DOCUMENT RESUME

ED 313 715 CS 212 180

AUTHOR Ljung, Magnus
TITLE Swedish Upper Secondary School English.
INSTITUTION National Swedish Board of Education, Stockholm.
PUB DATE Nov 89
NOTE 14p.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Collected Works - Serials (022)
JOURNAL CIT School Research Newsletter; n10 1989

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Basic Vocabulary; Comparative Analysis; *English; Foreign Countries; Secondary Education; Second Languages; *Standard Spoken Usage; *Textbook Content; Textbook Evaluation; Textbooks; Word Frequency
IDENTIFIERS Sweden; Text Analysis

ABSTRACT

A study examined the English vocabulary of English textbooks used in Swedish upper secondary schools to distinguish differences between the vocabulary of the textbooks and modern, everyday English as represented by newspapers, books and the colloquial language. Fifty-six books were included in the sample. In the selection process, account was made not only of the popularity of each book, but also of its distribution among different types of school and different grades to determine the rising level of difficulty in vocabulary between grades. These texts, referred to as the GYM corpus, were then compared to the COBUILD Corpus which represents the largest computerized English collection of texts available. Results revealed a different vocabulary profile in GYM texts from that found in normal English prose as represented by COBUILD texts. Proportions between abstract and concrete, between complicated and simple, were found to be shifted in GYM texts in favor of the concrete and simple. To some extent this is understandable, since the COBUILD texts are mainly addressed to adult, native speakers of English. Additionally, GYM texts appeared insufficiently progressive, with the result that the difficult words were very often randomly distributed among the grades, instead of growing successively more common. (One figure and four tables of data are included.) (KEH)

SCHOOL RESEARCH NEWSLETTER

NOVEMBER 1989 FIELD-TOPIC 1989:10
Language teaching

Concluded project:
SWEDISH UPPER SECONDARY SCHOOL ENGLIS..

Swedish Upper Secondary School English, was an NBE-funded project, started in 1986, in the Department of English, University of Stockholm, to evaluate the English vocabulary of English textbooks used in Swedish upper secondary schools.

The evaluation took the form of a comparison between the vocabulary of the textbooks and modern, everyday English.

PROJECT CONDUCTED AT:
The Department of English
Stockholm University
S-106 91 Stockholm

PROJECT LEADER:
Magnus Ljung

Figure 1 shows the structure of the corpus.

