The Tagged LOB Corpus について

乾　　隆

1．はじめに

近年のコンピューター技術の発達により，かつては膨大な時間と労力を要した言語の統計的分析や corpus と呼ばれるデーターベースの作成が容易となり，それらに基づいた研究が多く見られるようになった。LOB Corpus もそのようなコンピューターの力を駆使して開発された現代イギリス英語のデータベースの一つである。

1．1　LOB Corpus とBrown Corpus

LOB Corpus は，現代イギリス英語を，のべ語数にして約百万語，収集分析した一大言語資料である。一方，現代アメリカ英語の corpus としては，Brown Corpus が夙に有名である。LOB Corpus は計画の基本方針や分析方法において，既に高い評価を獲得していたこの Brown Corpus を踏襲しており，そのイギリス英語版と考えれば実像が浮かび易い。これら２つの資料が揃ったことにより，現代のアメリカ英語とイギリス英語の比較研究が促進されると思われる。

1．2　LOB Corpus の出現

1．2．1　研究開発者

LOB Corpus の研究開発は，1970年に University of Lancaster の Geoffrey Leecch の指揮のもとに開始され，1977年以後の作業を University of Oslo の Stig Johansson の指揮下に移し，最終的に Norwegian Computing Centre for the Humanities の Knut Hofland の協力を得て一応の完成

を見た。指導的役割を果たしたこれら三者の所属する研究施設の所在地名より，The Lancaster-Oslo/Bergen (LOB) Corpus と呼ばれる。

1．2．2 Hofland and Johansson (1982)

一般の研究者にも利用し易い資料として，LOB Corpus から得られるデータが，Hofland と Johansson の共著で *Word Frequencies in British and American English* の書名で1982年にロングマン社より出版された。この内容は，データの考察と「LOB Corpus の頻度順位表 (rank list)」，「アルファベット順頻度表」，「頻度順位上位百語の出典ジャンル別頻度表」，「アルファベット順出典分野別頻度表」，及び最も特徴的である「Brown Corpus との対比表」の4つの資料から成り立っている。更にこれには，逆引き語彙表 (Reverse-alphabetical word list) などが収録されている3枚のマイクロフィッシュ (microfiche)[1] が添付されている。この逆引き語彙表を用いれば，接尾辞を基準にした単語の頻度分析などが容易に行える。

1．2．3 Hofland and Johansson (1986)

Hofland and Johansson (1982) が基づいている LOB Corpus はコンピューターの処理上，同綴り語はすべて1種類の語として処理された。したがって，名詞 mine (鉱山) と所有代名詞の mine のように，異なる品詞で全く別の語でも，偶然に同綴りであれば，同一語として扱われてしまった。このような同綴り語の数はそう多くなく LOB Corpus 全体の資料的価値を下げるものではないが，同じ百万語を分析しても品詞の観点からも語を分類しておけば，単語レベルの資料から更に，文中におけるその役割，つまり統語レベルの資料としても使えることになる。

このような要求に合わせて，各語に tag を付与する方法で開発されたのが，tagged LOB Corpus である。tagged LOB Corpus は一般に利用で

きる資料として，コンピューターの磁気テープとマイクロフィッシュの形態で1986年に *The Tagged LOB Corpus: KWIC Concordance* の名で世に出された。しかしこれは，その形態及び，ノルウェーのコンピューターセンターで出されたなどの理由で，簡単に利用できるとは言い難かった。

1．2．4　Johansson and Hofland (1989)

そして1989年，tagged LOB Corpus は Stig Johansson と Knut Hofland の共著で，*Frequency Analysis of English Vocabulary and Grammar,* vol. 1, vol. 2 として，書物の形態で世に出された。以下では，主にこの Johansson and Hofland（1989）の内容に従い，tagged LOB Corpus を解説し言語的資料としてその有効性を紹介する。

2．tagged LOB Corpus の枠組み

2．1　tag とは

入力する多種類のデータをコンピューターが区別できるように各データに付けるプログラム上の記号を tag と言う。入力用データとして抽出したサンプルの英文をそのままコンピューターに入れると，通常のプログラムでは，コンピューターは各単語の形態上の区別しかできないので，同綴り語の区別をしたり，当該の単語が如何なる文法的範疇に属するかは分析できない。そこで，各単語の属する文法的範疇をコンピューターが判読できる記号にして各単語に付ける必要が生ずる。その記号が tagged LOB Corpus での154種の tag である。実際に He is a boy. の英文に tag を付けるとすれば次のようになる：

ˆ he PP3A is BEZ a AT boy NN.

