Vocabulary Frequency in TSIII: An Examination of Television News Transcripts

Tamara Swenson

トピックスタディーズIIIにおける語彙の頻度；TVニュースの文字化による検証

タマラ・スェンソン

Abstract

This paper examines the vocabulary usage of Topic Studies III television news transcripts from materials used from 1993 to 1998 to determine the frequency of the vocabulary used. It is hoped this information will help further understanding of the terms students need in order to be successful in this course.

Key words: vocabulary, frequency, concordance, news

(Received September 6, 1999)

抄　　録

本稿では、1993－1998年度のトピックスタディーズIIIにおいて使用したTVニュースに現われた語彙の頻度を検証する。これによって期待できることは、学生がこのクラスで必要とされる語彙をさらに理解できるようになることであろう。

キーワード：語彙、頻度、コンコーダンス、ニュース

(1999年9月6日　受理)

Introduction

The Topic Studies III course at Osaka Jogakuin Junior College was first implemented in 1988 when the then "new" curriculum was adopted, and retained following the college's 1998 curriculum revision. All second-year students are required to take, and pass, two-terms of the course in order to graduate. It continues to be, according to numerous post-graduate questionnaires, one of the college's most appreciated courses despite its difficulty.

In brief, TSIII is a current news course which covers a single current events topic during two 70-minute periods. Classes meet approximately 20 times each term. (Prior to the 1998 curriculum revision, the course was a full-year course held two 50-minute periods during the 26-week school year.)

Instructors, on a rotating basis, select a news topic from commercially broadcast television reports. The selected news story, about a current event, an economic trend or problem, or an on-going news issue, is between two and three minutes long. This broadcast is then transcribed, a print article about the topic is selected, and a series of comprehension worksheets, vocabulary lists, discussion questions, and quizzes are prepared. These materials are used by all classes.

In spite of its prominence within the OJJC curriculum, little has been done to analyze the course's contents. One aspect of the course that warrants examination is the vocabulary which appears in the television news broadcasts. As all OJJC students are required to take TSIII, understanding the vocabulary most likely to appear during the course is not an unrealistic expectation. In other words, what vocabulary should OJJC students be expected to know at the end of the first-year in order to be more successful in TSIII.

To this end, I decided to build and analyze a concordance of the vocabulary used in the news broadcasts selected by TSIII instructors.

Procedure

TSIII news broadcast transcripts used during the last six complete years, 1993 to 1998, were first collected. As no computer version existed for some transcripts, those from 1993 through 1997 were then scanned using a Canon CanoScan 600 scanner and the E. Typist version '97 (1997) computer program. This program converted the digital image into text files. The resulting texts of the transcripts scripts were then compared to the original versions and corrections made to ensure the scanned text matched the original transcript. The scanned text files and the text files of the 1998 transcripts were then combined into files containing the transcripts for each year's

news broadcasts and a file containing all transcripts from the six-year period. Concordances of these files were then built using Conc version 1.80 (Thomas & Hatton, 1996) to arrive at the corpus of TSIII vocabulary.

These files were then indexed by word frequency. These word frequency lists were then transferred into Microsoft Excel98 files (1998) and the lists sorted by frequency. These lists were then edited to eliminate the line number attached by Conc and errant characters misinterpreted as words (i.e. quotation marks and apostrophes). Lists were also edited to combine singular and plural instances of the same word (e.g. war, wars), different forms of the same verb (e.g. take, takes, took, taken), comparative and superlative forms of a word (e.g. short, shorter, shortest), and possessives (e.g. company, company's).

These sorted lists were then transferred into Microsoft Word98 (1998). After building a list based on the frequency of all words, it was decided that for the purposes of building a list appropriate for use with first-year OJJC students, further editing was needed. This necessitate the generation of three different lists. The first was the unedited list of words in order of frequency regardless of word type. The second list followed Bauman's (1999a) compilation standards for the General Service List (GSL), which meant that different forms of a word were combined. The third list was developed in order to provide a basic list of the frequent content words in the TSIII transcripts. For this list, it was decided to eliminate pronouns, prepositions, articles, conjunctions, modals, cardinal and ordinal numbers, negative markers, interrogatives, honorific titles, months, and days of the week. Also eliminated were the verbs be, have, do, go and say in all forms. In addition, non-English words (e.g. *nyet, shigataganai*), proper nouns (i.e. names), and acronyms were eliminated because they generally appeared in only one broadcast over the six years examined. Speech markers such as 'ah' were also deleted from the word list. The resulting list, though much shorter, was then heavily weighted toward content words (i.e. nouns and verbs). This content list was felt to more accurately indicate the types of vocabulary students would need to know to be successful in TSIII.

