LOB-Corpusにおける カテゴリーの特徴について

——多変量統計解析法による分析——

古　橋　　聰
高　橋　　薫

1　はじめに

コーパス言語学の発展により，欧米ではさまざまな英語に関するコーパス(1)が構築され，多方面からの解析が行われている。

日本においても英語コーパスに関しては，平成5年4月に英語コーパス研究会が発足し(2)，コーパス言語学を次のように定義している。

「コーパス言語学とはコーパスのデザイン，構築のプロセスなどに関する方法論，またはコーパス処理に伴うソフトウェアの開発などの技術論，およびコーパスとコンピュータを活用した言語分析を包括する言語学の一領域」。

コーパス言語学とは言語学の一領域であることはもちろんのこと，情報処理に関する知識をも必要とする分野である。コンピュータの使用は主にコーパス解析のための手段であり，コーパス言語学における主たる研究テーマではないが，文系の研究者にとっての活用方法という点からすると，コーパス言語学の研究分野のひとつであると言い得るであろう。

研究者はコンピュータの使用が必須条件であり，コーパスという特殊なデータベースを扱うために，問題点として明らかにすべき部分も多い。このようなことから，情報処理をも含めた学際的要素が強い分野であるといえるであろう。

ところで，コンピュータ活用の目的のひとつは，文字データの数量化で

ある。さらに，数量化されたデータをどのように解釈するかは統計処理の問題となる。

コーパスの解析の方法論として演繹的，帰納法的な二種類のアプローチが挙げられる。文法学者がある語法の研究をコーパスを利用して行う場合，具体的にその語法の活用部分に着目し，理論についての検証をおこなうのであれば，演繹的な方法論であると言えるし，大規模な英語文章を統計的手法に基づいて解析し，ある総括的な特徴を見いだす方法論は帰納法的であると言える。

コーパスは益々大容量化の方向に向かっており，近々，英国では総単語数1億語からなるコーパスがほぼ完成される等，大規模コーパスの編纂の予定が続々と報告されている。このようにコーパスが爆発的な膨張をしているなかで，コーパスの大規模さの特徴を生かした解析方法が実際には未熟であると言ってよい。

ところで演繹的方法論の観点からすると，コーパスの大容量化はあまり意味を持つものではない。そのような研究報告を見る限り，解析するデータは大容量の中から抽出されたごく一部の文章を持ち出して，小規模な出現頻度の比較等を行うものであることが多い。特にコンピュータにより自動的な検索等ができない手動の語法等の検索の場合には益々その傾向が強いと言える。実際には小規模なデータの解析で必要十分であることが多い。

それにもかかわらず，コーパスは依然として大容量化しているが，巨大化するデータの解析方法については後回しであるといった状況である。

そこでそれを解決する方法論が帰納法的なアプローチである。

コーパスの文字データや品詞情報の出現頻度等を数量化し，個体と変数を設定することによって多変量解析法というアプローチのしかたに注目できる。そして，その多変量解析法を使い，いままでの文法，語法に関する観点とは全く次元の違う特質をコーパスのなかに見いだすことが可能になる。

本論文はコーパスの文字データ等を数量化し，それについて多変量解析法を用いて，帰納法的なアプローチでの研究報告をするものである。

2　多変量解析法の一例

代表的な多変量解析法のひとつである因子分析法はもともと心理学の文系の分野で生まれたものであったが現在も言語，文学，絵画，歴史，法律の研究など，これまで数学とはほとんど無縁と考えられてきた領域に用いられ，しばしば，その有効性を示している。

すでに日本語の文章集について因子分析がなされた報告があるのでそれを紹介する(3)。一例として，因子分析を用いた日本の小説に関する研究を挙げる。これは日本の百人の作家の小説の任意の部分を抽出して，その名詞・動詞の長さ，色彩語・直喩・声喩・人格語・名詞・漢字の使用，句点・読点についての数値化を行い，それらを多変量解析法のひとつである因子分析にかけた結果，説明のつく三つの因子が現れた。ちなみにそこでは第1因子として，「体言型—用言型因子」あるいは「叙事—叙情因子」，第2因子として「修飾語—非修飾型因子」あるいは「青年性—壮年性因子」，第3因子として「会話型—文章型因子」あるいは「短文型—長文型因子」である。

大規模データの英語コーパスについても多変量解析法を活用することは当然考えられることであるが，その分野の研究報告は少ない。

本論文においてはLOB-Corpusについて，そのタグの出現頻度と小説，論述文などのカテゴリー分類に着目して多変量解析法を用い，解釈のつく因子がどのように出現するかを探索的に分析するものである。また，それらの出現した因子にもとづき，各カテゴリーの特徴づけを行うことも可能である。また，その解析の過程でのパソコンを用いたアプリケーションソフトの活用についても言及することにする。

ところで，このような報告を行う場合，多変量解析法とはどのようなものなのかが重要である。そのことについては数学的，演繹的に説明した文献は数多くある。当然，多変量解析法というひとつの数学的なモデルについての知識を与え，そのあとで，その実地への適用法を説明するというプロセスが自然である。

しかし，文系の研究者にとっては難解な統計数学や応用数学を理解する

ことは，かなり骨の折れるところであるので帰納法的に理解しようと考える。すなわち，まず，適用例をみることからはじめ，具体例を通じて多変量解析法のイメージを帰納法的に理解しようとすることになる。

本論文においては文系の研究者が行う報告であるため，多変量解析法の内容それ自身はあえてブラックボックスとして，どのような活用が可能で，そこから何が生まれてきて，どのような解釈ができるのかといった観点で研究報告を進めることにする。