各単語及びカンマの横に添えられた ˆ, PP3A, BEZ, AT, NN, . の記号が tag であり，それぞれ，〈文頭のスペース〉，〈人称代名詞 3 人称単数形〉，〈is または 's〉，〈単数冠詞（a, an, every）〉，〈単数普通名詞〉，〈カンマ〉

を表している。

tag を付けて英文を入力することにより，単語の頻度だけでなく，tag 毎の頻度，文の位置による tag 頻度，語の組合せ頻度など多くの統計的データが得られる。

入力用データの英文のすべての単語に一語ずつこの tag を付けてから，コンピューターに統計的処理をさせる訳であるが，この tag 付与の作業には自動 tag 付与プログラム（automatic tagging program）とその前後の手作業による編集が併用された。それまでの tag の付いていない LOB Corpus 全体に tag を付ける作業は大変な時間と労力を要したものと容易に想像される。因に Brown Corpus の場合は，同様の tag 付与の作業に10年の歳月を要している。

2．2　サンプルソース

いかに大量のサンプルを分析しても，そのソースのジャンルに偏りがあれば，分析結果は，英語全体の姿を正しく反映したものとはならず，特定の分野における語彙実態を表すだけの結果になってしまう。

LOB Corpus は1961年発行の書籍，新聞・雑誌，政府刊行文書から，表１に示すＡ～Ｒの15分野500種類のテクストを分析している。[2] ただし，各テクストに記載されているすべての語を分析したのではなく，各テクスト毎に2000語ずつの単語を分析している。その2000語は英文のままで入力するのであるが，当該テクストの第１ページから入力を開始するのではなく，無作為にある記事を選定し，そのページから順次2000語に達するまで入力を続けるのである。１篇のまとまりある文章（これを１item とする。新聞であれば１件の記事が１item）が比較的短い場合は，2000語に達しないうちに１item が終ることがあるが，そのような場合は，機械的に次の item へ入力を続けていくのではなく，内容及び文体が似ている item が現れるまで入力を中止することにしている。したがっ

表1．LOB Corpus のサンプルソースの分野別部類

		テクスト数			テクスト数
A	報道：報道記事	44	K	小説一般，短編小説一般	29
B	報道：社説	27	L	推理小説，短編推理小説	24
C	報道：評論	17	M	空想科学小説	6
D	宗教	17	N	冒険小説，西部劇小説	29
E	技術，商業，趣味	38	P	ロマンス，恋愛小説	29
F	知識	44	R	ユーモア	9
G	純文学，文学論文，伝記	77		計	500
H	政府文書，財団・企業報告書，業界通報	30			
J	科学，人文科学，社会科学	80			

て，4～5の item にまたがって2000語が抽出されている場合もある。

表1のA～Rの分野は2分され，A～Jが情報供給的（informative），K～Rが想像的（imaginative）なテクストより構成されている。また，A～Rの各分野は更に細かく下位分類されており，例えばAの報道記事の44のテクストは，表2のような内訳になっている。

表2．分野A（報道）のテクストの種類別内訳

A01-06	日刊全国紙	政治	A27-31	日刊地方紙	政治
A07-08	〃	スポーツ	A32-33	〃	スポーツ
A09-10	〃	社会	A34-37	〃	スポットニュース
A11-14	〃	スポットニュース	A38	〃	財政
A15-16	〃	財政	A39-40	〃	文化
A17-19	〃	文化	A41	週刊地方紙	スポーツ
A20-21	日曜刊全国紙	政治	A42	〃	社会
A22-23	〃	スポーツ	A43	〃	スポットニュース
A24	〃	スポットニュース	A44	〃	文化
A25	〃	財政			
A26	〃	文化			

Brown Corpus の場合は，上記の EFG の分野のテクスト数がそれぞれ，

38, 44, 77ではないが，この点を除けば，テクスト発行年数も含めてLOB Corpus と Brown Corpus のサンプリングの方法は同じである。

2．3　分野別入力データ語数

各テクストから2000語ずつ抽出したと述べたが，これは tagged LOB Corpus の基になった本来の LOB Corpus でのことである。LOB Corpus では，例えば，分野Aの分析語数は単純計算で次のようになる。

分野Aの分析語数＝2,000×テクスト数＝2,000×44＝88,000

しかし，tagged LOB Corpus の分野Aの分析語数は89,138でこの数値よりも若干多い。A以外の分野でも同じで，表3で示す各分野の分析語数は単純計算で得られる数値よりも若干多く，結果的に全体の語数も百万語を超えている。これは，tagged LOB Corpus では元の LOB Corpus の2000語を使っているが，例えば，I'll や She's など1語扱いをしていた縮約形を，tagging の関係上，分割して2語として登録するなど，語の境界の定義に僅かながら変更点があったことによる。

表3．tagged LOB Corpus の分野別 running words 数

A	89,138	D	34,387	G	155,336	K	59,204	N	59,391
B	54,447	E	76,913	H	60,761	L	49,145	P	59,382
C	34,321	F	89,090	J	161,900	M	12,119	R	18,203
							総計		1,013,737

2．4　tag の構成

tagged LOB Corpus の特徴は，その名の示す通り tag にある。tag は本来コンピューター処理のための記号なので，そのままの表示では，通常の使用に適していない。ところが，Johansson and Hofland（1989）で紹介されている膨大なデータは tag の表示がそのまま用いられているので，初めての使用者には即座に利用できるとは言い難く，特徴であるは