TEXTS	WORDS	HAP	TYPES	VAR	NAT	UST	PUB	GR
Action 1	12,494	1,535	1,750	.56	Sw.	2us	83	1
Action 2	24,261	2,096	3,944	.53	Sw	2us	84	2
All in one 1	34,496	2,924	5,232	.56	Sw	3us	74	1
All in one 2	39,683	3,599	6,346	.58	Sw	3us	75	2
All in one 3	62,040	4,343	7,901	.55	Sw	3us	76	3
As You Like It 1	21,799	1,381	2,942	.47	Sw	2us	79	1
As You Like It 2	24,689	1,851	3,666	.51	Sw	2us	85	2
Authentic English	3,760	936	1,418	66	OUP	int	80	0
Business World	32,798	2,804	5,472	.51	OUP	0	83	0
Challenge to think	14,609	1,530	2,792	.55	OUP	0	82	0
Cross Section 2	38,302	2,866	5,322	.54	Sw	3us	85	2
Cross Section 1	24,495	1,871	3,697	.51	Sw	3us	84	1
Crossroads	35,203	2,807	5,079	.55	Sw	3us	79	1
Echoes	43,901	3,607	6,491	.56	Sw	3us	80	2
Encounters	37,310	3,403	5,878	.58	Sw	3us	75	0
Free choice 1	36,370	2,509	4,794	.52	Sw	3us	76	1
Free choice 2	32,965	2,762	5,039	.55	Sw	3us	81	2
Hitchhiker	33,945	1,447	3,447	.42	Sw	us	81	1
Impressions	32,565	3,170	5,448	.58	Sw	ad	82	2
Insight 1	26,388	2,246	4,201	.54	Sw	2us	78	1
Insight 2	33,404	2,880	5,323	.54	Sw	2us	78	2
Lifelines 1	11,204	1,255	2,322	.54	Sw	2us	82	1
Lifelines 2	8,083	1,147	2,701	.43	Sw	2us	83	2
Listen to this	3,693	727	1,166	.62	OUP	int	75	0
Manage with English	17,458	1,085	2,514	.43	OUP	int	81	0
Modern short stories	37,152	3,504	5,508	.64	OUP	adv	81	0
More modern short st	44,093	3,307	6,014	.55	OUP	adv	81	0
New openings	33,103	2,156	4,220	.51	Sw	us	85	1
Now you are talking	6,846	897	1,605	.56	Sw	0	75	0
Outlook 1	29,051	2,443	4,527	.54	Sw	3us	74	1
Outlook 2	32,141	3,049	5,351	.57	Sw	3us	76	2
Outlook 3	24,597	2,852	4,939	.58	Sw	3us	79	3
Over to you	2,826	677	1,012	.67	Sw	3us	84	0
Pace 1	15,679	1,513	2,848	.53	Sw	2us	83	1
Pace 2	18,224	1,723	3,172	.54	Sw	ad	83	2
People in action 1	27,128	1,905	3,832	.50	Sw	2us	73	1
People in action 2	24,584	1,848	3,593	.51	Sw	2us	75	2
Prospects	35,667	3,521	6,163	.57	Sw	ad	77	3
Quartet 1	15,489	2,013	3,437	.59	OUP	int	82	0
Roadrunner 1	21,191	2,254	3,963	.57	Sw	3us	85	1
Scope 1	12,810	1,180	2,290	.52	Sw	2us	85	1
Scope 2	19,054	1,690	3,552	.48	Sw	2us	86	2
Side by side 1	22,766	1,926	3,687	.52	Sw	3us	78	1
Side by side 2	29,469	2,713	4,829	.56	Sw	3us	79	2
Side by side 3	35,385	3,224	5,744	.56	Sw	3us	80	3
Spinoff 1	34,076	2,020	4,148	.49	Sw	3us	77	1
Spinoff 2	26,882	2,395	4,280	.56	Sw	3us	77	2
Spinoff 3	29,711	2,486	4,494	.55	Sw	3us	78	3
Visions	24,221	2,142	3,806	.56	Sw	ad	76	1
Voices 1	5,660	729	1,370	.53	Sw	2us	73	1
Voices 2	13,565	1,349	2,573	.52	Sw	2us	74	2
Waiting for the police	15,597	1,981	2,274	.48	Sw	2us	59	1

Table 1. The texts of the GYM corpus.

The table headings require some explanation. "WORDS" refers to the number of words in each book, regardless of any repetitions.

HAP is short for hapax legomena, i.e. words occurring only once, a figure sometimes used to measure the size of the vocabulary in a text.

TYP refers here to each word in the text, counted only once. The TYPE:WORDS ratio is often used as another yardstick of the size or variety of the vocabulary in a text. The VAR column shows the results of this calculation for the different texts. As will be seen, Over to you has the most abundant vocabulary, but like many other books with high variation values, it is very short, which favours high values. More interesting in this context, therefore, is Modern Short Stories, which, surprisingly enough, comes a long way ahead of More Modern Short Stories.

The NAT column shows the nationality of the publisher, and UST shows types of upper secondary school. Here, apart from 2us and 3us (2-year and 3-year lines of upper secondary school respectively), we have the abbreviations int, upin and prein, standing for intermediate, upper intermediate and pre-intermediate. PUB and GR, finally, indicate publishing year and grade respectively. A zero in the GR column means that no grade is specified, as for example in the case of all the OUP books.

Sub-quantities

As explained earlier, the entire corpus was divided up into a number of sub-quantities. First and foremost, a distinction was made between the entire GYM corpus and the books passed as basic teaching materials. Excluded from the basic teaching material group are the 12 books from the Oxford University Press and eight Swedish books not passed as basic teaching materials. The basic teaching materials themselves were then divided into GRADE 1, GRADE 2 and GRADE 3 groups. Here one finds all texts intended for Swedish upper secondary schools, regardless of type. This group, for example, includes all books intended for 2-year and 3-year lines of upper secondary school, both in youth and in adult education.

