DOCUMENT RESUME

ED 343 133 CS 213 217

AUTHOR Ljung, Magnus
TITLE Swedish TEFL Meets Reality.
PUB DATE 91
NOTE 18p.; For an earlier report of this study, see ED 313 715.
PUB TYPE Reports - Research/Technical (143)

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

ABSTRACT

The work reported in this paper is an offshoot of a project started in early 1986 and designed to produce a comparison between the vocabulary in the TEFL (Teaching English as a Foreign Language) texts and the vocabulary found in contemporary non-technical English writing. This study investigated the unique words and differences in frequency between shared words in two large corpora, the GYM corpus (from "gymnasium" the Swedish term for upper secondary education) and the COBUILD corpus, a huge collection of machine readable English texts collected at the University of Birmingham (England). The GYM corpus was created by converting the nearly 1.5 million words in 56 English TEFL texts used in Swedish upper secondary education into computer-readable form. The COBUILD corpus contains 18 million words used in the development of the COBUILD dictionary. Comparison of the 1000 most frequently used words in the two corpora indicated that they have 796 words in common. Comparison of the unique words indicate that: (1) words unique to the GYM corpus are concrete terms denoting physical objects and processes, physical characteristics, and emotions; and (2) words found only in the COBUILD corpus are predominantly abstract. Comparison of the words in both corpora indicated a larger number of contractions and use of third person pronouns in the GYM corpus than in the COBUILD corpus, indicating a preponderance of narrative text in the GYM corpus. Findings suggest that it is in fact possible to obtain a fair amount of information about texts merely by looking at word lists. (Seven tables of data are included; 13 references are attached.) (RS)

SWEDISH TEFL MEETS REALITY

Magnus Ljung

Department of English, University of Stockholm

1 The aim of this study is to account for some of the findings made in an evaluation of the vocabulary found in English texts used in Swedish upper secondary education (the "gymnasium") and to demonstrate a few of the techniques that can be used in corpus comparison.

The work presented here is an offshoot of a project which started in early 1986 with financial support from the Swedish National Board of Education. The project's objective was to produce a comparison between the vocabulary in the TEFL texts and the vocabulary found in contemporary non-technical English writing. A more detailed description of the original project will be found in Ljung (1990).

Since the number of textbooks currently in use is far too great to allow an examination of all the relevant TEFL texts, it was decided to investigate a representative sample. Originally a selection was made comprising 60 titles, a number which was subsequently reduced to 56 on account of certain technical difficulties.

The books in the sample were converted to computer-readable form with the aid of a scanner. The end result of the scanning process was a corpus containing a total of 1437474 running words, which I will henceforth refer to as the GYM corpus (GYM for gymnasium). Some of the statistics for the corpus are shown in Table 1.

Table 1. Basic statistics of the GYM corpus

Number of tokens	1437474
Number of types	44066
Type: token ratio	.03

A few comments on Table 1 are in order. The type: token ratio is fairly low. On its own this value is difficult to interpret. It may just be a natural consequence of the length of the corpus (it is well known that the type: token ratio is sensitive to text length). On the other hand it may be a reflection of the lexical composition of the TEFL texts.

In order to make ;ense of the statistics above we have to compare t: e GYM type: token ratio with corresponding values from comparable corpora. The results of such a comparison have been set out in Table 2.

Table 2. Comparison between the type: token ratio in certain computerized corpora

The GYM corpus	.030
Carroll	.017
Brown	.051
LOB	.049

In Table 2, Carroll refers to the corpus used for The American heritage word frequency book; cf Carroll et al (1971).

With the aid of Table 2 we can now gain a better perspective on the GYM type: token ratio. The corresponding value for a very large corpus like Carroll et al - which contains 5,088,721 running words - is as low as .017, while the ratios for Brown and LOB are in the vicinity of .050.

Given that the length of the last two corpora is roughly two thirds of the length of the GYM corpus, the higher type: token ratios found here are only to be expected. It is of course impossible to tell to what extent the difference between the ratios is due to the differences in text length, rather than to the complexity of the Brown and LOB corpora. However, the comparison shows beyond any doubt that the type: token ratio found for the GYM texts must be due in large measure to the size of the corpus.

2 An evaluation of the GYM texts presupposes the existence of a suitable standard of comparison Such a standard must meet a number of criteria. Chief among these are the requirements that it be comprehensive, machine-readable, up-to-date and represent a fair number of written genres.

