アメリカ英語の統計的研究（1）

清　川　英　男

Ⅰ．この調査の目的

語いの統計的研究は数多く行なわれているが，その中で最も有名な研究は Thorndike, E. L. and Lorge, I. (1944)[1] であろう。その後この研究方法をモデルとして多くの研究がなされたが，最近の研究としては，成人用読書材料をサンプルとした Kučera and Francis (1967)[2]，3年生から9年生までの種々の教科書をサンプルとして分析した Carroll et al. (1971)[3] があげられる。

Thorndike の語い統計の研究はそれ自身価値があり，この研究結果は教育，とくに国語教育および日本における外国語教育に広く応用されている。Kučera and Francis (1967) は頻度の他にレンジを求め，それを並記してあるが，Carroll et al. (1971) は頻度の他に情報理論にもとづいて計算されたU値などでランクをつけている点に研究の特長がある。この2つの研究はきわめて貴重な研究であり，これらの研究結果は Thorndike と並んで広く利用されるであろうと考える。しかし，これらの研究は homographs（綴りは同一であるが，発音や意味が異る語）の区別を明らかにしていないこと，homologs（綴りと発音は同一であるが，意味が異る語）の区別もしていないこと，が欠点である。

本研究は，一連の語い統計の一部であるが，最終的には，話しことばを含めて現代のアメリカ英語の種々のジャンルの英語を統計的に分析し，語い統計をまとめ，その結果から基本語を選出することをねらっている。さらに，homographs の最終的なまとめを行い，品詞別の使用頻度を調査し，教科書，副教材の編集などに役立つ資料を提供しようとするものである。

基本語を選定することは，実用的な見地から言えば，上記のような教材の選定，編集やカリキュラム，到達目標の立案などに役立つが，学問的には，リーダビリティ公式を作成するための最も基本的な資料ともなるので，それもこの調査の目的である。

Ⅱ．調査方法と内容

語い統計のサンプルは，広いジャンルにわたる必要がある。これからの一連の研究は，いくつかの種類にわたってサンプルをとる予定であるが，今回は *Newsweek* の 1978 年 1 月 9 日号から 4 月 3 日号にわたって，U. S. Affairs の中から毎号 300 ないし 450 語を抜き出した，合

計5368語をサンプルとした。抽出する単位は，センテンスごとでは意味のつながりが切れることがあるので，パラグラフを一単位とし，連続した4ないし5パラグラフをひとつのグループとした。

単語の単位は lexical unit ではなく，Kučera and Francis (1967) にならって変化形もそれぞれ異った語と数えた。ハイフンでつながれた語は一語とみなした。記号，数字はいずれも一語と数えた。品詞の区別はすべてカードに記入したが，Kučera and Francis (1967) と比較する便宜上，本論のための集計にあたっては，二つ以上の品詞を持つ，同じ語形の単語は，同一の単語としてカウントした。

本研究では，上記の言語サンプルを一語づつカードに書き抜き，品詞をそえ，これを基礎資料とした。これを集計，分析して，次の項目についてチェックした。

（1） 使用頻度の分布状態
（2） 頻度と順位の関係
（3） 頻度と延べ語数の累積頻度の関係
（4） 累積異語のサンプルに対する割合
（5） 異語数/延べ語数の割合
（6） 頻度の高い語の選出
（7） 延べ語数の品詞別の構成比

Ⅲ. 結果と考察

（1） 使用頻度とその分布

表1は，総延べ語数の0.1%以上を占める頻度を持つ128語の異語を使用頻度の高いものから順に並べ，さらにその異語の頻度の累計と累積百分率を求めたものである。表2はそれ以下のランクを求つ異語に関するデータである。

表 1 頻度の分布

RANK		X	SUM FX*X	CUM% FX*X	RANK		X	SUM FX*X	CUM% FX*X
1	the	360	360	6.706	10	for	49	1270	23.658
2	and	172	532	9.910	11	is	48	1318	24.553
3	to	162	694	12.928	12	on	39	1357	25.279
4	of	139	833	15.517	13	was	39	1396	26.006
5	a	128	961	17.902	14	who	33	1429	26.621
6	in	84	1045	19.467	15	with	32	1461	27.217
7	he	61	1106	20.604	16	it	30	1491	27.776
8	that	60	1166	21.721	17	but	28	1519	28.297
9	his	55	1221	22.746	18	by	28	1547	28.818