3 LOB-Corpusについて

解析の対象となるLOB-Corpusはイギリス英語の百万語からなるタグ（品詞情報を表す記号）を付したコーパスで，特にタグの種類が百種類以上と豊富なことと，社説，評論に始まり，小説にいたる15種類のカテゴリーからなる点にある。ここでそれらを，論述的内容に近いinformative proseと口語的表現に近いimaginative proseに区別してそのカテゴリーを示す。

INFORMATIVE PROSE

A. Press : reportage
B. Press : editorial
C. Press : reviews
D. Religion
E. Skills, trades and hobbies
F. Popular lore
G. Belles lettres,biography,essays
H. Miscellaneous : government documents, foundation reports, industry reports, college catalogue, industry house organ
J. Learned and scientific writings

IMAGINATIVE PROSE

K. General fiction
L. Mystery and detective fiction
M. Adventure and western fiction
N. Romance and love story
P. Humour

タグは総数139個であるがその詳細な分類についてはすでに中京大学教養論叢第33巻第3号[4]にて説明済みであるのでここでは省略する。ただし，分析の過程で出てくるタグについてはその文法的役割について逐次説明す

る形をとる。

表2にタグ一覧を示す。

タグはいくつかのベースタグより成り，さらに細分化されている。タグの先頭の1文字，あるいは2文字目までが大まかな分類を示している。さらに，接尾辞が以下のように加わる。

＊　限定詞，代名詞，名詞，数詞に付く接尾辞

A：主格　O：目的格　I：単数または複数　S：複数　$：所有格　R：関係詞

＊　動詞に付く接尾辞

D：過去形　G：現在分詞，動名詞　N：過去分詞　Z：3人称単数

＊　形容詞，副詞に付く接尾辞

R：比較級　T：最上級

表2　タグの種類

A. . .　限定詞
　ABL　前位限定語
　ABN　前位数量詞
　ABX　相関接続詞
　AP. .　後位限定詞（AP"）
　　AP, AP$, APS, APS$
　AT　単数冠詞
　ATI　単数，複数冠詞
BE. . .　動詞，助動詞 be
　　BE, BED, BEDZ, BEG, BEM, BEN, BER, BEZ
CC　等位接続詞 (CC")
CD. . .　基数
　　CD, CD$, CD-CD, CD1, CD1$, CD1S, CDS
CS　従位接続詞 (CS")
DO. . .　本動詞，助動詞 do
　　DOD, DOZ
DT. . .　限定詞
　　DT, DT$, DTI, DTS, DTX
EX　存在の there
HV. . .　本動詞，助動詞 have
　　HV, HVD, HVG, HVN, HVZ
IN　前置詞 (IN")
JJ. . .　形容詞 (JJ")
　　JJB, (JJB"), JJR, (JJR"), JJT, (JJT"), JNP
MD　法助動詞
N. . .　名詞
　NC　引用語
　NN. . .　普通名詞
　　NN, (NN"), NN$, NNP, NNP$, NNPS, NNPS$, NNS, (NNS"), NNS$, NNU, (NNU"), NNUS
　NP. . .　単数固有名詞
　　NP, NP$, NPL, NPL$, NPLS, NPLS$, NPS, NPS$
　NPT. .　称号等の名詞
　　NPT, (NPT"), NPT$, NPTS, NPTS$
　NR. . .　副詞的名詞
　　NR, NR$, NRS, NRS$

OD. . . 序数
OD, OD$
P. . . 代名詞
PN. . . 不定代名詞
PN, (PN"), PN$, (PN$")
PP$ 所有限定詞
PP$$ 所有代名詞
PP. . . 人称代名詞
PP1A, PP1AS, PP1O, PP1OS, PP2, PP3, PP3A, PP3AS, PP3O, PP3OS, PPL, PPLS, (PPLS")
QL. . . 修飾語（副詞）
QL 修飾語
QLP 後位修飾語
R. . . 副詞
RB. . . 副詞
RB, (RB"), RB$, RBR, RBT
RI 前置詞的副詞
RN 名詞的副詞
RP 副詞辞
TO 不定詞の to (TO")
UH 感嘆詞
VB. . . 本動詞 (VB")
VB, (VB"), VBD, VBG, VBN, VBZ
W. . . WH-word
WDT. . . 限定詞的
WDT, (WDT"), WDTR
WP. . . 代名詞，疑問詞
WP, WP$, WP$R, WPA, WPO, WPOR, WPR
WPB 副詞的
XNOT not
ZZ 文字

また，コーパスのデータベースは表3のようにパソコンのテキストファイルとして使用可能である。各単語一語一語にタグが付加された一定のフォーマットをとっているため，コンピュータによる文字データの取り込みは容易である。

表3 タグ付き LOB Corpus の一部

^the_ATI film_NN is_BEZ a_AT well-made_JJ variation_NN on_IN that_DT sinister_JJ yarn_NN in_IN
which_WDTR half_ABN the_ATI cast_NN try_VB to_TO persuade_VB heroine_NN that_CS she_PP3A is_BEZ out_RP of_IN her_PP$

4 解析方法

LOB-Corpus の 139 種類のタグの頻度に注目する。タグの頻度は 14 種類[5]のカテゴリー毎に算出するが，タグの頻度はそれぞれのタグの出現回

数を総センテンス数で除したものとする。

実際のタグのカウントはC言語によりプログラミングを行い，カテゴリー毎にそれぞれの総センテンス数で除した数値データを求める。これにより14行139列の行列が完成する。これを多変量解析法のひとつである数量化Ⅲ類により分析する。解析の詳細は以下のとおりである。