The lists were then compared to Bauman's version of the GSL (1999b) of English word frequency, revised from the list originally developed by West in 1953.

Results

First, overall statistics regarding the word frequency in the six years under consideration indicate that 35,971 words were used, of which 5,547 were different words. The highest number of words appeared in 1997 (6,591) when 18 topics were covered, the lowest in 1998 (4,928) when 14 topics were used (see Table 1).

Table 1: Word Count by Year

Year	Number of topics	Number of words
1993	19	6,546
1994	18	5,782
1995	18	5,927
1996	18	6,197
1997	18	6,591
1998	14	4,928

The overall frequency list was then examined to see which words were the most common overall (see Table 2). The most frequent word, "the" occurred 2,024 times. No other word had more than 1,000 instances over the six years examined. The next five most frequent word were "to" (941 times), "of" (809), "and" (716), "a" (688), and "in" (632). The twenty most frequent words in the unedited list appear in Table 2. This list was not edited to combine various forms of a word at this point.

Table 2: The 20 Most Frequent Words from the Unedited List

Rank	Word	Frequency Count	Rank	Word	Frequency Count
1	the	2024	11	on	248
2	to	941	12	it	236
3	of	809	14	they	223
4	and	716	14	this	212
5	a	688	15	was	211
6	in	632	16	have	202
7	that	392	17	with	198
8	is	388	18	as	193
9	for	298	19	but	175
10	are	257	20	not	160

As with the GSL (Bauman, 1999b), which examined a much larger corpus, the most frequent word was "the" in the TSIII transcripts. Bauman's number two word, "be" was ranked 21st on the overall list of frequency, but when all forms of the verb were combined, following the GSL compilation guidelines, the frequency count was 986, enough to make this verb second on the overall frequency list. When revised to combine forms of the same word, the TSIII word frequency more closely, though not identically, mirrors the frequency ranking on the GSL (see Table 3).

Two words from the 20 most frequent on the GSL list were missing from the top 20 of the TSIII list edited following the GSL compilation standards. These words were

Table 3: Comparison of GSL and TSIII (Combined) Word Frequency Rank

GSL Rank	TSIII Word	Count	TSIII Rank
1	the	2024	1
2	be	986	2
6	to	941	3
3	of	809	4
5	a	807	5
4	and	716	6
7	in	632	7
9	have	431	8
11	that	392	9
10	it	382	10
13	they	325	11
12	for	298	12
18	on	248	13
22	this	212	14
8	he	211	15
15	with	198	16
16	as	193	17
26	but	175	18
17	not	160	19
20	at	152	20

"I" (GSL 14; TSIII 30) and "she" (GSL 19; TSIII 69).

The examination of the edited word list revealed that the most frequently appearing "content" words were "today" and "year" (117 each), followed by the verb "say" (111). No other word recorded more than 100 instances over the six years examined. The next 10 most frequent words were "now" (91), "all" (90), "people" (88), "government" (86), "sail" (81), "about" (79), "new" (74), "president" (71), "country" (67), and "some" (67). A list of the most frequent words, the 374 words with at least 10 instances on the edited list of content words, appears in the Appendix.

Discussion

As expected, the most frequent words used in the TSIII news broadcasts during the six years surveyed were those which also appeared near the top of the GSL list, primarily articles, prepositions, conjunctions, and high frequency verbs such as "be" and "have." While this group of words are essential for OJJC students' overall English

ability, it also represents primarily grammatical markers that are not the focus of the TSIII course.

As the focus of the course is on understanding world events, and improving English listening and discussion ability, the content list (Appendix) provides far more information about the words students should be expected to know in order to be successful in the TSIII course. Examination of this list reveals that many of the words are those which appear in high school English material (e.g. people, government, time, child). However, some occur far less often and should be considered as words that are essential to cover during the first-year curriculum (e.g. minister – 36, policy – 32, military – 29, representation – 26, violence – 24).