RANK		X	SUM FX*X	CUM% FX*X	RANK		X	SUM FX*X	CUM% FX*X
19	Carter	27	1574	29.322	64	says	10	2232	41.580
20	be	25	1599	29.787	65	there	10	2242	41.766
21	are	24	1623	30.235	66	what	10	2252	41.952
22	from	23	1646	30.663	67	will	10	2262	42.139
23	at	21	1667	31.054	68	$	10	2272	42.325
24	have	21	1668	31.446	69	first	9	2281	42.493
25	has	20	1708	31.818	70	general(G)	9	2290	42.660
26	had	19	1727	32.172	71	only	9	2299	42.828
27	one	19	1746	32.526	72	she	9	2308	42.996
28	said	19	1765	32.880	73	so	9	2317	43.163
29	last	18	1783	33.215	74	time	9	2326	43.331
30	they	18	1801	33.551	75	were	9	2335	43.499
31	an	17	1818	33.867	76	year	9	2344	43.666
32	him	17	1835	34.184	77	back	8	2352	43.815
33	most	17	1852	34.501	78	how	8	2360	43.964
34	their	17	1869	34.817	79	Humphrey	8	2368	44.113
35	as	16	1885	35.115	80	I	8	2376	44.262
36	or	16	1901	35.414	81	if	8	2384	44.411
37	president(P)	16	1917	35.712	82	its	8	2392	44.560
38	would	16	1933	36.010	83	Nixon	8	2400	44.709
39	more	14	1947	36.270	84	off	8	2408	44.858
40	this	14	1961	36.531	85	old	8	2416	45.007
41	into	13	1974	36.773	86	park	8	2424	45.156
42	national	13	1987	37.016	87	people	8	2432	45.306
43	new	13	2000	37.258	88	three	8	2440	45.455
44	than	13	2013	37.500	89	trees	8	2448	45.604
45	coal	12	2025	37.724	90	up	8	2456	45.753
46	miners	12	2037	37.947	91	White	8	2464	45.902
47	now	12	2049	38.171	92	befove	7	2471	46.032
48	out	12	2061	38.394	93	Congress	7	2478	46.162
49	work	12	2073	38.618	94	government	7	2485	46.293
50	against	11	2084	38.823	95	health	7	2492	46.423
51	all	11	2095	39.027	96	Koch	7	2499	46.554
52	CIA	11	2106	39.232	97	may	7	2506	46.684
53	house(H)	11	2117	39.437	98	party	7	2513	46.814
54	no	11	2128	39.642	99	still	7	2520	46.945
55	over	11	2139	39.847	100	then	7	2527	47.075
56	some	11	2150	40.052	101	after	6	2533	47.187
57	week	11	2161	40.257	102	also	6	2539	47.299
58	when	11	2172	40.462	103	become	6	2545	47.411
59	about	10	2182	40.648	104	campaign	6	2551	47.522
60	been	10	2192	40.835	105	can	6	2557	47.634
61	could	10	2202	41.021	106	Colson	6	2563	47.746
62	Hollywood	10	2212	41.207	107	contract	6	2569	47.765
63	not	10	2222	41.393	108	director	6	2575	47.969

RANK		X	SUM FX*X	CUM% FX*X	RANK		X	SUM FX*X	CUM% FX*X
109	down	6	2581	48.081	119	strike	6	2641	49.199
110	early	6	2587	48.193	120	through	6	2647	49.311
111	even	6	2593	48.305	121	two	6	2653	49.423
112	far	6	2599	48.417	122	union	6	2659	49.534
113	get	6	2605	48.528	123	Washington	6	2665	49.646
114	Jimmy	6	2611	48.640	124	Watergate	6	2671	49.758
115	other	6	2617	48.752	125	we	6	2677	49.870
116	seemed	6	2623	48.864	126	where	6	2683	49.981
117	Senate	6	2629	48.938	127	without	6	2689	50.093
118	state(S)	6	2635	49.087	128	years	6	2695	50.205