Brown and LOB, which are without doubt the best-known among the English language corpora, both meet the requirements of machine-readability and diversity of genre. On the other hand they are too limited to permit interesting comparisons of vocabulary and have the additional drawback of representing the British and American English of the early 1960s.

Carroll (1971) is more than five times the size of LOB and Brown, but contains texts drawn from a rather special area, i.e. "the range of required and recommended reading to which students are exposed in school grades 3 through 9 in the United States" (Carroll et al. 1971:xiii).

The only corpus which can be claimed to be both representative of contemporary English, to contain a variety of genres and to be comprehensive enough for our purposes is the huge collection of machine-readable texts collected at the university of Birmingham in the course of the work on the COBUILD dictionary. (For a presentation of the contents of the COBUILD database see Renouf (1987) and other contributions to Sinclair (1987).

There are several COBUILD databases. The Main Corpus, which was the main input to the dictionary, contains 7.3 million words. In addition there is the Reserve Corpus which contains roughly another 13 million words (cf. Renouf 1987:11 ff). Together the Main and Reserve Corpora comprise almost 18 million words of (mostly written), mainly British texts representing a broad spread of current language use. It is true that the COBUILD database also has certain drawbacks in comparison with, for example, the Brown and LOB corpora, in particular the lack of genre-indicators for the different texts. However, there is no doubt that the COBUILD material constitutes the largest and most up-to-date corpus of English presently available. Consequently an agreement was reached with the Research and Development Unit of Birmingham university under which the unit undertook to lemmatize the GYM corpus and the combined Main and Reserve Corpora. The lemmatization was carried out on a mainframe computer in Birmingham. The computer programs for the lemmatization were written by Jeremy Clear, at the time the Unit's Senior Computing Officer. In the absence of subsequent manual editing, the resulting lemmatized lists are naturally a halfway house between a simple list of word types and a real lemmatization, i.e. one involving homograph separation.

3 It is not immediately clear how a comparison at the level of vocabulary between large corpora should proceed. In the present case, the problem is also compounded by the great discrepancy in length between the two corpora, the GYM corpus containing roughly 1.5 million words as against the almost 18 million words of the COBUILD corpus.

Among the approaches that suggest themselves, a fairly obvious one is to find out how many of the words in each corpus are unique to that particular corpus. Here we have to proceed with caution. Clearly, once we leave the one or two thousand most frequent words in each corpus, the selection of individual words in either corpus is largely due to chance. However, if words from the top frequency band in one corpus are missing from the other, that is a fairly strong indication that the two corpora differ in ways which are not due to chance.

Another comparative technique is to study the differences in frequency among the words which are shared and to establish a measure which can be used to evaluate these differences. Such a comparison should also be restricted to the highest frequency band in each corpus.

Frequency counts can also be used to establish stylistic differences: a high incidence of contracted forms and of words like e.g. mum, dad is indicative of texts which are low on the formality scale. As will be shown later, frequency data also provides helpful information about the textual types contained in a corpus.

In addition to the studies of individual words just described, there are other important areas of comparison of a more collocational or even grammatical type. Quoting Winter 1978, Carter & McCarthy 1988 point to a number of words with discourse function, like e.g. achieve, explanation, method, point and problem. They refer to words like these as signposts which subsume large chunks of text and which, when appropriately arranged, can represent the entire macrostructure of a text.

The ability to understand and use such textual cues is an important part of full reading competence in a language. An important task in any evaluation of TEFL texts will therefore be to investigate whether these texts make use of such discourse signposts in the same way and to the same extent as the native English standard of comparison.

In their contribution to Carter and McCarthy (1988), Sinclair & Renouf point to the importance of teaching central patterns of usage in mainstream English. They stress in particular the importance of delexicality - "the tendency of certain commoner transitive verbs to carry particular nouns or adjectives, which can in most cases themselves be transitive verbs." (Sinclair & Renouf 1987:153).

Familiar verbs of this kind are e.g. *give*, *have*, *make*, *put* and *take*, which normally have little independent meaning and mostly occur in collocations like *give information*, *have a look*, *make a decision* etc. An obvious task for the evaluator of TEFL texts will be to find out whether the central patterns in the corpus serving as standard of comparison are present in like measure in the texts that are being evaluated.