ずの tag が，かえって tagged LOB Corpus を一般の利用者に使い難いものにしている。そのような tag が容易に判読できるように，その構成を簡単に解説しておく。

2．4．1　基底 tag と「接尾辞」

tagged LOB Corpus の154種の tag の表示は恣意的ではなく，表4に示す基底 tag とその下位範疇で用いる「接尾辞」との組合せで構成されている。基底 tag と「接尾辞」を念頭に置けば，tagged LOB Corpus は相当利用し易くなる。しかし，これだけでは十分ではなく，例えば，2．1で，He is a boy. の tag を ^, PP3A, BEZ, AT, NN,. と紹介したが，これら6個の tag のうちで，表4の基底 tag と表5の「接尾辞」の組合せで容易に判読できるのは，BEZ の tag だけである。つまり，BEZ は〈be 動詞〉を表す基底 tag の BE と〈3人称単数〉を表す「接尾辞」z から構成されているので「3人称単数で用いる be 動詞」ということで is とわかるのである。

tag の構成は，これら2つの表で表される基本組織の他に，品詞によって，更に細かな取り決めがある。因に，代名詞としての機能しか持たない語は，表4に示すように，その tag はどれも P で始まるが，それらの下位区分は次のように決められている：

PN	代名詞（以下の代名詞は除く）	例：anybody, none
PP＋数字	人称代名詞（数字で何人称か示す）	例：I, you
PPSS	所有代名詞	例：mine, yours
PPL	再帰代名詞	例：myself, yourself

これらの tag の後には，数や格を示す「接尾辞」が付くことがある。たとえば，PP3AS であれば，PP3（3人称の人称代名詞）＋A（主格）＋S（複数）であり，they を表す。PP3A はこれから接尾辞S（複数）を除いたものなので，he か she ということになる。

表4．基底 tag

A....	決定詞，代名詞	HV...	have（動詞，助動詞）	R....	動詞
BE...	be動詞（含助動詞）	IN	前置詞	TO	不定詞の to
CC	等位接続詞	J....	形容詞	UH	間投詞
CD...	基数	MD	法の助動詞	VB...	動詞
CS...	従属接続詞	N....	名詞	W....	疑問詞
DO...	do（動詞，助動詞）	OD...	序数	ZNOT	not
DT...	決定詞，代名詞	P....	代名詞	ZZ	文字
EX	存在の there	QL...	限定詞		

表5．tag に用いる「接尾辞」

A	主格	$	所有格	N	過去分詞
O	対格	R	関係詞	Z	3人称単数
I	単数 or 複数	D	過去時制	R	比較級
S	複数	G	現在分詞，動名詞	T	最上級

2．4．2 ditto tag

in order to，and so on などのような熟語的表現は，一つのまとまりとしての文法機能を持ち，構成要素の各単語の文法的機能とは別である。このような熟語的表現に tag を付与する場合は，構成要素の各単語にその本来の文法機能を示す tag を付与するのではなく，次に示すように，句としての文法機能を示す tag を付与する方針がとられている。

in order to → TO TO" TO"

and so on → RB RB" RB"

この例では，in order to を複合不定詞標識（complex infinitve marker）として扱い，1語目の in に不定詞を示す tag である TO を付与し，2語目以後の語には TO" を付与している。同様に，and so on は複合副詞（complex adverb）として扱い，1語目の in に副詞を示す RB を付与し，2語目以後の語には RB" を付与している。ここで示した TO" や RB" のように「"」のついた tag を ditto tag と呼んでいる。

ditto tag の導入によって，語の分析だけにとどまらず，熟語の文法機

能を踏まえた分析も可能になっている。

3．tagged LOB Corpus から得られるデータ

tagged LOB Corpus はコンピューターに入力されているデータベースなので，コンピューターのプログラミング次第で多種多様なデータが得られるが，Johansson and Hofland（1989）では，主に次の１～４のデータを表の形式で提示している。

Vol. 1：　１．tag 頻度　　２．語頻度

Vol. 2：　３．tag 組合せ頻度　４．語組合せ頻度

以下ではこの１～４のデータの解説を行いながら，考察を加える。なお説明に用いた表６～表12は Johansson and Hofland（1989）からの抜粋である。

3．1　tag 頻度

表６は，tag 頻度表の一部であるが，左欄には tag 名（ABL～ZZ をアルファベット順に，その後に！～？の記号 tag），１行目には入力した英文の出典の分野（A～R）を配している。各欄の数字は左欄で示されている tag の分野別頻度を示し右端の欄でその合計を出している。各欄の数値は２段になっているが，上段では絶対頻度，下段では相対頻度を示している。