表 2 129 位以下の頻度の分布

RANK	X	FX	SUM FX*X	CUM% FX*X
129～ 163	5	35	2870	53.46
164～ 220	4	57	3098	57.71
221～ 328	3	108	3422	63.75
329～ 575	2	247	3916	72.95
576～2027	1	1452	5368	100.00

X：頻度
FX：Xの頻度を持つ異語の数
SUM FX*X：頻度Xの延語数およびそれ以前のランクのXの合計
CUM% FX*X：SUM FX*Xのサンプル全体に対する割合

この表から次のことがわかる。

a. 異語は 2,027 語である。

b. ランクが 127 位，すなわち延べ語数の累計が2689語でサンプルの50％を越える。

c. 1回のみあらわれる異語は1452語もあり，サンプル全体の 27.05％，異語全体の71.6％に達する。

d. 次の語は *Newsweek* あるいは時事英語のみにみられる現象であろう。 Carter (19位)，president(P) (37位)，coal (45位)，miners (46位)，CIA (52位)，Hollywood (62位)，Humphrey (79位)，Nixon (83位) など。

TTR

異語数の延べ語数に対する割合（ＴＴＲ）を計算し，他の言語サンプルと比較したのが表３である。これによると，ＴＴＲは 0.378 である。Kučera and Francis の total corpus の TTR が 0.050であり，大学生にスピーチをさせ，その語い統計を調査した Black and Ausherman (1955)[4] のＴＴＲが 0.024 であるが，これと比べるときわめて高い数値である。この数値は，Sample A21 の数値に近いことから，このように高い割合を示すのは，サンプル数が小さいためかもしれない。しかしこれだけのデータからは断定できない。

表 3 サンプル中の異語数／延べ語数

サンプル名	異語数	延べ語数	TTR
Newsweek	2,027	5,368	0.378
Kučera & Francis (Sample A21)	883	2,002	0.441
Kučera & Francis (Total Corpus)	50,406	1,014,232	0.050
Black & Ausherman	6,826	288,152	0.024

最も頻度の高い語の比較

次に，最も頻度の高い語20語を Kučera and Francis (1967) および Black and Ausherman と比べてみることにする(表4および表5)。表4では，20語のうち17語までが両リストに共通

表 4 最も頻度の高い 20 語の比較 (1)

RANK	Present study	Kučera and Francis
1	the	the
2	and	of
3	to	and
4	of	to
5	a	a
6	in	in
7	he	that
8	that	is
9	his	was
10	for	he
11	is	for
12	on	it
13	was	with
14	WHO	AS
15	with	his
16	it	on
17	BUT	be
18	by	AT
19	CARTER	by
20	be	I

大文字の語はどちらか一方にしか現れない語

表 5 最も頻度の高い 20 語の比較 (2)

RANK	Present study	Black & Ausherman
1	the	the
2	and	and
3	to	of
4	of	to
5	a	a
6	in	in
7	he	that
8	that	is
9	HIS	it
10	for	THEY
11	is	YOU
12	on	THIS
13	was	WE
14	WHO	HAVE
15	WITH	ARE
16	it	was
17	BUT	be
18	BY	he
19	CARTER	for
20	be	on

大文字の語はどちらか一方にしか現れない語

に現われている。また6位までの語は，順位は入れ替るが，同じ語であることは興味ある事実である。なお，両リストの片方にしか現れない語（大文字の語）について比べてみると，Kučera and Francis (1967) の as, at, I は本研究ではそれぞれ35位（頻度16），23位（頻度21），80位（頻度8）となっている。本研究の who, but は，Kučera and Francis (1967)では，それぞれ46位と25位である。さらに49位（本研究では50位が9語並んでいるために49位までとした）までを比べてみると，このうち共リストに共通している語は 35 語，すなわち 71.4

%であった。

上記の20語と Black and Ausherman (1955) の20語とを比較する(表5)と，20語のうち14語が共通である。また6位までの語は表4と同じように，同じ語であり，to と of の順位が入れ替っているだけである。表5 において，14 語しか共通していないのは，ひとつが，written English，もうひとつが spoken English の言語サンプルであるためであろう，と推察されるが，これだけのデータでは結論は出せない。