4－1　頻度の算出

筆者等は「検索処理における頻度についての問題」(英語コーパス研究会紀要第1号)[6] にて頻度算出方法について次のように結論づけた。すなわち，それぞれのカテゴリーに語彙数の差異があるため，実際の頻度ではなく正規化した「出現率」を用いる場合，その出現頻度を総単語数で除すか，あるいは総センテンス数で除すかが問題となるが，分析の結果，総センテンスで除す方がそれぞれのカテゴリーの持つ特質を顕著化する傾向があることが判明した。たとえば，口語文と文語文を比較した場合，よりお互いの差異が明確に現われ，このような差異は後の多変量統計解析を行うためには有効であると言える。

また，センテンスをどのように区切るかが問題となるが，厳密なセンテンスの定義に基づいて区切り，カウントすることはデータが大量なため不可能であるし，プログラムを作成しパソコンで自動的に判別することも技術的に難しく，今回は単純にピリオド数をセンテンス数と定義づけた。

4－2　C言語の使用について

解析にはC言語を用いた。筆者等は以前にMicro-OCPによる検索処理を行い，Micro-OCPの有効性を明らかにしたが，その後少なからず問題が生じたため，現在ではC言語を用いた解析を中心としている。検索処理等の解析を行うソフト等の必要条件としては第一に誰でも簡単に目的とする処理が行えることと，第二には処理速度である。

Micro-OCPはプログラムの組み易さでは大変優れていると言える。特に句構造の検索等には威力を発揮する。しかしながら，処理速度の点では，解析するデータのレコード長などの形式に左右されることが多く，必ずしも最良の手段であるとは言えない。LOB-Corpusのようにテキストファイ

ルで10メガバイトを越えるようなデータの解析ではかなりの時間を必要とする。

このようなことは，C言語を用いることで解消される。また，C言語は文字関数が非常に豊富であること，ひとたびC言語のプログラムを組んでしまえば解析の速度は飛躍的に向上すること，バッチ処理が可能なため，複数のカテゴリーの解析も簡単に行えるなどの利点がある。

さて，LOB-CorpusはMS-DOS上のテキストファイルになっていて，ハードディスクに収められているので，容易に出し入れが可能である。入力データの出典や原文での行数を明示した数値が各行の定められたコラムにあり，タイトル部は独特の書式になっているため，C言語によりこの部分を自動的に排除して，タイトル部を除いた文章のみをデータとして扱った。

4－3　表計算ソフト・Lotus1-2-3の使用

C言語により14種類のカテゴリーについて139個のタグの出現率を算出した後に表計算ソフトであるLotus1-2-3を用いてデータを整理する。後の多変量統計解析法ではタグの頻度の総カテゴリーの和，そのソーティング後の数値を必要とするため，Lotus1-2-3を活用した。

その場合，必要な知識としては，C言語によって出力された数値データをあらかじめ，Lotus1-2-3のデータとして取り込み易いように設定することであり，それが不可能である場合は，入力のための単純な手作業を強いられることになる。

4－4　多変量統計解析法

タグの頻度についての数値は14行（カテゴリーの種類），139桁（タグの種類）のデータ数となる。これを多変量統計解析法の一つである数量化Ⅲ類を用いて解析する。この解析方法は予想すべき外的基準のない場合の数量化法の一つであり，個体の種々のカテゴリーの反応の仕方に基づいて，個体とカテゴリーの両方を数量化し，さらにその数量を用いて分類を行なおうという方法である。このことについてはさらに以下のように説明できる。

すなわち，タグ（個体）のカテゴリー（変数）に対する使用頻度数（反応）に基づいて出現頻度（数値）を与え，似ているもの同士を近くに，似ていないものを遠くに空間的に配置しようとする方法論である。

あるタグ（個体）がともに反応を示したカテゴリー（変数）は，そのタグ（個体）にとって何等かの意味で互いに似ているはずであり，また，同様に，あるカテゴリー（変数）にともに反応を示したタグ（個体）にもそれがいえる。この類似性を数量化の手がかりとして，互いに類似したタグ（個体）には似た数値を，異なったタグ（個体）にはできるだけ異なった数値を付与する。

コーパス言語学の立場から注目すべき点は，タグ間，カテゴリー間の類似性を手がかりとすることによって，反応が依存している潜在的な根拠を発見することができ，文章について体系的な分類が可能となると考える。

すなわち，2章の多変量解析法のなかで示した日本語文章の因子分析による解析のように，説明のつく因子が LOB-Corpus の解析によって現れることが十分予想される。

さらに，その出現した因子とタグあるいはカテゴリーの関係を明らかにすることによって，英語文章を体系的に分類することが可能になると考えられる。

5　分析結果

5－1　タグの出現頻度について

14個のカテゴリーについて，139個のタグの出現頻度を正規化して，頻度の高い上位20個のタグについて，数量化Ⅲ類の解析を行なった。言語学の立場からすると，139個のタグが設定されている以上，そのすべてのタグについての体系的な分析結果に興味の涌くところであり，残りの119個の出現頻度のデータを完全に無視することには抵抗はあるが，頻度の高いタグについてのみ解析することで，使用の極めて少ない特殊なタグが解析のための誤差になることを避けるため，上位20個のタグについての解析を進めることにする。

表4に20個のタグを示す。

表4 タグの出現頻度

（総カテゴリーについて，総センテンス数で除したもの）
数値2.99（順位1）は1センテンス中2.99個の割合で単数普通名詞が出現することを示す。

順位	タグ	数値	品詞	順位	タグ	数値	品詞
1.	NN	2.99	(単数普通名詞)	11.	AT	0.53	(a, an, every)
2.	IN	2.47	(前置詞)	12.	VBD	0.49	(動詞過去形)
3.	ATI	1.42	(the, no)	13.	CS	0.38	(従属接続詞)
4.	JJ	1.33	(形容詞)	14.	PP$	0.35	(所有限定詞)
5.	NNS	1.01	(複数普通名詞)	15.	TO	0.32	(不定詞)
6.	CC	0.75	(等位接続詞)	16.	MD	0.31	(法助動詞)
7.	RB	0.71	(副詞)	17.	VBG	0.26	(Ving)[*7]
8.	NP	0.70	(単数固有名詞)	18.	BEZ	0.26	(is)
9.	VB	0.67	(本動詞)	19.	PP3A	0.25	(he, she)
10.	VBN	0.54	(動詞過去分詞形)	20.	CD	0.25	(基数)