Usage of some of these words varies from how they are currently introduced the in first-year curriculum. For instance, "minister" appears in the TSIII curriculum in reference to "prime minister," "cabinet minister," and other such government positions. However, in the first-year courses, it is introduced in Unit 2 in reference to the leader of Protestant denominations. Introducing students to alternate definitions of words is one concern that needs to be addressed. In other words, this list deserves further consideration during future revisions of the first-year curriculum. Closer examination of the list generated, and comparison of it to both the vocabulary expected for Japanese high school students and that used in first-year courses, would indicate which of the words are necessary to include and alternate definitions of words that need to be covered.

In addition, there were some surprises among those words occurring 9 times, words in first 500 in frequency on the TSIII edited list.[1] These included "ammonia," "clone," "discrimination," and "nicotine," all words with far less frequency on the GSL. This, too, deserves further examination as a list of essential words is developed.

At the other end of the frequency list, a number of words with much higher frequencies on the GSL occurred only once. These included, among many others, "admire," "brain," "coat," "favorite," "lesson," "owe," "police," "shine," and "wind."

Conclusion

The examination of the TSIII corpus seems to indicate that the lists developed for other sources, such as the GSL, are inappropriate for the purposes of preparing OJJC students for Topic Studies III. As Sauvignon (1983) so aptly pointed out, commercially-available materials, because they are written for a general audience, cannot meet the needs of all teachers and learners. Obviously, a TSIII specific vocabulary list, one which eliminates commonly occurring words already learned, would benefit learners.

As such, the development of the concordance of words used in TSIII transcripts from 1993 to 1998 is only the first step in a much larger project. A complete concordance covering all transcripts used in every year of the course needs to be created. In addition, the articles used in the course, as well as the additional work-sheets and quiz materials developed for use in the course, need to be added to this concordance. Furthermore, all lists need to be analyzed more closely to determine which of the vocabulary items need to be covered during the first year of study at OJJC and the context in which these items should be taught. These projects will bring us closer to understanding what vocabulary items are essential for every student to be successful in TSIII.

Note

1 Because of space requirements, the more than 5,000 words with less than 10 occurrences are not included in the Appendix. A complete list of the words found in this examination of the TSIII corpus can be obtained from the author upon request.

References

Bauman, J. (1999a). About the general service list. Internet , July 10, 1999, 〈http: plaza3.mbm.or.jp/~bauman/aboutgsl.html〉.

Bauman, J. (1995b). The actual 2.284 words, with frequency numbers. Internet , July 10, 1999, 〈http: plaza3.mbm.or.jp/~bauman/gsl.html〉.

E. typist version '97 [Computer program]. (1997). Tokyo: Media Drive Corporation.

Microsoft Excel98 [Computer program]. (1998). Seattle, WA: Microsoft Corp.

Microsoft Word98 [Computer program]. (1998). Seattle, WA: Microsoft Corp.

Sauvignon, S. (1983). *Communicative competence: Theory and classroom practice. Texts and contexts in second language learning.* Reading, MA: Addison-Wesley.

Thomas, J. & Hatton, J. (1996). Conc version 1.80, beta 3 [Computer program]. NY: Summer Institute of Linguistics.

Appendix A: Content Words with 10 or More Instances (N = 374)