英語の単語の標準曲線

図1は，本研究，Black and Ausherman, Kučera and Francis のそれぞれの語い統計における異語の使用頻度と順位の関係を両対数方眼紙にプロットしたものである。Miller (1951)[5] は Zipf, G. K. の *The Psycho-biology of Language* を引用して，英語の単語の標準曲線

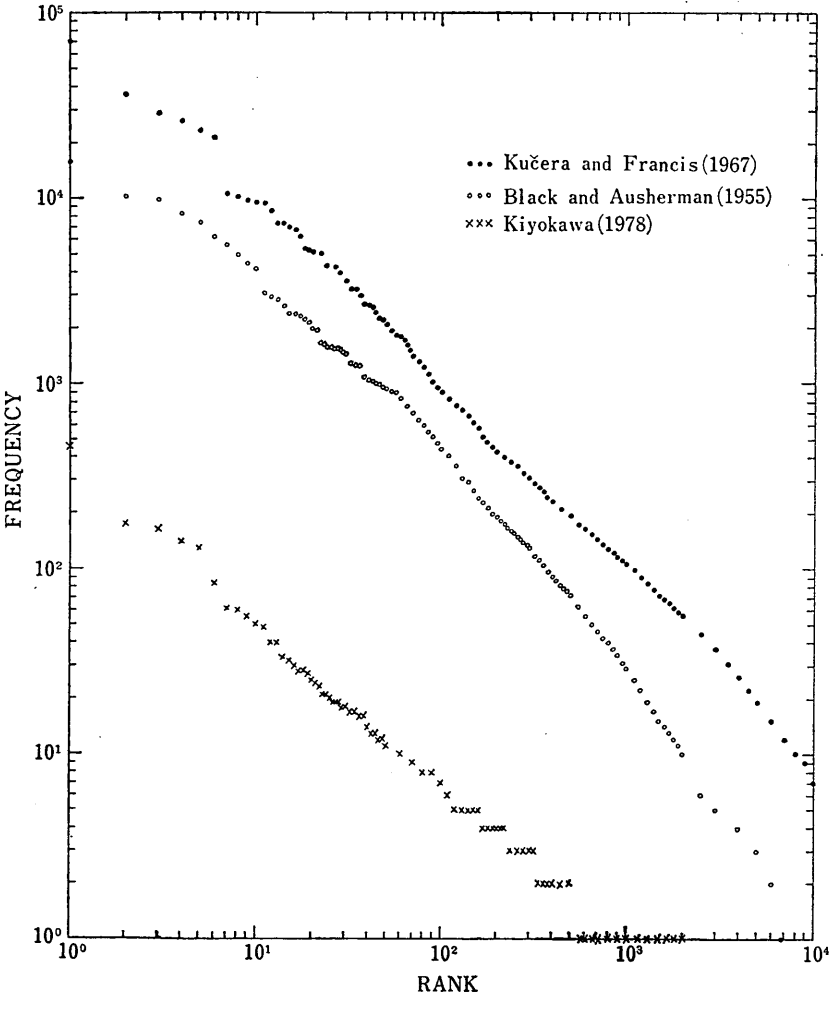

図1 語頻度と順位

(standard curve for English words) と呼ばれる線を紹介している。これは，語い統計において異語の使用頻度(f)とその順位(r)の積は一定(C)，すなわち，

$$f \times r = C$$

の関係があるというものである。

図1では，Kučera and Francis (1967) はrが10以上ではほとんど一直線でかつ勾配も45度に近い。本研究でもrが10以上100まではほぼ一直線で，かつ勾配も45度に近い。さらに，rが1から9までは，Kučera and Francis も本研究も同じような曲線を描いていることは興味深い。Black and Ausherman (1955) もほぼ一直線で，かつ勾配も45度に近い。また，rが1から9までは他の二曲線とかなり似かよった曲線である。