頻度の最も高いタグは単数形普通名詞で1センテンス中2.99個の割合で出現する。LOB-Corpusでは普通名詞は表2より，13種類に分類されている。接尾辞としてS：複数，$：所有格，NNPは文頭大文字で始まるEnglishman，German，etc.，NNUは測量単位，hr，lb等となっている。

また，別に固有名詞としてNPがあり，さらに上述のような13種類の分類がある。NPLでAbbey，Bridge等の大文字で始まる場所に関する名詞，NPTとしてArchbishop，Captain等の大文字で始まる称号等を示す名詞が代表的な名詞である。

このことにより，NNは上述以外の単数普通名詞であると言える。また，その複数形が5位にあり，両者で普通名詞は1センテンス中4個である。さらに単数固有名詞が8位であり，このことより，当然ながら文章は主に名詞で成り立っていることがわかる。次に頻度の2位は前置詞である。前置詞がこのように上位にある理由としては，名詞のような細かな分類がなされていないためで，他の動詞，限定詞，接続詞，代名詞において細かな分類があるので，それぞれにおいては頻度が分散されるなかにあって，前置詞だけはただひとつであるためである。

第3位の ATI は定冠詞等である。これは普通名詞が頻度が高く，それと共起するためである。このことが11位の AT の不定冠詞等においても言える。第4位の形容詞は他に限定的用法のみに用いるもの chief, entire, main, 比較級，最上級の形容詞，固有名詞から派生した，English, German 等を除いた形容詞であり，この形容詞も冠詞同様，限定的な用法が主であると推測できる。第6位の等位接続詞は第13位の従属接続詞と合わせると接続詞としては1センテンス1個程度の出現率となる。

次に注目すべきものは動詞，準動詞，法助動詞である。

動詞は本動詞（9位），過去分詞形（10位），過去形（12位），is（18位）となっている。ちなみにこれらすべての頻度の合計は1センテンス中2個程度の頻度となる。また，準動詞としては，不定詞（15位），現在分詞・動

表5　各カテゴリーのタグの頻度

（総センテンス数で除した値）
（.6 は 0.6 を表す）

	"A"	"B"	"C"	"D"	"E"	"F"	"G"	"H"	"J"	"K"	"L"	"M"	"N"	"P"
NN	3.0	3.2	3.5	3.2	3.8	3.3	3.6	4.8	4.6	2.0	1.9	1.8	1.8	1.7
IN	2.5	2.6	2.8	2.5	2.8	2.7	3.3	4.4	4.0	1.6	1.4	1.4	1.3	1.3
ATI	1.4	1.6	1.5	1.8	1.7	1.6	1.7	2.3	2.3	.9	.8	.8	.8	.7
JJ	1.1	1.4	1.9	1.2	1.6	1.4	1.7	2.0	2.0	.9	.8	.9	.7	.7
NNS	1.0	1.3	1.0	.8	1.4	1.3	1.3	2.2	1.7	.6	.4	.6	.4	.4
CC	.6	.7	.9	.8	.9	.9	1.0	1.2	1.0	.6	.5	.5	.5	.5
RB	.5	.7	.9	.7	.8	.8	.9	.8	.9	.7	.6	.5	.5	.6
NP	1.5	.8	1.2	.6	.6	.7	.8	.5	.6	.4	.5	.6	.4	.5
VB	.5	.8	.6	.7	.9	.7	.7	.9	.7	.7	.5	.5	.5	.6
VBN	.5	.6	.5	.6	.7	.6	.6	1.0	.9	.4	.3	.3	.3	.3
AT	.6	.5	.8	.5	.6	.6	.7	.6	.6	.4	.4	.4	.4	.3
VBD	.5	.2	.3	.2	.2	.4	.6	.2	.2	.8	.8	.7	.8	.8
CS	.3	.5	.3	.5	.4	.4	.5	.5	.6	.3	.3	.3	.2	.3
PP$	.3	.3	.5	.3	.2	.4	.5	.3	.2	.5	.3	.4	.4	.4
TO	.3	.4	.3	.3	.4	.4	.4	.4	.4	.3	.2	.2	.2	.3
MD	.3	.4	.2	.3	.4	.3	.3	.5	.4	.3	.3	.2	.2	.3
VBG	.2	.3	.2	.2	.3	.3	.3	.3	.3	.3	.3	.3	.3	.3
BEZ	.2	.4	.4	.5	.4	.3	.3	.3	.4	.1	.1	.1	.1	.1
PP3A	.2	.1	.2	.1	.1	.2	.3	.1	.1	.5	.4	.2	.4	.6
CD	.4	.3	.2	.3	.4	.3	.2	.7	.6	.1	.1	.1	.1	.0

名詞（17位）となっている。ただし，現在分詞，動名詞は文法上区別すべきものであるが，LOB-Corpusにはその区別がない。そのため，個々には頻度は低くなる。

ところで，ここまでのタグの頻度は全てのカテゴリーについての頻度であり，実際にはカテゴリー毎に頻度に差異が現れるはずであり，数量化Ⅲ類により，その差異を総括的に解釈することが可能であるとも言える。ここで各カテゴリーのタグの頻度を表5に示す。

