.

In another calculation, we listed the texts according to the number of words underlined by all readers and compared the listings thus obtained (with corrections for variations in length in the texts) with the individual students' rankings. The readers were in agreement about which texts were "most difficult". We interpret this as an indication that readers used some of the same criteria for underlining the words, and individually did so consistently throughout the texts.

### 3. Discussion: the readers' vocabularies



A Copenhagen mailman

#### 3.1 The "Core assumption" and the Frequency Bands

In this discussion we use the Thorndike and Lorge frequency bands. Although there are

major agreements between different counts in the highest frequency bands there are also variations in the order of the words in high frequency bands (e.g. Harris and Jacobsen, 1973; Dinnan, 1975). ...// 26 ...

Our choice of Thorndike and Lorge was determined by its comprehensiveness i.e. it reaches far into the low-frequency bands, which therefore opens up the possibilities of including "rare" words in the discussion. Nevertheless, we think that the identity of the frequency count used is actually immaterial to our conclusions on questions of theory and principles.

It is generally accepted that the less frequent a word is, the smaller the chance that readers will know it. We can check this assumption with our data, ranking all words underlined according to the number of readers who underlined them.

A listing of one random word in each group looks as follows (with an indication of the frequency band in parenthesis):

1 reader:  
2 readers: 3 readers: 4 readers: 5 readers: 6 readers: 7 readers: 8 readers: 9 readers:  
10 readers: *pitching* (2-3,000).  
11 readers: *spine* (5-6,000).  
12 readers: *magma* (not listed i.e. 30,000+).  
13 readers: (*a*) *stoop* (2-3,000).  
14 readers: *sewer* (7-10,000).  
15 readers: *immortelle* (not listed).  
16 readers: *parched* (7-10,000).  
17 readers: *pods* (10,000+).  
18 readers: *buccaneers* (10,000+).  
19 readers: *molten* (6-7,000).  
20 readers: *semi-arid* (arid: 7-10,000).  
21 readers: *scuppers* (20,000+).  
23 readers: *corroborate* (10,000+).  
24 readers: *sedateness* (10,000+).  
27 readers: *incandescent* (10,000+).  
29 readers: *seedie* (not listed).

In general, the list corroborated the "core assumption": few readers are unfamiliar with frequent words, and more readers with infrequent words.

### 3.2 The "good" readers

We also posited that if the "core assumption" holds good, there must be more "poor" readers than "good" readers who will be unfamiliar with a specific word: every time a word has been underlined by "good" readers, we must expect it to be underlined by even more "poor" readers. ...// 27 ...

Our five best readers had underlined 42 different word types 72 times. Only 2 of the 42 words failed to follow the pattern shown above: in *The Plague* two "good" - but no "poor" - readers underlined the word "lather"; and four "good", but only three "poor" readers underlined "careened".

With these exceptions, the results also strengthen the core assumption.

The list of the words unfamiliar to our "best" readers deserves a closer scrutiny; in parenthesis we list the number of "good" readers that found any given word unfamiliar:

Switzerland: *sedateness* (1).

Miners: *shotfirers* (1); *combustion* (2).

Energy: *fissures* (1), *harness* (1), *brine* (1), *crud* (up) (1), *feasible* (1), *sulfer*(1); *ample* (2), *(non)corrosive* (2), *molten* (2); *fiscal* (3), *rudimentary* (3).

Disasters: *devastating* (1), *prodigious* (1), *squander* (1); *parched* (2), *semiarid* (2), *rampaging* (2); *inundated* (3).

The Plague: *clattering* (1), *pods* (1), *hibiscus* (1), *buccaneer* (1), *berth* (1); *boisterously* (2), *sewer* (2), *lather* (2), *incandescent* (2), *scuppers* (2), *immortelle* (2); *careen* (4), *seedie* (boy) (4).

The Special: *stoker* (1), *ascertain* (1), *stoop* (1), *spine* (1), *dispatch box* (1); *oscillation* (2); *corroborate* (4).

Most of these words are rare: "ample" - the most common and frequent word underlined by our "best" readers is in the 3-4,000 word band; "combustion" and "molten" in the 6-7,000; and "parched" in the 7-10,000 word band. The majority of the unknown words are in the 10-20,000 word band, with "scuppers" and "rampaging" in the 20,000+ band. And, as mentioned, "immortelle" and "seedie (boy)" are not listed by Thorndike and Lorge at all.