絶対頻度はその tag で表される語が実際に用いられた度数のことであり，相対頻度は絶対頻度を百万語当りの頻度に換算した数値である。表６の PP3A（he，she）の欄で見ると，A（報道記事）の欄では，上段が1053，下段が11813となっているので，he，she の三人称単数の人称代名詞は，報道記事で実際に1053回使われ，百万語当りでは11813回使われることになることがわかる。

表6．tag 頻度

Tag	A	B	C	D	E	F	G	H	J	K	L	M	N	P	R	A-J	K-R	Total
PP3A	1053	336	357	226	199	819	1617	125	563	1761	1527	185	1902	2251	240	5295	7866	13161
	11813	6171	10402	6572	2587	9193	10410	2057	3477	29745	31071	15265	32025	37907	13185	7001	30554	12983
PP3AS	316	249	108	156	296	456	508	210	302	328	124	78	227	248	79	2601	1084	3685
	3545	4573	3147	4537	3849	5118	3270	3456	1865	5540	2523	6436	3822	4176	4340	3439	4211	3635
PP3O	156	62	89	55	37	251	462	34	91	606	441	52	580	816	58	1237	2553	3790
	1750	1139	2593	1599	481	2817	2974	560	562	10236	8973	4291	9766	13742	3186	1636	9917	3739
PP3OS	114	99	62	77	122	182	274	83	149	183	74	32	128	104	32	1162	553	1715
	1279	1818	1806	2239	1586	2043	1764	1366	920	3091	1506	2640	2155	1751	1758	1536	2148	1692
PPL	49	41	45	54	44	91	236	29	110	124	104	29	130	149	22	699	558	1257
	550	753	1311	1570	572	1021	1519	477	679	2094	2116	2393	2189	2509	1209	924	2167	1240
PPLS	19	36	17	27	21	59	95	26	58	29	17	11	14	22	13	358	106	464
	213	661	495	785	273	662	612	428	358	490	346	908	236	370	714	473	412	458
PPLS′	9	6	8	2	2	20	16	0	23	19	6	3	9	14	3	86	54	140
	101	110	233	58	26	224	103	0	142	321	122	248	152	236	165	114	210	138
QL	318	282	278	176	453	483	954	248	760	366	234	74	268	381	100	3952	1423	5375
	3568	5179	8100	5118	5890	5421	6142	4082	4694	6182	4761	6106	4512	6416	5494	5225	5527	5302
QLP	16	10	9	2	25	31	44	7	22	26	40	6	21	20	4	166	117	283
	179	184	262	58	325	348	283	115	136	439	814	495	354	337	220	219	454	279
RB	2213	1861	1270	1108	2791	3168	5684	1642	5486	2358	1967	458	2239	2435	677	25223	10134	35357
	24827	34180	37004	32221	36288	35560	36592	27024	33885	39828	40024	37792	37699	41006	37192	33351	39364	34878
RB″	79	109	58	55	112	139	333	72	247	118	119	22	129	157	36	1204	581	1785
	886	2002	1690	1599	1456	1560	2144	1185	1526	1993	2421	1815	2172	2644	1978	1592	2257	1761
RBS	1	0	0	0	0	0	0	0	0	1	1	0	1	2	0	1	5	6
	11	0	0	0	0	0	0	0	0	17	20	0	17	34	0	1	19	6
RBR	106	54	42	31	76	141	271	50	159	87	81	18	114	96	31	930	427	1357
	1189	992	1224	902	988	1583	1745	823	982	1469	1648	1485	1919	1617	1703	1230	1659	1339
RBT	3	13	6	1	9	9	17	8	15	4	2	1	10	4	1	81	22	103
	34	239	175	29	117	101	109	132	93	68	41	83	168	67	55	107	85	102

表6の右から3番目の欄にはA～J（情報供給的テクスト）における頻度，同じく2番目の欄ではK～R（想像的テクスト）における頻度が示されている。この2つの欄を比較すると，PP3A では相対頻度（下段）はそれぞれ，7001と30554なので，K～Rの分野で，はるかに多用されることがわかる。またこの比は7001：30554＝1：xより，1：4.4と求められるので，「he，she は想像的テクストで情報供給的テクストでよりも4倍以上用いられる」と解釈できる。

表6の右端の欄では全分野における頻度が示されている。PP3A の相対頻度を見れば，12983の数値が得られるので，これも1000000：12983＝x：1よりx＝77が求められ，「he，she は英文中で77語に1語の頻度で用いられる」と解釈できる。

3．2　語頻度

語頻度（word frequency）表ではアルファベット順に各語の頻度を示している。Johansson and Hofland（1989）ではこの表が最大の資料であ

る。表7はその中から for, since, because, student, oxygen の項を抜粋した。それぞれの語の，3つの数値欄は左から，頻度，出現分野数（分野A～Rのうち，その語またはtagが出現した分野の数，最多は15），出現テクスト数（最多は500）を示している。