5－2　数量化Ⅲ類による分析結果

まず，反応が依存している潜在的な根拠を発見することにある。ここで数量化Ⅲ類による分析結果を表6に示す。縦にカテゴリー，横に第1，2，3と3つの有効な成分を示す。

表6　数量化Ⅲ類による分析結果

固有値・固有ベクトル
寄与率・累積寄与率
タグ：上位20個

カテゴリー	第1成分	第2成分	第3成分
A	0.0164834	0.6228050	0.5884380
B	− 0.1442979	0.0563109	− 0.2524279
C	− 0.0479257	0.5540510	− 0.3638556
D	− 0.1502096	0.0070070	− 0.4693444
E	− 0.2128764	− 0.0821324	− 0.1150804
F	− 0.0671009	− 0.0114931	0.0006689
G	− 0.0166700	0.0086576	− 0.0165896
H	− 0.3621966	− 0.3923394	0.4054971
J	− 0.3258331	− 0.2211968	0.0741961
K	0.3570041	− 0.2322717	− 0.0769020
L	0.3623384	− 0.0866668	0.0528840
M	0.2305501	0.0314748	0.1693526
N	0.3852964	− 0.1388976	0.0749579
P	0.4541460	− 0.1018128	− 0.1043266
固有値	0.49476	0.085210	0.038054
寄与率	0.72366738	0.12463297	0.05566001
累積寄与率	0.72366738	0.84830036	0.90396036

最初に注目すべきことは，それぞれの成分の寄与率である。それぞれの成分の意味付けを行うまえにそれぞれの成分が潜在的根拠全体のどの程度を説明しているかの数値として第1成分より順に 0.72，0.12，0.06 である。

これを 72%，12%，6%とみなすことができる。合わせて，90% となり，第4成分が4%であることより，この第3成分について今後考察することが妥当である。

次に A から P のカテゴリーの第1，2，3成分の数値に注目する。

それぞれの絶対値が大きければ，その成分での特徴が顕著であるといえる。また，符号のあるなしでその成分でのプラス方向，マイナス方向といった特徴を示している。とりあえず，第1成分を横軸に第2成分を縦軸にとって散布図を図1に示す。また，個体すなわち，タグそれぞれの数量

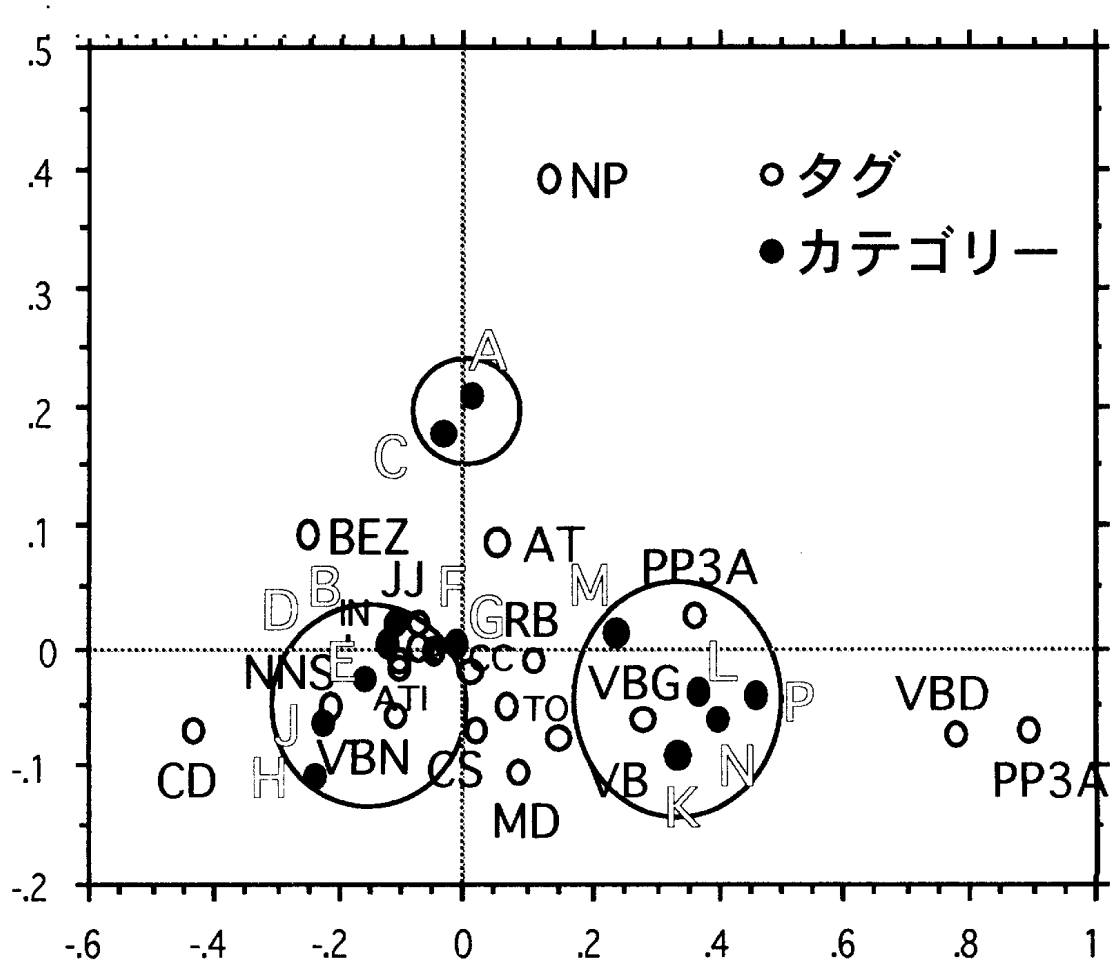

図1　散布図

（第1成分：横軸　第2成分：縦軸）

化得点の数値により，同じ散布図上に上位20個のタグを表示する。また，各タグは表7の因子得点を参考に散布図に位置される。

これにより，カテゴリーは3つの独立したグループに分かれる。

散布図の右側にimaginative proseのすべて，すなわちK，L，M，N，Pが，また，informative proseはA，Cが散布図の上部に，他のB，D，E，F，G，H，Jは中心よりやや左寄りにひとつのグループを形成している。