Conversely, if the words in the texts serve as the point of departure the five "best" readers had no problems with numerous words in the 10,000+ range, e.g. *collated*, *deforestation*, *ecological*, *geological*, *hectare*, *jumbo jet* (set), *nuclear*, *reforest*, *technological*, *savanna(h)*, *round-the-clock*, *supplemental*, *unsparing*, *thermonuclear*, *overblown*, *supercargo*.

Some of these words are undoubtedly more common today than when the corpus of the Thorndike and Lorge count was written e.g. *nuclear*. Even so, the best readers know many highly infrequent words: their vocabularies are very large, and not confined to words from their own specialist areas. It is true that some of these words, e.g. *hectare*, *savanna(h)* also exist in Danish. ...// 28 ...

But if we uncritically accepted that Danish readers would know English words which looked like Danish ones we would miss a point: these words are not very frequent in Danish either, so the impression that some readers have large receptive vocabularies is not weakened.

### 3.3 The five "poorest" readers

We would expect our poorest readers to know only "core-words" and then only odd words above a certain boundary (which would, in turn, depend on the readers' knowledge of English). As mentioned, our "poorest" reader underlined 187 words. Among unfamiliar words were *current* (1-2,000 word band); *acknowledge* in the 3-4,000 word band; *complex* (5-6,000) etc. But curiously, words like

*available, code, and economy* (3-4,000 word band); *dilemma* (10,000+); *depopulate* (20,000+), and many similarly infrequent words were not underlined.

### 3.4 All thirty readers

*Affair, bright, forest, c(ent), and guard* in the 0-1,000 word band were each underlined by only one reader. So were *bore, bound, current, firm, flat* in the 1-2,000 word band. In the 2-3,000 band *attach, commit, and depth* were likewise unknown to one reader each - only *application* was unfamiliar to 5 readers. In the 3-4,000 word band *apparent* was unknown to one reader; two readers underlined *available, contribution, decrease, emergency*: and no less than 12 readers indicated that *ample* was unknown to them. However, if we look at the texts in another way, the list of words from the low frequency bands unfamiliar to only one of the thirty participants looks as follows:

3-4,000 word band:	<i>amaze, chapel</i>
4-5,000:	<i>barrier, cargo</i>
5-6,000:	<i>banana, breathless, complex</i>
6-7,000:	<i>balcony, bamboo</i>
7-10,000:	<i>breakdown, annual, conservation, comical, dependence, first-class, fragile, market-place, phenomena, rainfall, sensational, ski, skipper, smear, spokesman, spontaneous, underlying, urban.</i>
10,000+:	<i>bazaar, centre, dilemma, efficiently, ensure, exotic, fantastically, geyser, inefficient, inexplicable, middle-aged, monsoon, deforest, depopulate, geological, nuclear, overblown, reforest, supplemental, technological, periodically, physique, potentially, seasonal, second-class, turbine, upstream, washerwoman.</i>
30,000+:	<i>hectare, round-the-clock, breakthrough ...</i>
... // 29...	

## 4. Discussion

The "core assumption" appears to hold good as very few Danish readers of English at an advanced level met with unfamiliar words in reading below the 3-5,000 word boundary.

The exact boundary however, can not be defined. Even if we had established it, we could not claim that it would apply to all learners of EFL: in other words, we cannot and will not argue that all learners of EFL must know any specific number of words in order to manage.

In addition, there is an equally important result: many undergraduates appear to know even infrequent words, and this cannot be explained by simply combining the "core assumption" with frequency bands. Many of the words discussed would be very infrequent in any general frequency count of the English language.

## 5. Vocabularies and reading strategies



The "Sprogtest" programme comprises other studies than the vocabulary study, including an introspection study where 28 other readers - 7 undergraduates and 21 students in the modern language stream at the gymnasium ('high school') - reported on their reading and test-solving techniques during the reading (Dollerup, Glahn and Rosenberg Hansen, 1982).

This particular study leads us to suggest that the "core assumption" should be supplemented with reading and decoding strategies. This would explain why our readers had fewer difficulties with low-frequency words than expected.

These strategies include the following:

1. Etymological, morphological, and (transparent) semantic decoding using

1a. Components of words they know from another language (mostly Latin): Text: "*Decreasing the Inconvenience*" Reader's comment : *I don't know how to translate 'decreasing'. Then I think of Latin 'convenio' ...*

1b. Components from English words familiar to the readers.

1c. A knowledge of a Danish word which looks more or less like the one read: for example, the English word flood (inundation) was often taken to mean 'river' which translates as Danish '*flod*' (a so-called 'false friend').

2. Translation into Danish. The speech cited at 1a. illustrates this strategy which applies to both passages and words (compounds).