表7．語頻度表

for	9306	15	500	**since**	541	15	291	**because**	777	15	329
CS	488	15	238	CS	305	15	194	CS	635	15	293
IN	8694	15	500	IN	182	14	134	IN	142	15	110
IN″	81	15	70	RI	54	14	47				
NC	1	1	1					**student** NN	64	14	39
RB	42	11	32					**oxygen** NN	65	3	8

各語の頻度は左欄を見るだけでわかるが，中欄と右欄を見れば，その語の使用範囲の広さがわかる。for, since, because は機能語なので頻度も高く使用範囲も広い。これに比べて，名詞の oxgen と student は頻度が低い。しかしこの2語は，頻度が65，64とほぼ同じにもかかわらず，student は比較的広く使われ，oxygen の使用範囲には極めて偏りがあることがわかる。

更に，左欄（頻度）を右欄（出現テクスト数）で割れば，その語の出現テキスト中での相対頻度も算出でき，それぞれ above（1.6），because（2.4），oxygen（8.1），since（1.9），student（1.6）の数値が得られる。

oxygen の数値の突出は，この語が特定のテクストで極めて高頻度に用いられることを意味し，専門用語に属することを示している。

各語の tag の頻度も示唆に富んでいる。for の tag は，CS（従属接続詞），IN（前置詞），IN″（熟語的前置詞句中の1語），NC（引用），RB（副詞）であるが，従属接続詞としての頻度が488，前置詞としての頻度が8694なので，〈for＋句〉は〈for＋節〉の約18倍の高頻度であることがわかる。これとは逆に，because は CS（従属接続詞）が635，IN（前置詞）が142なので，〈because＋節〉が〈because of＋句〉よりも約4倍多く使わ

れることがわかる。since も，because と同様に，CS（従属接続詞）の頻度がIN（前置詞）の頻度より高い。

これらの結果を応用言語学の観点から考えてみると，学習者に新語を導入する場合，単に頻度の高い語を優先的に導入すべきと考えるのは危険であり，使用範囲の分散度も考慮に入れなければならないことがわかる。また辞書などの編集では，多品詞語の扱う品詞の順の決定に，tag の頻度を参考に入れるべきであることが理解できる。

3．3　tag 組合せ頻度

英文を構成する各単語の意味がわかっても，それらの統語関係がわからないと文意がわからないことが多い。英語の統語関係はほとんどの場合語順で決まる。したがって，語と語の組合せの傾向を知ることは英語の統語関係を知る上で重要である。

Johansson and Hofland（1989）の tag 組合せ頻度表は，語と語の組合せの傾向を，tag の組合せの頻度として，ditto tag を除く全 tag をアルファベット順に明確な数値で示している。表8はそのうちの JJT（形容詞の最上級）と RB（副詞）の項を抜粋したものである。

先ず，JJT（形容詞の最上級）の項目でこの表を説明することにする。太字で書かれた JJT の tag 名を中心に欄が左右に分かれているが，左側は英文中で JJT の直前に位置する tag，右側は同じく JJT の直後に位置する tag を JJT との組合せの頻度順に上から下へ並べてある。3つの数値欄は，ATI（the，no）を例にとると，左欄の730は〈ATI＋JJT〉の組合せの度数，中欄の70.33はすべての〈～＋JJT〉の組合せに占める〈ATI＋JJT〉の百分率，右欄の1.04はすべての〈ATI＋～〉の組合せに占める〈ATI＋JJT〉の百分率を表す。

JJT（形容詞の最上級）の直前には定冠詞が最も多く位置することは，経験的に知る所であるが，tag 組合せ頻度により，JJT の直前には（no

も含めてだが）約70％の高頻度で定冠詞が位置することがわかり，経験的知識の正しさを裏付けている。逆に，ATI（the, no）から見ると，JJT との組合せは１％程度に過ぎず，JJT はそれ程重要な組合せ相手でないことがわかる。

組合せの度数が10以下の tag はすべて≦10の行にまとめられている。一番下の行は各欄の数値の合計である。

副詞の位置は比較的自由であるが，この tag 組合せ頻度表を利用すると，副詞の位置にはある程度の傾向があることがわかる。表８の RB（副

表８．JJT, RB との tag 組合せ頻度表[3)]

JJT

ATI	730	70.33	1.04	NN	490	47.21	0.33
PP$	128	12.33	0.75	NNS	175	16.86	0.34
NP$	36	2.50	1.05	IN	105	10.12	0.09
BEZ	15	1.45	0.12	JJ	67	63.45	0.10
CC	14	1.35	0.04	,	40	3.85	0.07
NN$	13	1.25	0.83	CC	23	2.22	0.06
`	12	1.16	0.02	TO	22	2.12	0.14
≦10	10100	9.63		.	22	2.12	0.04
	1038	100.00		JNP	11	1.06	0.35
				≦10	10 83	7.99	
					1038	100.00	