累積異語のサンプルに対する割合

異語を頻度順に並べ，その頻度を累積し，その延べ語数がサンプルに対してどの位の比率を占めるか，をグラフにしたものが図2である。それと同時に Kučera and Francis (1967) との比較を試みた。このグラフによると，二つの線はほぼ同じ曲線を示し，順位が115位までは

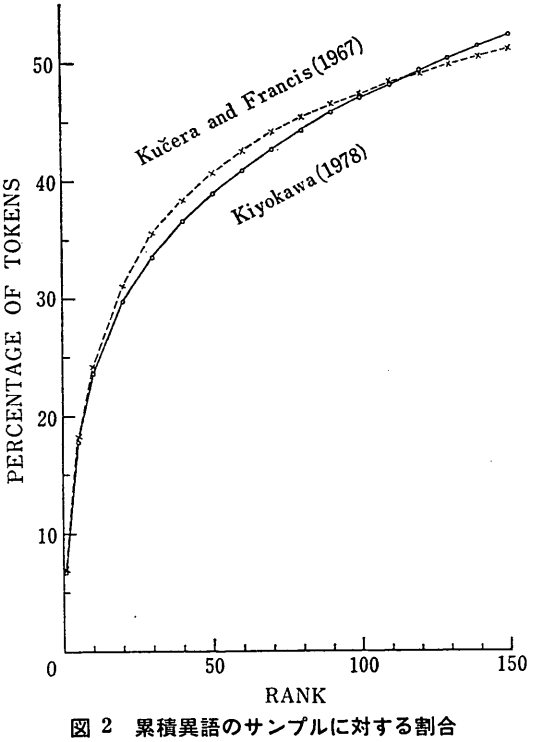

図 2 累積異語のサンプルに対する割合

Kučera and Francis (1967) が少し比率が多い。116以上は本研究のサンプルが，Kučera and Francis を上まわっていて，この比率の差は少しづつ大きくなっていく。また，Kučera and Francis は135語の異語で全サンプルの50％に達するのに対し，本研究では127語で50％を越

えている。

以上のことから，written English においては，最も頻度の高い約130語で読書材料の約50％の単語をカバーできる，といえるようである。また，時事英語においても同様のことが言えるようである。

頻度と延べ語数の累積頻度

頻度を小さい順に並べ，その異語のもつ頻度の延べ語数を累積した数字がサンプルの中で占める割合を求めたのが図3である。この図では，Kučera and Francis (1967) と Black and

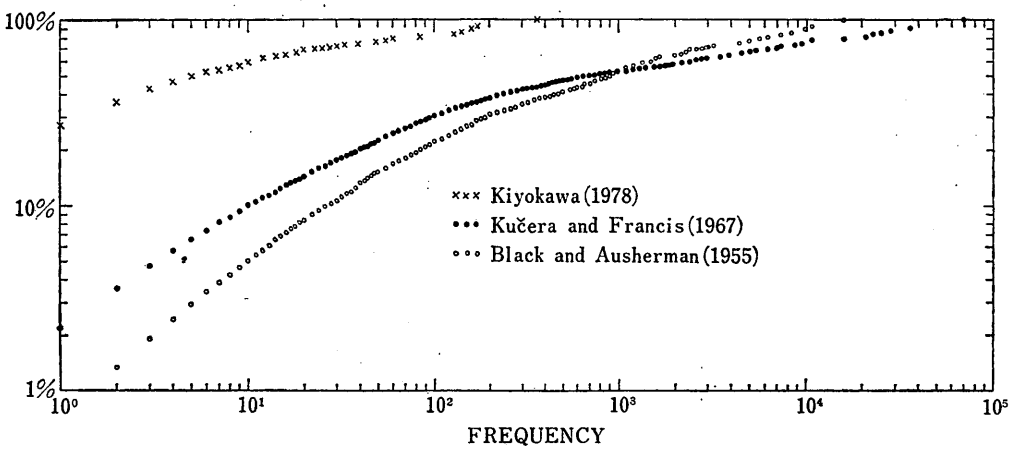

図 3 頻度と延べ語数の累積頻度

Ausherman (1955) が written English と spoken English の相異にもかかわらず，ほぼ同じ傾向を示していることがわかる。しかし，本研究のデータの曲線は，頻度が1の語がサンプルの27.5％を占めているため，他の二曲線に比べてゆるやかである。サンプル・サイズを大きくすると，低い頻度の語の割合が減少し，より急な勾配になるのではないか，と予測できる。