4 In the present study, the emphasis is placed on the first two types of comparison discussed above, i.e. investigations of unique words and of differences in frequency between shared words.

A comparison between the 1,000 most frequent words in GYM and COBUILD reveals that within the confines of this frequency band, the two corpora have 796 words in common. A study of the 204 words which are not shared points to important differences between the corpora.

Among the words unique to the GYM corpus, the majority of the nouns denote physical objects, processes and actions. The verbs denote physically observable phenomena, most of them human actions, and the adjectives express either emotional judgement (*terrible*, *wonderful*), physical characteristics (*soft*, *bright*), or feelings (*angry*, *glad*).

It is also clear that there are certain themes which are predominant in the GYM texts, like e.g. family life, travel, etc. The following 25 words, which could be used to write a number of fairly simple-minded stories of a familiar kind, all share the distinction of appearing among the 1000 most frequent words in the GYM corpus, while being absent from the first 1000 COBUILD words: *accident*, *beach*, *boat*, *bus*, *Christmas*, *cigarette*, *driver*, *football*, *flight*, *holiday*, *island*, *passenger*, *plane*, *policeman*, *pub*, *restaurant*, *Saturday*, *sun*, *Sunday*, *sport*, *taxi*, *ticket*, *tea*, *traffic*.

When we turn to the 204 words found exclusively in the COBUILD material, the nature of the words becomes strikingly different. The majority of the nouns in the COBUILD material are either abstract, like e.g. *argument*, *decision*, *difficulty*, or interpretations of behaviour, like e.g. *action*, *activity*. Quite a few denote aspects of social organisation, like *authority*, *community*, *council*.

Several nouns denote dimensions used to evaluate physical objects, like *size*, *shape*, *amount*, *design*, *value*. Others are what appear to be political terms: *campaign*, *labour*, *tax*, *trade*, *vote*. The majority are difficult to characterize in terms of content areas, but share the characteristic of being abstract.

Most of the verbs express evaluations of human behaviour, like *achieve*, *argue*, *assume*, for example. Not many of the adjectives denote physical characteristics. The majority are relational, organisational or technical terms such as *apart*, *basic*, *central*, *industrial*, *international*, *nuclear*.

The comparison of unique words indicates that there are important differences between the corpora being investigated. The unique words in the GYM material are predominantly, and in the case of certain categories almost exclusively, concrete terms denoting physical objects and processes, physical characteristics and emotions. The words found only in the COBUILD material, on the other hand, are predominantly abstract.

5 After the COBUILD frequencies had been adjusted to account for the difference in corpus size[1], the GYM and COBUILD frequencies for the shared words were compared. The results of the comparison were expressed by means of the difference coefficient used in Hofland & Johanson 1982 and originally suggested in Yule 1944. The formula for computing the coefficient is

$$(GF- CBFADJ)/(GF+ CBFADJ)$$

where GF stands for the frequency found in the GYM corpus and CBFADJ represents the adjusted COBUILD frequency.

The coefficient, which ranges from 1.00 to -1.00, expresses the relative frequency of a lemma in the two corpora. A positive value indicates that the lemma is over-represented in the GYM corpus (in comparison with its frequency in the COBUILD corpus), while a negative value is an indication of under-representation.

The exact location of the cut-off points between over- and under-represented words is of course arbitrary. In the following Table, it has been assumed that words with coefficient values between 0.9 and -0.9 are more or less equi-represented in the two corpora. Values higher than 0.9 and lower than -0.9 have been taken as indications of over- and under-representation respectively. When applied to the 796 shared words, these principles imposed the classification found in Table 3.

Table 3. Over-, under- and equi-represented words in the GYM texts in comparison with COBUILD

Type of repres.	N	Coefficient
Words which are over-represented in GYM	361	0.56 - 0.10
Equi-represented words	260	0.09 - -0.09
Words which are under-represented in GYM	175	-0.10 - -0.54
Total	796	

The majority of the shared words are, naturally, lexical words, a study of which confirms the impression formed on the basis of the unique words. Rather than pursue this line of inquiry with its more or less predictable results, I will turn to a study of the closed class items.

Table 4 below presents some of the closed class items which are over-represented in the GYM material together with their difference coefficients.