RB

,	3372	9.54	6.18	IN	4418	12.50	3.58
`	2578	7.29	5.00	,	3816	10.79	7.00
CC	2114	5.98	5.73	JJ	3549	10.04	5.56
NN	2035	5.76	1.37	VBN	3222	9.11	11.92
BEZ	1709	4.83	14.05	.	2218	6.27	4.41
QL	1590	4.50	29.58	VB	1621	4.58	4.96
VBD	1484	4.20	6.01	VBD	1346	3.81	5.45
MD	1428	4.04	9.60	RB	1180	3.34	3.34
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
≦10	10126	0.34		≦10	10119	0.36	
	35357	100.00			35357	100.00	

詞）の項で，RB の直前に位置する tag は，（カンマ），^（文頭のスペース），CC（等位接続詞）が最も多く，それぞれ9.54%，7.29%，5.98%である。一方，RB の直後に位置する tag の中で，（カンマ）．（ピリオド）は最多ではないが高い頻度であり，それぞれ10.79%，6.27%である。したがって，副詞は文頭・文末，あるいは節頭・節末のように文の前後の端に位置する傾向が見られる。もちろん BEZ（is，'s），VBD（動詞の過去形），VBN（過去分詞）など動詞類の tag も多く見られるので，副詞が動詞の前後に位置する傾向も強いことは言うまでもない。

表８には掲げなかったが，^（文頭のスペース）の項を見ると，副詞は引用符，The または No，He または She，単数固有名詞に次いで，文頭に位置する頻度が高く5.00%であることがわかる。同様に，．（ピリオド）の項を見ると，その直前の副詞の頻度は4.41%である。この２つの事実から，およそ５%の英文は副詞で始まるか，副詞で終ると言えよう。

３．４　語組合せ頻度

Johansson and Hofland（1989）の語組合せ（word combination）頻度表には４種類あり，それぞれ，動詞，形容詞，名詞，副詞小辞の次に位置する tag 及び語の頻度表である。ここでの「語組合せ」は「コロケーション」と呼ぶ方が理解され易いかもしれないが，いずれにせよ，この４種類の表は見出し語のアルファベット順に編集されているので，「頻度付きコロケーション辞典」あるいは「頻度付き単語活用辞典」としても使用可能である。表９～表12に従って，各頻度表を説明するが，表中で大文字による表記の語はその語のすべての語形を代表している。

３．４．１　動詞との組合せ

動詞との組合せ頻度表では，tagged LOB Corpus の中で少なくとも１種の語形（形態）が11回以上現われた動詞を見出し語としている。

表9の見出し語 ADVISE は大文字なので，ここで動詞 advise の全語形を扱っていることがわかる。なお，動詞 ADVISE の項に限らず，動詞の表の見出し語はすべて大文字である。1行目の TOTAL の欄の数値は，ADVISE の VB（原形），VBZ（3人称単数現在形），VBD（過去形），VBN（過去分詞形），VBG（ing 形）の各語形の度数及びそれらの合計度数を示している。これらの数値は3．2で説明した語頻度表でも得られるが，この表は一括して表示しているので全体像が把握し易い。

左欄の P*（代名詞），TO（to 不定詞）……CC（等位接続詞），CS（従属接続詞）の tag は tagged LOB Corpus で動詞 ADVISE の直後の位置を2件以上のテクストで占めた tag である。つまり，動詞 ADVISE と組合せになり得た語の tag であり，その組合せの度数の多い順に上から下に並んでいる。一番多い P* を例にすると，P* の行の TOT の欄を見ると11なので，〈ADVISE＋代名詞〉の組合せは11回出現したことになる。ADVISE の各語形と P* との組合せの度数は，P* の VB（原形）…VBG（ing 形）の各欄に示されており，その6，0，3，2，0の数値は〈advise＋代名詞〉の組合せが6回，以下〈advises＋代名詞〉0回，〈過去形 advised＋代名詞〉3回，〈過去分詞 advised＋代名詞〉2回，〈advising＋代名詞〉0回を表している。

組合せ頻度表の下部は実際の語との組合せの度数を示したもので，2件以上のテクストで見出し語の直後の位置を占めた語をアルファベット順に載せている。表9によると動詞 ADVISE には〈ADVISE＋人称代名詞＋to do〉，〈ADVISE＋that 節〉，〈ADVISE on～〉，〈advised to do〉などの表現があることがわかる。

紙幅の関係でここに取り上げなかったが，BECOME の項を見ると，この語には形容詞が最も多く後続し，計算によるとその確立は約30％である。もちろん，〈BECOME＋to 不定詞〉の組合せは見当らない。BEGIN の項では to 不定詞が後続することが最も多く，52.5％の高頻度で

ある。これに比べ，〈BEGIN＋～ing〉の組合せは意外に少なく4.9％である。このような情報は辞書や教科書の編集に極めて有用であろう。

3．4．2　形容詞との組合せ

形容詞との組合せ頻度表は，tagged LOB Corpus の中で10回以上現れた形容詞を見出し語としている。この表の構成は基本的に動詞の場合と同じであるが，動詞のような語形変化がない分簡単になっている。