また，第1成分の固有ベクトルの数値の高いものはPP3A（he，she），VBD（動詞過去形），最も低いものはCD（基数）であり，このようなタグが第1成分との関連が特に高く，後の因子の解釈の時に役立つものである。

また，imaginative proseの付近に現れているタグはこのグループ内で

表7　数量化得点

	第1成分	第2成分	第3成分
AT	0.049206	0.085949	− 0.026731
ATI	− 0.10659	− 0.0164302	− 0.08074
BEZ	− 0.25019	0.095127	− 0.23207
CC	0.011317	− 0.022174	− 0.023383
CD	− 0.43468	− 0.072508	0.20949
CS	0.018925	− 0.073048	− 0.046570
IN	− 0.10452	− 0.012158	0.025962
JJ	− 0.072610	0.020641	− 0.040237
MD	0.083273	− 0.10383	− 0.028886
NN	− 0.072081	− 0.0016405	0.0018588
NNS	− 0.21110	− 0.049594	0.087487
NP	0.13276	0.39154	0.059569
PP$	0.36631	0.025746	− 0.075142
PP 3 A	0.89964	− 0.072530	− 0.015202
RB	0.10845	− 0.0096176	− 0.092652
TO	0.067656	− 0.051002	− 0.030870
VB	0.15053	− 0.077265	− 0.057767
VBD	0.77945	− 0.073386	0.12684
VBG	0.28454	− 0.062532	0.053814
VBN	− 0.11154	− 0.057214	0.021986

の使用頻度が高いことを示している。そのタグとして，PP$（所有限定詞: my, your, etc），VBG（現在分詞，動名詞）がある。CD（基数）は第1成分のマイナス側を特徴づけているタグである。また，A，C以外の informative prose のグループ付近にあるタグとしては，BEZ (is, 's)，NNS（複数普通名詞），VBN（過去分詞），ATI (the, no)，IN（前置詞: about, above, etc），JJ（形容詞），NN（普通名詞）が挙げられる。

第2成分に関しては NP（単数固有名詞）が高い値を示している。

また，第3成分はプラス方向に CD（基数），VBD（過去形），またマイナス方向に BEZ（is, 's），RB（副詞）が挙げられる。

6 考　察

6－1　因子の解釈

それぞれの成分を特徴づけているタグについての共通の役割について考察することによって，その因子の解釈が可能となる。

第1因子は PP3A，VBD 対 CD，BEZ，NNS 等から「口語的文章型」対「論述的文章型」と名付ける。英語に限らず，文章を体系的に分類しようとする時には，その文章が口語的であるか，あるいは論述的であるかといった発想は誰にでも思い浮かぶところである。ここで多変量解析によって，漠然と抱いていた文章の体系的分類のイメージが支持されたといってよい。具体的にはカテゴリーの分類のところで示した informative prose と imaginative prose がこの第1成分により，区分されることが判明した。

また，第2成分については，特徴を示しているタグが出現頻度の高い上位20個のなかでは，単数固有名詞が高い数値を示しているのみであり，この成分の特徴を特定するのが難しいため，出現頻度の高い上位50個のタグに分析を拡張した結果，次のタグに固有ベクトル値が高く現れた。

NPT　称号等の名詞　：Archbishop, Captain
JNP　形容詞（大文字で始まるもの）：English, German, etc.
WPR　〃　（主格，目的格）：that, who
VBZ　動詞（3人称単数現在形）
HVZ　has, 's

また，マイナス側には以下のタグが現れた。

PP 2	you，thou，thee
PP 1 AS	we
PP 1 A	I
DTI	any，some
BE	be

これらのタグより，第2成分を「特殊主題型」対「一般的主題型」と解釈できる。すなわち，特殊主題型は固有名詞を中心とした特定の主題に関する名詞，そこから派生した形容詞，また，固有名詞が単数であることが多いため，それに関連する述語動詞で説明がつく。

その対極として，主題がごく一般的内容となるタグとして，代名詞が多くなる。さらにそのような特質を持ったカテゴリーとして，特殊主題型ではA（Press：reportage），C（Press：reviews），また，一般的主題型としてはH（Miscellaneous）がある。Hは政府書類，学生便覧，工業報告書，大学紹介，社内報などで，規則集であるとか，報告事項，連絡事項といった内容であるため，一般的主題型の典型的な文章内容である。

第3成分についても，第2成分同様上位50個のタグに拡張して考察する。それによると，プラス方向にCD（基数），VBD（過去形）のほかにNPT（称号等の名詞），JJB（限定的用法のみの形容詞），マイナス方向にBEZ（is，'s），RB（副詞）のほかにVBZ（動詞のs形），QL（副詞）が挙げられる。

また，プラス方向に特徴が顕著なカテゴリーとしてはA（Press：reportage），マイナス側ではC（Press：reviews），D（religion：宗教）となる。

これらのことから第3成分についても解釈すべきところであるが，現段階では挙げたタグのなかに共通性を見いだすことは困難であり，今後の課題とするところである。いずれにしても，寄与率が6%程度の成分なので考察に値しない成分であるかもしれない。