Table 4. Some over-represented closed class items in the GYM corpus

Item	Diff. coeff.
contractions	+0.37
she	+0.35
nobody	+0.32
me	+0.22
you	+0.22
anything	+0.21
he	+0.21
everything	+0.20
here	+0.20
maybe	+0.20
I	+0.19
him	+0.18
yourself	+0.18
her	+0.16
mine	+0.16
nothing	+0.16
something	+0.15
your	+0.14

In Table 4 the item "contractions" is a cover term for all the contractions found among the 1,000 most frequent words in GYM and COBUILD. Contractions are by far the most over-represented category in GYM. The average number of contractions per 1,000 words is 18 for GYM, 9 for COBUILD.

To a certain extent this may simply be an indication that there is a far greater proportion of dialogue in the GYM texts than in the texts underlying the COBUILD lists, an assumption which receives further support from the high coefficients for the first and second person personal pronouns. It is probably also a measure of the stylistic difference between the two corpora, however.

I have no ready explanation for the preponderance of indefinite pronouns in _-thing_ in the school texts. The high incidents of the third person pronouns _she_ and _he_ - but not of _it_ - presumably reflects an interest in "human interest stories" and further strengthens our earlier conclusions about the kind of themes that are most widespread in the text books.

6 The frequent use of third person pronouns can also be linked to differences in text type. In a number of publications (e.g. 1986, 1988 and 1989), D. Biber has shown how a number of linguistic features can be used to place texts along a number of dimensions of textual typology. In the remainder of this paper I will attempt to show how certain of the criteria used in Biber 1986 can be used to establish textual differences between GYM and COBUILD.[2]

Biber regards third person pronouns as being among the criterial features for two typological dimensions. One of these is what, in his 1986 paper, he calls the dimension of _abstract vs. situated content_. Basically texts with abstract content are characterized by a "highly abstract, nominal content and a highly learned style" (1986:345), while texts of the situated type have concrete content and "a greater reliance on an external situation" (1986:346). Situated content texts make frequent use of third person pronouns and of place and time adverbs (among other things). Texts with abstract content, on the other hand, score high on prepositions, by-passives and certain conjuntcs and disjuncts (cf. Biber 1986:346).

The preponderance of third person pronouns in the GYM texts has already indicated that the GYM texts should be regarded as having situated rather than abstract content. Further support for this conclusion comes from the data set out in Table 5, which show that time and place adverbs are heavily over-represented in GYM.

Table 5. Over-represented time and place adverbs in GYM in comparison with COBUILD

Adverb	Diff. coeff.
here	+0.20
away	+0.20
down	+0.20
then	+0.19
again	+0.18
soon	+0.18
never	+0.17
outside	+0.15
today	+0.13
there	+0.10

A third and final source of evidence for the "situated" nature of the school texts is the distribution of prepositions in the two corpora. There are only three potential prepositions among the 361 over-represented words in GYM, i.e. inside, off and across. Of these, the last two are also frequently used as adverbs and as particles in phrasal verbs.

When we turn to the under-represented words, on the other hand, we find the nine prepositions in the following table.

Table 6. Under-represented prepositions in GYM in relation to COBUILD

Prep.	Diff. coeff.
within	-0.39
between	-0.20
by	-0.20
upon	-0.20
against	-0.16
of	-0.16
since	-0.12
than	-0.11
during	-0.10

As Table 6 indicates, prepositions are much less frequent in the "gymnasium" texts than in COBUILD, a fact which provides additional evidence of the "situated" character of the former.

Up till now we have concentrated on the content dimension and our findings have provided ample evidence that the GYM texts should be regarded as representing situated rather than abstract content, whereas the opposite is true of COBUILD.

But the third person pronouns are also instrumental in placing texts along a second dimension, which Biber refers to as that of reported vs. immediate style. Reported style is what we find in texts with a "narrative emphasis marked by a considerable reference to a removed situation", while immediate style is what we find in texts with non-narrative emphasis, more elaborate content and "immediate reference" (Biber 1986:346).

In Biber's view, reported texts are characterized by high concentrations of third person pronouns, verbs in the past tense and perfect aspect. Texts with immediate style, on the other hand, make frequent use of present tense verbs but have relatively few instances of third person pronouns, past-tense verbs and perfect aspect constructions.

The frequent use made of third person pronouns in the GYM texts has already given us reason to believe that these texts are of the reported rather than the immediate type. However, in order to make as strong a case as possible for categorizing the GYM texts as reported rather than immediate, we need more evidence pointing in the same direction. As the following table shows, it is in fact possible to obtain such evidence from the distribution of past tense forms in the two corpora.