表10では busy の項を示したが，この見出し語は表9の動詞の場合のように大文字ではないので，ここで扱う busy は代表形ではなく，busier や busiest のような他の語形を含んでいない。表10より，busy の後には N* （名詞類）が15回，IN （前置詞）が13回，PUNCT （句読点）が9回，VBG （～ing）が6回，TO （to 不定詞）が2回出現したことがわかる。

ここで留意すべきは，N* （名詞類）の tag が一番多いことから，形容詞 busy は〈busy＋名詞類〉，つまり付加用法 （attributive usage） での使用が一番多いと即断するような資料解釈上の過ちを犯さないことである。N* 以外の tag はすべて叙述用法 （predicative usage） で用いられると考えられるので，それらの tag の各々の度数は低いが合計すれば30回となり，N* の15回よりも多く，形容詞 busy は叙述用法で用いられる方が多いという結果が得られる。このことは，度数を次のように百分率に換算すると一層明白であり，N* が過半数を超えないことがよくわかる。

N* (28.8％)，　IN (25.0％)，　PUNCT (17.3％)，
VBG (11.5％)，　TO (3.8％)

表10の後半では，2件以上のテクストで形容詞 busy の直後の位置を占めた語をアルファベット順に掲載している。busy と前置詞の組合せは，頻度以外の情報であれば，通常の辞書でもある程度得られるが，名

詞との組合せの busy years は通常の辞書では例文としてほとんど記載されておらず，貴重な情報と言える。

表9．動詞 ADVISE との組合せ

ADVISE

	VB	VBZ	VBD	VBN	VBG	TOT
TOTAL	25	3	9	20	2	59
P*	6	0	3	2	0	11
TO	3	0	1	4	0	8
DET	5	1	0	0	1	7
IN	3	1	0	3	0	7
PUNCT	3	0	2	1	0	6
N*	3	0	1	1	0	5
CC	4	0	0	0	0	4
CS	1	0	2	1	0	4

advised	*VBN*	by *IN*	2
ADVISE		her *PP30* to *TO*	2
ADVISE		on *IN*	4
ADVISE		that *CS*	4
ADVISE		them *PP30S* to *TO*	3
advised	*VBN*	to *TO*	11
advised	*VBN*	to *TOG take VB*	2

表10.

形容詞 busy との組合せ

busy *JJ*

TOTAL	52
N*	15
IN	13
PUNCT	9
VBG	6
TO	2
-for *IN*	2
-in *ING*	3
-on *IN*	3
-to *TO*	2
-with *IN*	3
-years *NNS*	2

5．4．3　名詞との組合せ

名詞との組合せは頻度表は，Johansson and Hofland（1989）の動詞及び形容詞の組合せ頻度表の中に出現した名詞類（Nで始まる tag）との組合せのすべてを扱っている。表11からわかるように，前半は〈形容詞＋名詞(＋前置詞)〉，後半は〈動詞＋(冠詞＋)名詞(＋前置詞)〉の組合せと度数をそれぞれ形容詞・動詞のアルファベット順に示している。

この頻度表により，当該の名詞を修飾する形容詞や，それを目的語とする動詞の種類といった統語上の傾向が明示的にわかる。

5．4．4　副詞的小辞との組合せ

ここで言う副詞的小辞とは tagged LOB Corpus で RP の tag を持つ語である。RP には形態上同綴りの前置詞が存在し得る。しかし RP は常に動詞と対になって使われ，単なる〈動詞＋副詞〉とは意味的に別の新たな動詞構造を形成する。その構造を句動詞（phrasal verb）と呼ぶことが多い。したがって，「副詞的小辞との組合せ」は「句動詞」と同値と考えてよい。

副詞的小辞との組合せ頻度表も，Johansson and Hofland（1989）の動詞及び形容詞の組合せ頻度表の中に出現した副詞的小辞のすべてを扱っている。数が多くないのでそのすべてを示す：

about, across, along, around, aside, away, back, behind, by, down, forth, in inside, off, on, out, outside, over, past, round, through, up.

表11．名詞 look との組合せ

look *NN*	85
good *JJ* look *NN*	5
good *JJ* look *NN* at *IN*	3
new *JJ* look *NN*	2
quick *JJ* look *NN*	2
come *VB* and *CC* take *VB* a *AT* look *NN*	2
GET a *AT* good *JJ* look *NN*	2
GET a *AT* good *JJ* look *NN* at *IN*	2
GIVE him *PP30* a *AT* look *NN*	2
LIKE the *ATI* look *NN*	2
LIKE the *ATI* look *NN* of *IN*	2
SEE the *ATI* look *NN*	3
SEE the *ATI* look *NN* of *IN*	3
TAKE a *AT* look *NN*	9
TAKE a *AT* look *NN* at *IN*	6