以上のことから，特に第1，2成分に特徴を限定して，カテゴリーについて，総括すると図2のような位置関係となる。

原点付近にあるF：Popular lore　G：Belles lettres，biography，essaysは文章を構成しているタグが全てのカテゴリーの文章全体からみ

原点付近　F : Popular lore
　　G : Belles lettres, biography, essays

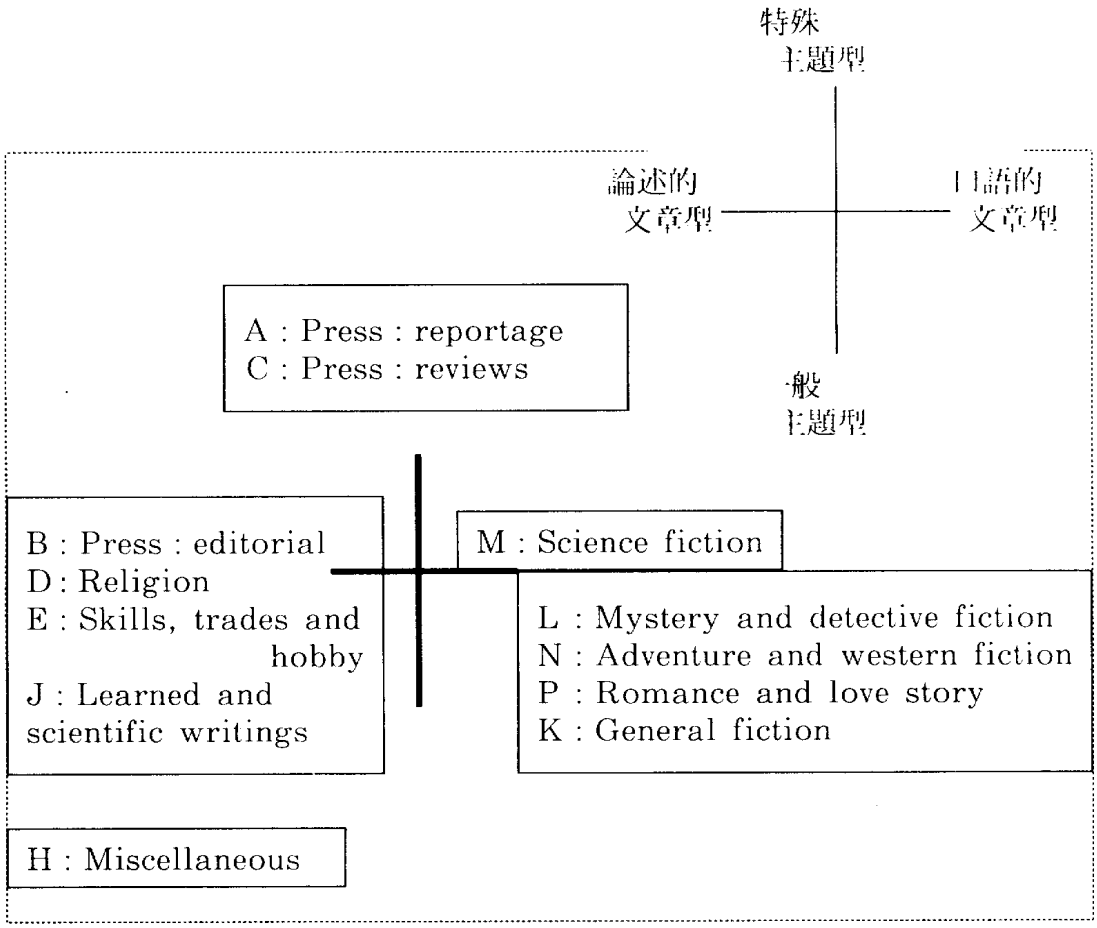

図2　カテゴリーのグルーピング

ると平均的に出現していると言える。すなわち，文章形態に口語文，論述文であると印象づけるような部分が少ないと言える。

また，第1成分では小説の中でもM: Science fiction が論述的傾向を持つことも納得がいくところである。

第2成分については，報道関係の文章で固有名詞が多く，それに伴って三人称単数現在形の動詞等が共起することが特殊主題型の大きな特徴をなしている。

小説類ではM: Science fiction が中立的である。他の小説類はむしろ一般的主題型傾向が強いと言える。これは固有名詞を使うよりも，you, I, we を使用することが多いことから，そのような傾向になるためであると推測できる。

7 最後に

今回の解析で一つの論点となるのは，解析したタグを使用頻度の高い上位20個に限定したことにある。そのように限定したことによって，むしろ，他の重要な因子の出現を無くしてしまった可能性も考えられる。しかし，139個を最初の段階から，解析の個体として考えては，解釈が難しいのが実状である。タグを絞り込むことによって第一段階の解釈が今回報告できたわけであり，今後，さらに段階的にタグを増し解析を進めていくことにする。

また，タグの設定については次のような考え方も成り立つ。

多変量解析法はその結果は変数と個体がもたらすものであるため，今回の報告において，139個のタグを個体と設定したことによって，上述のような成分が現れてきた。しかし，タグがこれほどの種類がないようなコーパスについて解析を行えば，違った成分が浮きでてくることも考えられる。たとえば，LOB-Corpusでは第2成分に「特殊主題型」対「一般的主題型」の成分が現れたが，仮に名詞をひとつのタグとして，まとめて解析を行えば，このような第2成分が現れないことも十分考えられる。しかしながら，現存する英語コーパスのなかでこのような豊富なタグが付されたコーパスを十分生かすためにも今後，タグはさらに細分化する必要さえある。そして，最良のタグ数でどのような成分が浮きでてくるか考察すべきであろう。