Table 7. The use of past tense verb forms among the 1000 most frequent word types in GYM and COBUILD

	No of past tense forms	Accum. past tense freq.s
COBUILD	100	73727
GYM	116	98209

There is only a slight difference in the actual number of different past tense forms encountered in the two corpora. However, when it comes to the use made of these forms in the texts, the accumulated COBUILD frequencies amount to only 75% of those for the GYM corpus. (The COBUILD figures represent adjusted frequencies.)

As we have already seen, a high proportion of past tense forms is an indication of reported rather than immediate style. Adding this bit of evidence to that already obtained from the third person pronouns, we are now in a position to claim with a certain amount of confidence, that the GYM and COBUILD texts differ also along the dimension of reported vs. immediate style, with the former oriented towards the reported end of the dimension, and the latter towards the immediate.

7 In this short study I have demonstrated some of the techniques that can be used in evaluating TEFL texts. I would like to conclude my study by stressing three points. One is that it is in fact possible to obtain a fair amount of information about texts merely by looking at word lists.

The second point is that the closed class items have turned out to be as revealing as - or perhaps even more revealing than - those from the open word classes. A natural extension of the limited study of closed-class items conducted in this study would be a full-scale study of the texts in the two corpora in terms of the text-typological models of Biber 1988 and 1989.

The third and final point has to do with the conclusions that can be drawn from the results of the comparison between the TEFL material and the COBUILD texts. Finding texts which are both linguistically satisfactory and suitable for the age-range involved here is never easy. It will always involve an uneasy compromise between proficiency goals on the one hand, and an assessment of the kind of texts the majority of the students can reasonably be expected to read on the other.

However, such a balance does not seem to have been struck in the TEFL texts under investigation. There is heavy over-emphasis on concrete and uncomplicated matters and a dearth of abstractions and words relating to the organization of society. There are also indications that the texts included tend to be of a fairly simple, narrative kind.

But an understanding of abstractions, societal terms and non-narrative genres is precisely what many students will need, once they have left school. It is a prerequisite for activities like reading (quality) newspapers, reading and producing reports and manuals and following newscasts on the media. It is hardly unreasonable to demand that the TEFL texts used in the final three years of a total of nine years of English studies should prepare the students for these tasks.

Notes

1. The adjusted COBUILD frequencies were calculated in accordance with the following formula:

$$\text{(CB-freq/CB-n) * GYM-n = CB-adj}$$

Here CB-freq denotes the original COBUILD frequency, CB-n the total number of words in the COBUILD database, GYM-n the number of words in the GYM corpus, and CB-adj the adjusted COBUILD frequency.

2. The argument here is based on Biber 1986 and does not take account of his subsequent development of the model.

BIBLIOGRAPHY

Biber, Douglas. 1986. "Spoken and written textual dimensions in English". Language 1986.

Biber, Douglas. 1988. Variation across speech and writing. Cambridge. Cambridge University Press.

Biber, Douglas. 1989. "A typology of English texts". Linguistics 27.

Carroll, B. et al. 1971. The American Heritage Dictionary. Boston.

Carter, Ronald & Michael McCarthy. 1987. "Lexis and discourse: vocabulary in use". In Carter & McCarthy (eds.) 1987.

Carter, Ronald & Michael McCarthy (eds.). 1987. Vocabulary and Language teaching. London. Longman.

Hofland, Knud & Stig Johansson. 1982. Word freqencies in British and American English. Bergen. The Norwegian Computing Centre for the Humanities.

Ljung, Magnus. 1990. A Study of TEFL Vocabulary. Stockholm Studies in English 78. Stockholm: Almqvist & Wiksell International.

Renouf, Antoinette. 1987. "Corpus Development". In Sinclair (ed.) 1987.

Sinclair, John (ed.). 1987. Looking Up: An Account of the COBUILD Project in Lexical Computing. London. Collins.

Sinclair, John & Antoinette Renouf. 1987. "A Lexical Syllabus for Language Learning". In Carter & McCarthy (eds.) 1987.

Winter, E.O. 1978. "A look at the role of certain words in information structure". In Jones & Horsnell 1978.

Yule, George U. 1944. The Statistical Study of Literary Vocabulary. Cambridge. Cambridge University Press.