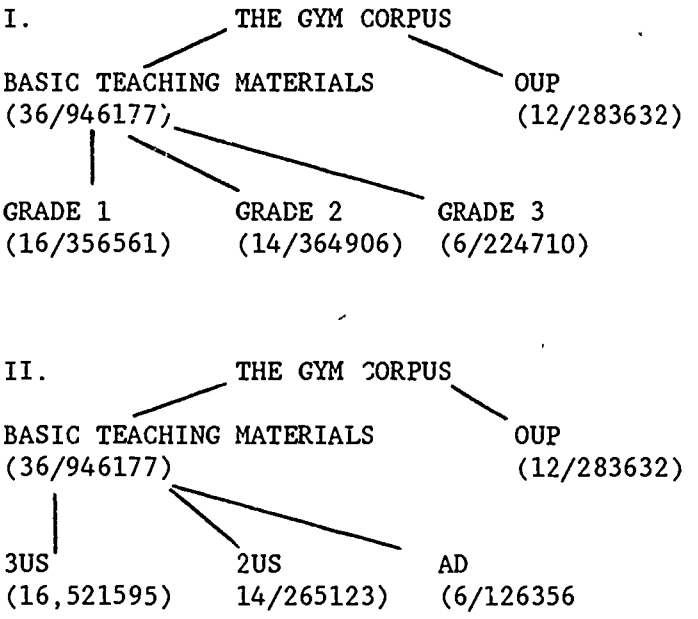

Fig. 1. Parts of the GYM corpus. Figures in parenthesis refer to number of books and words respectively.

In the 3US and 2US groups I have assembled all books intended for the 3-year and 2-year lines of upper secondary school respectively in youth education. AD contains six books intended for adult education at upper secondary level.

Preparations

The first step in processing the texts was to transform the entire corpus from the printed word to a mechanically readable format, using a scanner. With such a large corpus - running to about 1.5 million words - this was a time-consuming process which had to be farmed out to companies specialising in this type of work. First the books had to be read through and all material not to be included - such as word lists, items of grammar etc.- specially marked.

The actual scanner work took a very long time, and since, moreover, the percentage transfer error was as high as 5% or more, the transferred text had to be called up on the screen, compared with the original and fully corrected. All in all, this part of the project took about 18 months.

Selection of comparative text

The choice of a suitable English standard with which to compare the upper secondary school texts was restricted by certain definite criteria. The most important of these was that the comparative corpus had to be as extensive as possible, mechanically readable, comprehensive with regard to type of text and representative of modern English.

There are several mechanically readable corpuses of English texts on the market today. The best-known include the BROWN and LOB corpuses, which comprise one million words each and contain a carefully considered sample of different types of text from American and British English. These corpuses, however, have several disadvantages. They are too small to allow interesting comparisons of vocabulary, and the English they represent is somewhat outmoded, because all the texts they include were written and published in 1961.

A far more comprehensive corpus, with slightly more modern text, is The American Word Heritage Frequency Book, compiled by Carrol et al. Here again, though, the texts are somewhat outmoded, added to which they represent a rather specialised material, viz literature for American juveniles.

One special corpus which to some extent has also been used as comparative material in this study is the corpus of colloquial British English produced in London and Lund. This corpus contains 500,000 words and contains different types of colloquial language recorded (very often with a hidden microphone) during the 1960s and 1970s.

Purely generally speaking, however, it would be odd to use a purely colloquial corpus as one's standard when evaluating the GYM corpus; on the other hand, of course, it is interesting to see how close together or far apart these two corpuses are. It must be made clear, however, that most of the speakers in the London-Lund corpus are British graduates, a fact which leaves its mark on the vocabulary.

The only corpus ultimately proving to meet all our requirements was the text, containing 18 million words of modern English (written and spoken),

which form the basis of the COBUILD dictionary compiled by John Sinclair and his associates at the University of Birmingham.

The COBUILD corpus is at present the largest computerised English collection of texts available. Moreover, it contains a vocabulary which, to a great extent, is a good deal more modern than in other text collections. Its shortcomings are organisational. It does not have the strict division into types of text and sub-departments to be found, for example, in BROWN, LOB and the Lund-London corpus.

The advantages of the COBUILD material, however, are so great that they amply outweigh the disadvantages, and permission was therefore obtained to compare the GYM corpus and the COBUILD texts. This again took longer than expected, because permission had to be obtained from Collins, who own the copyrights for the texts.