（2） 延べ語数の品詞別構成比

図4は延べ語数の品詞別構成比をグラフ化したものである。ひとつの単語で2つ以上の品

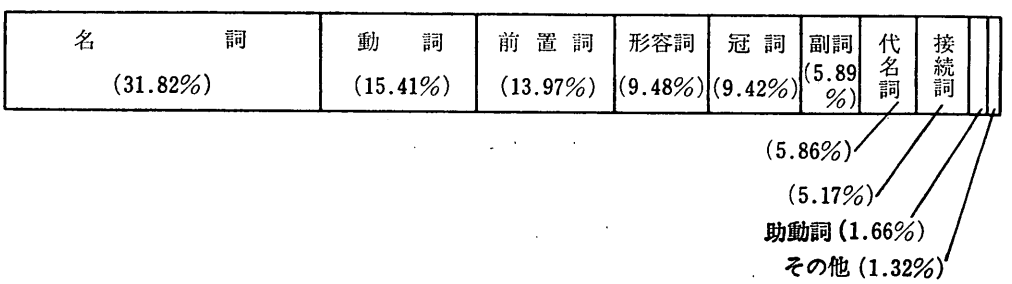

図 4 延べ語数の品詞別割合

詞にまたがっている場合は，それぞれについてカウントした。品詞の区別は「小学館ランダム

ハウス英和大辞典(1973, 小学館)」によった。現在分詞，過去分詞の形容詞利用法は，上記の辞書に形容詞としての表記があるもののみを形容詞として分類し，その他は動詞に含めた。

この図から，名詞が最も多く 31.82% を占めていることが分る。この名詞のうち，頻度が1回だけの名詞は 751 語で，1回だけ現れる語の51.72%，名詞の総数の44.02%，異語数の総計2,027語の37.05%にものぼる。

Newsweek のサンプルは written English であるが，spoken English と比べてみると，表6のような構成比となる。French et al (1930)[6] のデータは原著のデータとやや異っている

表 6 Written English と Spoken English (延べ語数) の品詞別構成比

	French et al.*	Fairbanks	Present study**
名詞	15.91%	15.39%	31.85%
代名詞	18.22	17.96	5.86
動詞	22.39	22.95	15.41
形容詞・副詞	10.06	16.85	15.37
前置詞・接続詞	12.62	18.83	19.14
冠詞	5.60	6.79	9.42
間接詞	8.08	1.26	0.

*Fairbanks (1944) より引用

**助動詞 1.66%，その他 1.32% を含まないため合計が 100%にならない

が，Fairbanks (1944)[7] より引用したものである。彼女は French et al. と異った観点から両研究の品詞を分類したためにこのようになっており，筆者も彼女の主張が正しいと判断したので，これを採用した。

この表から，電話をモニターして統計を出した French et al. および大学生に諺の解釈をさせ，それをサンプルとして分析した Fairbanks (1944) の両データにおいて，動詞が圧倒的に多く，いずれもサンプルの20%を越えている。これを本研究の名詞の比率31.82%と比べると，きわめて興味深い。また，spoken English のサンプルでは，いずれも代名詞が第2位であるのに対し，本研究では，前置詞・接続詞が第2位となっている。この理由は，異った観点から分析すれば解明できるかもしれない。

なお，Kučera and Francis (1967) および Black and Ausherman (1955) は，品詞別のデータがないのでここに取り上げなかった。

that, who, for の品詞別頻度

頻度が10位以内の語の中から，二つ以上の品詞にまたがっている語を品詞別に分類したところ，次の表7のような結果をえた。

この表から，that は見かけ上は頻度は多いが，指示代名詞はわずか4で，that の総数の7%，指示形容詞を加えても 17(28.4%) にしかすぎないことが分る。同様に，who も疑問代名詞としては6%しか使われていない。

表 7　that, who, for の品詞別頻度

that　60	接続詞　34，関係代名詞　16，指示形容詞　6，指示代名詞　4
who　33	関係代名詞　31，疑問代名詞　2
for　49	前置詞　49，接続詞　0