Word	Count
today	117
year	117
say	111
now	91
all	90
people	88
government	86
make	83
sail	81
about	79
new	74
president	71
country	67
some	67
get	66
just	65
time	65
world	62
over	61
last	60
many	57
other	57
take	57
day	51
little	49
think	49
come	47
like	47
good	46
still	44
city	43
peaceful	43
know	42
official	42
only	42
kill	41
nation	41
tonight	41
call	40
live	40
begin	39
man	39
leader	38
because	37
child	37
minister	36
news	36
political	36
work	36
another	35
most	35
student	35
use	35
much	34
north	34
want	34
against	33
company	33
vote	33
even	32
force	32
give	32
old	32
policy	32
state	32
war	32
way	32
law	31
try	31
build	30
week	30
back	29
military	29
seed	29
talk	29
late	28
report	28
tobacco	28
already	27
change	27
end	27
off	27
ago	26
down	26
month	26
nuclear	26
prince	26
representation	26
attack	25
believe	25
case	25
family	25
found	25
long	25
look	25
party	25
any	24
bomb	24
early	24
group	24
high	24
home	24
market	24
violence	24
car	23
great	23
issue	23
need	23
sheep	23
thing	23
troop	23
woman	23
action	22
dollar	22
every	22
fight	22
life	22
right	22
whale	22
become	21
big	21
hour	21
international	21
million	21
money	21
see	21
south	21
stop	21
through	21
around	20
computer	20
consider	20
continue	20
job	20
leave	20
meet	20
such	20
charge	19
decision	19
die	19
find	19
fire	19
grow	19
hold	19
house	19
kid	19
place	19
then	19
train	19
workers	19
young	19
death	18
debate	18
election	18
evening	18
few	18
general	18
hard	18
help	18
mother	18
rule	18
sanctuary	18
shoulder	18
strong	18
test	18
turn	18
weapon	18
well	18
while	18
white	18
agreement	17
allow	17
area	17
ask	17
black	17
council	17
hope	17
include	17
never	17
percentage	17
plant	17
problem	17
setback	17
trade	17
administration	16
agree	16
capital	16
crime	16
deal	16
far	16
feel	16
future	16
industry	16
night	16
open	16
own	16
partial	16
powerful	16
process	16
sexual	16
air	15
base	15
clear	15
control	15
despite	15
enough	15

foreign ·····15
free ·····15
half ·····15
keep·····15
lead ·····15
member ·····15
national ·····15
negotiation ·····15
put ·····15
return ·····15
secret ·····15
story ·····15
town ·····15
united ·····15
university ·····15
victim ·····15
yesterday ·····15
again ·····14
break ·····14
close ·····14
dead·····14
expect·····14
face ·····14
fear ·····14
happen ·····14
hospital ·····14
important ·····14
limit·····14
mean ·····14
question·····14
rear ·····14
start·····14
themselves ·····14
aid ·····13
animal·····13
away ·····13
battle ·····13
bring ·····13
fact ·····13
heavy ·····13
human ·····13
instead ·····13
lose ·····13
lot ·····13
often ·····13
order ·····13
planes·····13
racial ·····13
spend ·····13
street ·····13
study ·····13
support ·····13
threaten·····13
act ·····12
advertising
(ads) ·····12
affirmative ·····12
airbag·····12
bad ·····12
ban ·····12
both·····12
business·····12
care·····12
cause ·····12
chairman ·····12
cigarette·····12
crowd ·····12
drink ·····12
drop·····12
effect ·····12
ever·····12
former·····12
health ·····12
increase ·····12
involve ·····12
message ·····12
morning·····12
northern·····12
parliamentary ···12
population·····12
produce ·····12
protestant ·····12
safe ·····12
school ·····12
schoolboy ·····12
severe·····12
sniper ·····12
soon·····12
stay ·····12
sure ·····12
surrogate ·····12
television ·····12
warning ·····12
accept ·····11
almost·····11
appear·····11
body ·····11
buy ·····11
commercial ·····11
direct ·····11
dozen ·····11
each·····11
economic ·····11
gun ·····11
harassment ·····11
kind·····11
local·····11
nothing ·····11
number ·····11
offer·····11
part ·····11
patch ·····11
poll ·····11
position ·····11
publicity ·····11
quiet ·····11
relationship ·····11
retrained ·····11
send·····11
soldier·····11
strike ·····11
target ·····11
though ·····11
until ·····11
able ·····10
according ·····10
add ·····10
affect ·····10
aid ·····10
announce ·····10
army ·····10
blame ·····10
chip·····10
coalition·····10
community ·····10
credit ·····10
decide·····10
demand ·····10
destroy ·····10
effort ·····10
event ·····10
everything·····10
facility ·····10
girl ·····10
history ·····10
inspection·····10
interest ·····10
itself ·····10
land·····10
large ·····10
line ·····10
major ·····10
medical ·····10
move ·····10
operation ·····10
overseas·····10
product ·····10
program·····10
propose ·····10
remain ·····10
sign ·····10
signal ·····10
sit·····10
smile ·····10
speak ·····10
wait·····10
whether ·····10
word ·····10
yet ·····10