The mechanical side of the comparison was performed on a mainframe at the Research and Development Unit for English Language Studies, University of Birmingham. Before this work could actually begin, the two corpuses had to be lemmatised, as the term goes, i.e. conjugated forms of a word were gathered under a main form, a lemma.

The work done in Birmingham was confined to lemmatisation and to a lemma-by-lemma comparison of the GYM corpus as a whole with the COBUILD material. Other processing, such as the breaking down of the material into sub-divisions and lemmatisation and evaluation of those sub-divisions, took place at the Department of English, Stockholm University, and in the Stockholm University Information Processing Centre (QZ).

Results

There are various ways of comparing large masses of text. One can, for example, employ the relation between types and words as a quantification of the vocabulary of the texts.

The GYM corpus contains 1,437,474 words and 44,066 word types. This gives a type:word ratio of 44,066/1,437,474, i.e. 0.0306. We can round

this off to 0.03 and compare it with the text collections mentioned previously. Table 2 shows the results of the comparison.

GYM	0.003
Carroll	0.017
Lund/London	0.003
BROWN	0.051
LOB	0.049

Table 2. Vocabulary size of various English corpuses.

Another, somewhat more exact method is based on investigating the proportion of vocabulary within a certain frequency band which is common to different text masses. Comparing the 1,000 most frequent lemmana in GYM and COBUILD, one finds 789 common to both. This is a slightly lower figure than one is entitled to expect. Usually there is 80% agreement or more in the 1-1,000 frequency band.

There are, then, 211 words in the top frequency band of the GYM corpus which are unique to this corpus, i.e. do not occur among the 1,000 commonest COBUILD words. (Here I am using word to mean "lemma".)

A comparison of the 211 unique words - words not common to the corpuses- reveals interesting differences. Disregarding the function of words, one finds that practically all the words unique to the upper secondary school texts denote concrete objects or events (football, accident), observable processes (sing, sink), feelings (angry, glad), and value judgements (dangerous, safe).

The overwhelming majority of the unique COBUILD words denote abstracts (purpose, accord), ambiguous processes or relations (achieve, require) or adjectives not referring to physical properties (basic, nuclear).

Tables 3 and 4 show examples of words unique to each corpus (among the 1,000 most frequent words).

Table 3. Examples of unique GYM words

Noun	Adjective	Verb
accident	angry	breathe
aunt	bright	disappear
boat	electric	frighten
dinner	empty	repeat
dollar	expensive	shut
driver	glad	sing
knee	Irish	sink
mum	popular	steal
passenger	quiet	steal
policeman	safe	
pub	sick	
truck	soft	
TV	terrible	
uncle	wonderful	

Table 4. Examples of unique COBUILD words

accord	apart	achieve
activity	basic	apply
attitude	economic	assume
choice	industrial	assume
decision	nuclear	depend
difficulty	likely	establish
evidence	particular	occur
image	physical	prove
method	political	reduce
purpose	various	require
relation	similar	tend

Differences between other high-frequency words

Turning now to the 789 lemmana in the 1-1,000 frequency band, in both the GYM and COBUILD corpuses, we once again discover conspicuous differences. After adjusting for the difference in corpus size, each word was given a difference co-efficient, i.e. a value indicating the difference between the relative frequencies of that word in the two corpuses. This coefficient was calculated as follows:

(Frequency in A - Frequency in B)/(Frequency in A + Frequency in B)

with A and B denoting different corpuses. Thus what the coefficient does, quite simply, is to work out the difference between the frequencies of the word in the GYM and COBUILD corpuses and divide the result by the sum total of those same frequencies.

The coefficient assumes a value between +0.99 and -0.99, the first of these values indicating a very high excess representation of a given word in the GYM corpus and the second an equally pronounced under-representation in the same corpus.

Still confining ourselves to the 789 high-frequency words common to both sets of material, we find that 355 words have a much higher frequency in GYM than in COBUILD, 178 have a much lower frequency in GYM than in COBUILD, and 256 are more or less equally common in both corpuses.