関係代名詞の頻度

that と who の頻度のうち，関係代名詞が予想以上の数であったので，その他の who, whose, whom, what についても調査してみたところ，次の表8のような結果であった。

表 8　関係代名詞の頻度

	左の語の頻度	関係代名詞の頻度
that	60	16
who	33	31
what	10	6
which	5	4
whose	0	0
whom	0	0

この表から，関係代名詞は that よりも who が多く，who のほとんど（94%）が関係代名詞であり，that と who 以外の関係代名詞がきわめて少ないことが分る。

（3）　英語教育への提言

最頻語と発言上の困難点

表4の中でも，the, of, in, on, for, with, that は日本人にとって音声指導上困難となる発音を含んでいる。that, with, the は〔ð〕, of は〔v〕, for は〔f〕, in と on は〔n〕の発音練習につごうのよい単語である。発音の導入にはふつう名詞と絵を同時に示して（例，mother と〔ð〕など）行なわれるが，このような方法は key word として記憶させるためには必要であろうが，〔ð〕の練習のための教材ならば，その後の練習では，with, the を含む文を提示するような工夫が望ましいといえよう。

品詞と教材

自作，市販のいずれを問わず，教材を作成するにあたっては，品詞に関係なく，どの単語が大切かという，重要度が単語の選択の第一の要因となる。次に，「どの品詞を優先をさせるか」についてもかなり注目する必要があろう。教授目標によっても異るが，読解力養成の教材ならば，名詞の構成比が全体の 31.82% であることを考えると，名詞に力点をおくことが必要である，といえる。ただし，頻度が1回の語が名詞の総数の 44.02% を占めるという事実から，いわゆる recognition のための語いの拡張が重要となってくる。

最頻語と品詞，関係代名詞などの指導

表7の最頻語の頻度から，指示代名詞，指示形容詞としての that は比較的少なく28.4%であった。このことから，thatは，関係代名詞と接続詞に力点を置くことが必要である，といえる。同様に，who についても疑問代名詞よりも，関係代名詞としてとり上げることが望ましい，といえよう。

表8の結果から，関係代名詞の頻度は，who, that, what, which の順に大である。さらに，whose, whom が 0 であるという事実を考えると，サンプルがきわめて大きくなっても，この二語が出現する確率はかなり低いと考えられる。このことから，文法，作文の時間にこれらの語を従来ほどきびしくくり返す必要があるか，きわめて疑問である，といえる。

IV. 今後の課題

今後の課題はサンプルを多くすることであるが，他の時事英語の他に，フィクション，学術書，実用書，宗教関係，話しことばなどジャンルも広くする必要がある。フィクションも科学的な読物からラブストーリーまで種々にわたっているのでどのジャンルからいくつのサンプルをとるか，を決定するための資料を作ることが，次の課題となってくる。

サンプルの発行された年も1978年に限るか，できるだけ近い年に限ることが望ましいであろう。単行本はそれ以前に準備されたものであるが，「現代英語」というカテゴリーには属すると考えられるからである。

引用文献

(1) Thorndike, E.L. and Lorge. I. *The Teacher's Word Book of 30,000 Words*, Teachers College Press, 1944

(2) Kučera, H. and Francis, W. N. *Computational Analysis of Present-day American English*, Brown University Press, 1967

(3) Carroll, J. B., Davies, P., and Richman, B. *American Heritage Word Frequency Book*, American Heritage Publishing Co., Inc., 1971

(4) Black, J. W. and Ausherman, M. *The Vocabulary of College Students in Classroom Speeches*, Ohio State University, 1955

(5) Miller, G. *Language and Communicatin*, McGraw-Hill Book, Co., Inc., 1951, p.91

(6) French, N. R., Carter, C. W., and Koening, W. Jr., "The Words and Sounds of Telephone Conversation," *Bell System Technical Journal*, IX (April, 1930).

(7) Fairbanks, H., "The Quantitative Differentiation of Spoken Language," *Studies in Language Behavior*, (Psychological Monographs, Whole Number 255), 1944