3. Context: e.g. "*I have seen these words before, but I do not know what they mean: when it says 'the first carriage was solely ...' this must mean 'only'. ...//30 ... I go for the first answer to the multiple-choice question because I skim the text. It says that the carriage has only first and second class compartments.*"

We suggest that these and other strategies provide an explanation why the students' know low-frequency words in the vocabulary study.

One last point - also mentioned by Anderson and Freebody (1981) and Nation (1983) must be made, viz., that the concept of knowing a word is problematic. From our sample, it seems as if one strategy is to get a hazy idea of what a word means, assess that it is fairly unimportant, and then accept this vague impression as "familiarity"; thus only half the readers underline the word *immortelle*, presumably because it occurs in the sentence: "the *immortelle* that fills the valleys with crimson". The sentence signals that *immortelle* is a kind of red large flower, and in the wider context, it serves only to give flavour to the description of a tropical island.

## 6. Concluding remarks

We suggest that in reading, we are not dealing with a static entity when we speak about a

vocabulary but a changing and fluid mass.



There is a core of words, a word knowledge, which centres around the most frequent words in the language and the size of which may vary with readers' personalities and backgrounds. This word knowledge is, we suggest, relatively - but not completely - stable, and its size can be estimated, with the limitations imposed by the methods used and the definitions of vocabularies employed. *But this word knowledge is only part of a reader's receptive vocabulary.*

Another part of the vocabulary consists of *the strategies that individual readers use for decoding words and for gaining an overall comprehension.* This has been touched upon by others. Thus Amaud (1984) cites Denninghaus as having used the term "potential vocabulary" about words hypothetically known to learners. Nagy and Anderson (1984) suggest that knowledge of infrequent words increases with exposure to language, and refine this in Nagy, Herman and Anderson (1985) to an ability to learn by context. We wish to stress, however, that (a) the strategies are not identical with a learning process but that the words are understood and known in one particular context and perhaps only momentarily, and (b) that this applies to reading. We do not preclude that this approach applies to other situations as well, but leave this problem for others to solve.

*A third component of a vocabulary is the text which is actually being read:* it is only in the reading of a text that the strategies and the word knowledge can interplay. To be explicit: there are words which an individual reader will meet with and immediately understand only once in a lifetime.

In summary, readers' vocabularies in the reading process consist of (a) a "word knowledge store", (b) strategies for decoding words, and (c) the special linguistic context. ... // 31 ... It implies that individual vocabularies in reading exist instantaneously, and that they are, in effect, fluid entities which change every time they are generated by the reading of specific texts. *Vocabularies differ not only in time but also from text to text with the same reader.*

The following sketch indicates the nature of individual receptive vocabularies in reading.

 reader with a large core vocabulary and many strategies.  
 reader with a small core vocabulary and few strategies.

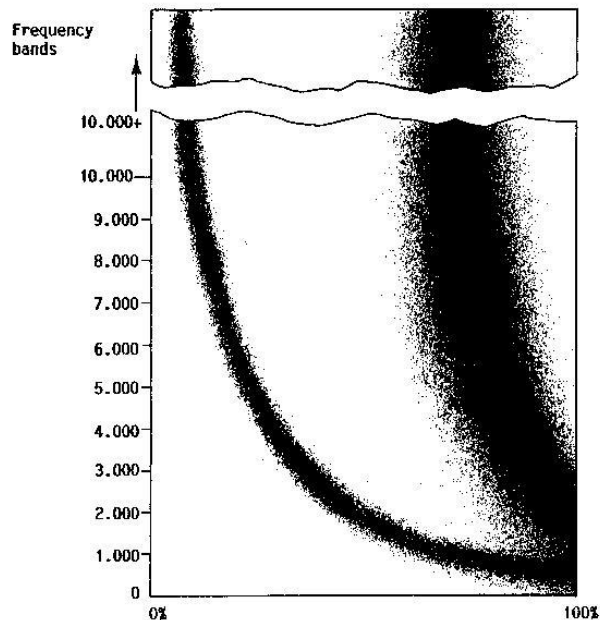


Figure 4

In this Figure the left hand column indicates the frequency bands. It includes all words in a specific language, even those not listed in the most comprehensive dictionaries: therefore we leave the upper limit open (which does not mean that vocabularies are infinite).

In adding the readers' reading and decoding strategies, we suggest that poor readers with few strategies at their disposal will know fewer words in any given frequency band than the good readers; yet they will still know some very rare words. ...// 32 ...

The results indicate that the importance of vocabulary coping strategies should not be overlooked; there should be a conscious instruction in the rules of word formation and word derivation. Most of all reading strategies should be taught as an integral part of these activities.