Closer study of the 355 over-represented and the 178 under-represented words in the GYM corpus reveals the same tendencies as we have already found among the unique words. Thus the over-represented words include a heavy preponderance of words denoting concrete, observable phenomena, e.g. chair, floor, kitchen, home, table.

The under-represented words, as expected, include many abstracts, but also many terms for social phenomena - such as society, government, party - and words used to evaluate different phenomena, such as condition, quality, rate.

The differences described above are further accentuated when the rest of the material is included in the comparison, i.e. the words outside the 1-1,000 frequency band. What is more, compared with the COBUILD corpus, the upper secondary school texts proved to have a preponderance of informal words, e.g. mum, dad. The upper secondary school texts also present a heavy dominance of words connected with certain subject fields, e.g. the family, outings and sport.

Another distinctive property of the upper secondary school texts is their predilection for contracted forms. Heavily over-represented contractions in the GYM texts include gonna (+62), what'll (+062), (+0.52), you'll (+44), I've (+41), and she's (+35).

Conclusions

As we have now seen, the two corpuses presented differences even as regards the 1,000 commonest words. Those differences would not have been

very startling if they had only been connected with individual words. The interesting thing is that the discrepancy betweer the two corpuses can be described in terms of opposites, such as abstract-concrete, complex-simple and so on.

The vocabulary of the upper secondary school texts, in other words, appears to have quite a different profile from the COBUILD texts, i.e. different from that found in normal English prose: the proportions between abstract and concrete, between complicated and simple, have been shifted in favour of the concrete and simple.

To some extent this difference is understandable and justifiable: the upper secondary school texts are intended for non-native speakers aged between 16 and 19, while the COBUILD texts are mainly addressed to adult, native English speakers.

One may ask, however, whether such great differences as those actually occurring are reasonable, especially considering that the upper secondary students, when they begin their studies, already have six years' English behind them. One of the stumbling blocks to students going on, after upper secondary school, to read English at university, is understanding ordinary texts in newspapers and magazines like The Observer, Newsweek and Time. It is an open question whether one should not be able, after a total of nine years' studies, to read this type of text without difficulty.

A comparison between texts intended for the different grades yields interesting results. One might expect the more difficult, more abstract words to grow more common as one moves up through the grades. This kind of progression, however, is often hard to find. Whereas, for example, social and evidence show rising frequencies from grade 1 to grade 2 and from grade 2 to grade 3, words like choice and private show much the same distribution in all three grades. It is also easy to find "difficult" words which are more common in grade 1 and/or grade 2 than in grade 3.

One tentative conclusion, therefore, is that the upper secondary school vocabulary as a whole presents a flatter profile than normal English texts. Another conclusion is that the textbooks are insufficiently

progressive, with the result that the difficult words are very often randomly distributed between the three grades, instead of growing successively more common as one proceeds from grade 1 to grade 3.

In addition, the upper secondary school texts, compared with COBUILD, display a clear over-emphasis of informal words and structures, which is interesting, considering that COBUILD includes about 25% spoken English.

Further development

The points mentioned above hint at the results which can be obtained by comparing word lists from the two corpuses. There are a host of other data which can be extracted from the material at word list level. One can, for example, see the material as a reflection of the English-speaking society described in the texts and investigate which parts of society are included and which omitted. One can also investigate the balance between British and American English in the choice of spelling variants, different types of expression and so on.

It would also be interesting, however, to supplement these data by means of a study based on searches in the actual texts from the upper secondary school corpus and in concordances based on them. Studies of this kind will be undertaken in the autumn of 1989 and will supply information, for example, concerning verbal phrases, idiomatic expressions and other units above word level.

Finally, we may add that the method and software used in this study are also applicable to new textbooks, so long as they are available in mechanically readable format, which in turn will make it possible for vocabulary profiles for new teaching materials to be compiled continuously.

SCHOOL RESEARCH NEWSLETTER

School Research Newsletter contains reviews of curent and terminated research within the field of education.

School Research Newsletter is published and distributed in an edition of 600 issues ten times per year. The Newsletter is obtained free of charge.

Published by Swedish National Board of Education,
Department for Coordination and Planning.
S-106 42 Stockholm. Telephone +46 8 783 20 00
Responsible Editor: Inger Marklund. Editor: Cecilia Dahlberg
ISSN 0345 5343