表12．副詞的小辞 through との組合せ

through *RP*	90
BREAK through *RP*	5
BREAK through *RP* at *IN*	2
BREAK through *RP* to *IN*	2
CARRY through *RP*	4
COME through *RP*	9
FALL through *RP*	3
FOLLOW through *RP*	2
GET it *PP3A* through *RP*	2
GET through *RP*	8
GET through *RP* to *IN*	2
GO through *RP*	10
went *VBD* through *RP* to *IN*	2
go *VB* through *RP* with *IN*	3
PASS through *RP*	2

この頻度表はJohansson and Hofland（1989）の他の膨大な頻度表と比べると，僅か7ページの短い資料に過ぎない。それは副詞的小辞そのものの数が少ないことと，隣接する動詞と副詞的小辞の組合せしか収録していないことによる。表11では例外的に動詞と句動詞が離れているGET it *PP3A* through *RP* のパターンが収録されているが，句動詞は実際の使用においては，このように動詞と副動詞的小辞が離れている場合が多いのである。

6．問題点とまとめ

tagged LOB Corpus の言語資料としての最大の問題点は，サンプル数の少なさである。もちろん，データから得られる結果の傾向は，サンプル数を何倍かに増やしても大きな点では変わらず，現状でも資料的価値を損なうものではない。しかし，低頻度の部分での現象がデータの少なさ故に判断が付きかねる事が多い。本稿で提示した資料でも一桁以下の数値が目立ち，それらの語彙では実際の差が数値の差として十分反映されていないことに気付くであろう。

テクスト500種，百万語のサンプルは一見多いように思われる。しかし，1テクストから2000語は少ない。因に，*Encyclopedia Americana* の挿絵や表のないページでは，約1500語の running word が使われている。したがって，2000語は *Encyclopedia American* の約1.3ページ分となり，百万語では700ページに満たないことになる。*Encyclopedia Americana* の1巻のページ数はどの巻も700ページを越えている。いかに500種のテクストから抽出したといえども，全体で百科辞典1巻分に満たないサンプルでは資料としての不安が残ると言わざるを得ない。この点では，5百万語以上を分析した the American Heritage Intermediate Corpus（AHI Corpus）などに劣る。

また，すべてのサンプルが，書かれた資料（written material）から採

取されているので，口語英語の視点の欠如という点で，現代英語の全体像を反映しているとは言い難い。ラジオ・テレビ放送からもサンプリングしている Collins Birmingham University International Language Database（COBUILD）に劣る点である。

本文中でも既に述べたが，文法範疇が tag の形態で表示されるのは不便である。tagged LOB Corpus がより広く使われるためには，データを読み易くする工夫が必要であろう。またマイクロフィッシュや磁気テープではなく tagged LOB Corpus をフロッピーデスクに収録して，検索や分析のし易い利用プログラムを付ければ専門の研究者以外の多くの人達も利用できるのではないか。

このような問題点はあるものの，既に本論でも見てきたように tagged LOB Corpus の言語資料としての有効性はそれらを補って余りがある。本論では，Johansson and Hofland（1989）の4種の頻度表を中心に考察を加えたが，他にも資料の解釈・分析のしかたで多くの考察が可能である。今まで経験的に知られていた言語事実の統計的証明に，また，気付かれずに見過ごされてきた言語現象の新たな発見に力を発揮すると思われる。

註

1) マイクロフィッシュとは多くのマイクロフィルムを一枚のフィルムシートにまとめたものであり，肉眼では読めず，マイクロリーダーと呼ばれる専用の機器を必要とする。Hofland and Johansson（1986）に添付のものは，15cm×10cm の長方形で一枚に208コマのマイクロフィルムを収録している。

2) A～R の分野記号のうち I，O，Q はコンピュータ処理の関係上，意図的に除外している。

3) RB（副詞）に隣接する tag の種類は大変多いので，表8では

Johansson and Hofland（1989）の上位8位までの頻度の tag だけを示した。

参考文献

Carroll, J.B, Davies,P, and Richman,B. (1971) *Word Frequency Book.* New York: American Heritage.

Francis,W.N., and H.Kučera (1982) *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston: Houghton Mifflin.

Garside,R.G., G.N.Leech, and G.R.Sampson (eds.) (1987) *The Computational Analysis of English.* London: Longman.

Hofland,K. and Johansson,S. (1986) *Word Frequencies in British and American English.* Bergen: Norwegian Computing Centre for the Humanities/ London: Longman.

____, ____ (1986) *The Tagged LOB Corpus: KWIC Concordance.* Bergen: Norwegian Computing Centre for the Humanities. Microfiche and computer tape.

Johansson,S. (1978) *Some Aspects of the Vocabulary of Learned and Scientific English.* Gothenburg Studies in English, 42. Gothemburg: Acta Universitatis Gothoburgensis.

____, and Hofland,K. (1989) *Frequency Analysis of English Vocabulary and Grammar.* Vol.1, Vol.2. Oxford: Oxford.