In vocabulary testing it should be more readily acknowledged that frequency lists may tell some part of the truth and a useful one at that - but sometimes they are a far cry from the whole truth.

We wish to thank all the readers who participated in the present study. We are indebted to Ethel Ussing for her unfailing help in setting up the material for "Sprogtest", and for the present vocabulary study; and to Anette Andersen and Marlene Bamer for having typed this article. We are grateful to our assistants over the years, Benedicte Holbak, Birte Kristensen, Karin Sigurdskjold and Eva Schaumann.

## References

- Afflerbach, Peter P., Richard L. Allington, and Sean A. Walmsley. (1980), A Basic Vocabulary of US Federal Social Program Applications and Forms. *Journal of Reading* 23, 332-336.
- Anderson, Richard C. and Peter Freebody. (1981), Vocabulary Knowledge. In Guthrie, John T. ed. *Comprehension and Teaching : Research Views* Newark: IRA, 77-117.
- Arnaud, Pierre J.L. (1984), A Practical Comparison of five types of vocabulary tests and an investigation into the nature of L2 lexical competence. Paper read at AILA, Bruxelles, 5-10 August, 1984.
- Brutten, Sheila R. (1981), An Analysis of Student and Teacher Indications of Vocabulary, Difficulty. *RELC Journal*, 12, 66-71.
- Dinnan, James A. (1975), A comparison of Thorndike-Lorge and Carroll prime frequency word lists. *Reading Improvement*, 12, 44-46.
- Dollerup, Cay, Esther Glahn, Carsten Rosenberg Hansen. (1980), Some Errors in Reading Comprehension. In Faber, H. von Ed. *Leseverstehen im Fremdsprachenunterricht* Munich: Goethe Institut.
- Dollerup, Cay. (1981), *Studies in the Major Modern Languages (English, German, French), at University Level in Denmark by 1980/81* (ERIC ED 203 681).
- Dollerup, Cay, Esther Glahn, Carsten Rosenberg Hansen. (1982), Reading Strategies and Test-Solving Techniques in an EFL-Reading Comprehension Test: a Preliminary Report. *Journal of Applied Language Study*, 1, 93-99.
- Farr, Roger and Robert F. Carey. (1986), *Reading : What can be measured* 2nd edition Newark: IRA. ...// 33 ...
- Goodman, Kenneth S. and Louis Bridges Bird. (1984), On the Wording of Texts: A Study of Intra-Text Word Frequency. *Research in the Teaching of English*, 18, 119-145.
- Hansen, Inge Gorm, and Karen Stetting. (1977), Specialsprog. In Glahn, Esther and Leif Kvistgaard. eds. *Fremmedsprogs pædagogik* Copenhagen: Akademisk forlag.
- Harris, Albert J, and Milton D. Jacobson. (1973-74), Some Comparisons between Basic Elementary Reading Vocabularies and Other Word Lists. *Reading Research Quarterly*, 9, 87-109.
- Johansson, Stig. (1977), Reading comprehension in the native and the foreign language: on an English-Swedish comprehension index. In Zettersten, Arne. ed. *Papers on English Language Testing in Scandinavia*. Copenhagen: Anglica et Americana 1, 43-58.
- Journal of Reading* (IRA), (1986: no 7. Special Issue on Vocabulary.)
- Nagy, William E. and Richard C. Anderson. (1984), How many words are there in printed school English? *Reading Research Quarterly*, 20, 233-253.
- Nation, L.S. Paul. (1983), *Teaching And Learning Vocabulary* Victoria University of Wellington : English Language Institute.
- Noesgaard, A. og Vagn Pedersen. (1949), *Hyppighedsundersøgelser over Engelsk som Fremmedsprog (fire begynderbøger)*. Copenhagen: Fr. Bagge.
- Oppertshausen, Otto. (1974), Absolute oder relative Häufigkeit? WortStatistik als Hilfsmittel zur Aufstellung eines verbindlichen Mindestwortschatzes für den Englischunterricht im Sekundarbereich I. *Praxis des neusprachlichen Unterrichts*, 31, 42-52.
- Richards, Jack C. (1970), A Psycholinguistic Measure of Vocabulary Selection. *IRAL*, 8, 87-102.
- Thorndike, Edward L. and Irving Lorge. (1944), *The Teacher's Word Book of 30,000 Words*. New York: Columbia University.
- Zettersten, Arne. (1979), *Experiments in English Vocabulary Testing*. Malmö: Hermods.



College graduates dancing around the statue on the King's New Square, Copenhagen in June on Graduation Day