注

(1) 日本において最も入手しやすいコーパスとして，Norwegian Computing Centre for the Humanitiesより発売されているICAME CORPUS COLLECTIONがある。これにはLOB Corpus, Helsinki Corpus, London-Lund Corpusのテキストデータが CD-ROMに納められていて，容易にデータを引き出せる。

(2) 英語コーパス研究の概要については，英語コーパス研究会会長である大阪大学斎藤俊雄教授が英語青年（1994年2月号）にて「英語コーパス研究の最近の動向」として掲載されている。

(3) 「因子分析法」 安本美典・本多正久著（1981）培風館より引用。

(4) 中京大学教養論叢第 33 巻第 3 号 (1992) pp. 117－146 に「タグ付きコーパスの句構造の解析について」を掲載。主に受動態といった文体の出現頻度に注目した報告である。

(5) カテゴリーは 15 種類であるが，入手したコーパスの最後の R：Humour はデータが不備のため，解析不可能である。

(6) 英語コーパス研究　第 1 号 PP. 35－48 に掲載。

(7) Ving とは現在分詞，動名詞を示す。ただし，文法上，現在分詞，動名詞は同類とは考え難いが，LOB-Corpsu のタグ設定においては，自動タグ付けでの区別の困難さからこのようになっている。これはこのコーパスの改良点として指摘できることである。他にも，this，that などの指示代名詞が単体で動詞の目的語となった場合にも，限定的用法と区別が無い等，若干の問題点がある。

参考文献

古橋聰，高橋薫「タグ付きコーパスの句構造の解析について」『中京大学教養論叢』第 33 巻，第 3 号 (1993)，pp. 117－146

三浦敏明『英語副詞の研究―副詞の多様性―』文化書房博文社 (1991) Oxford University Press

長瀬真理訳『Micro-OCP 文章解析プログラム』沖田電子技研 (1988)

斉藤武生，鈴木英一『冠詞・形容詞・副詞』研究社 (1984)

高橋薫「英語教育のパソコン利用における諸問題」計測自動制御学会，『教育工学論文集』(1991)

高橋薫，古橋聰「タグ付きコーパスの文章解析について」中部地区英語教育学会紀要 22 号 (1992)

高橋薫，古橋聰「コーパス言語学の英語教育への応用」中部地区英語教育学会紀要 23 号 (1993)

竹蓋幸夫『コンピュータの見た現代英語―ボキャブラリーの科学―』エデュカ株式会社 (1981)

竹蓋幸夫『英語科の CAI』エデュカ株式会社 (1987)

竹蓋幸夫『英語教師のパソコン』エデュカ株式会社 (1992)

田中豊，脇本和昌『多変量統計解析法』現代数学社 (1983)

渡辺登士『英語の語法研究・十章』大修館 (1989)

Aijmer, Karin & Altenberg, B (eds.), *English Corpus Linguistics : Studies in honour of Jan Svartvik*, Longman, GBR. (1991)

Armstrong, Susan (ed.), *Using Large Corpora, ACL-MIT Press Series in Computational Linguistics*, MIT. (1994)

Atkins, B. T. S and A. Zampolli (eds.), *Computational Approaches to the Lexicon*, Oxford U. P., GBR, (1994)

Fries, Udo et al. (eds), *Creating and Using English Language Corpora* :

Papers from the fourteenth International Conference on English Language Research on Computerized Corpora, Language and Computers. (1994)

Hockey, Susan and N. Ide (eds.), *Research in Humanities Computing: Research in Humanities Computing*, Oxford at the Clarendon Pr., GBR, Vol. 3: Papers from the 1991 ACH-ALLC conference. (1994)

ICAME NEWS Nos. 1, 2, 3. (1978–1979)

ICAME Journal–Computers in English Linguistics, Nos. 4–18, Norwegian Computing Centre for the Humanities, Bergen. (1980–1994)

John Sinclair, *Corpus Concordance Collocation*, Oxford University Press (1991)

Kaoru Takahashi, *Problem Areas in the Tagged LOB Corpus*, Journal of Toyota College of Technology. (1982)

Karin Aijmer and Bengt Altenberg, *Corpus Linguistics*, Longman. (1991)

Kytö, Merja et al. (eds.), *Corpora across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora St Catharin's College Cambridge, 25–27 March 1993*, Language and Computures, 11) (Editions Rodopi). (1994)

Martin Vide, *Current Issues in Mathematical Linguistics: Selected paper from the 1st International Conference on Mathematical Linguistics*, Tarragona, Spain. (1993)

Saint Dizier, Patrick, *Advanced Logic Programming for Computational Linguistics*, Academic Pr. (1994)

Stig Johansson, *Computer Corpora in English Language Research*, Norwegian Computing Centre for the Humanities, Bergen. (1982)

Stig Johansson and Knut Hofland, *Frequency Analysis of English Vocabulary and Grammar, Based on the LOB Corpus, Volume 1 : Tag Frequencies and Word Frequencies*, Clarendon Press, Oxford. (1990)

Stig Johansson and Knut Hofland, *Frequency Analysis of English Vocabulary and Grammar, Based on the LOB Corpus. Volume 2: Tag Combinations and Word Combinations*, Clarendon Press, Oxford. (1989)

Jan Svartvik (ed.), *Directions in Corpus Linguistics, Trends in Linguistics*, Proceedings of Nobel Symposium 82. (1991)

Walker, Donald E., A. Zampolli and N. Calzolari (eds.), *Automating the Lexicon: Research and practice in a multilingual environment*, Oxford U. P., GBR. (1994)

謝辞

本研究は文部省統計数理研究所の1994年度・共同研究A「タグ付き英語文章

コーパスの統計的解析」ならびに文部省科学研究費「英語コーパスのパソコンによる文章解析」の一環であり，統計的指導については統計数理研究所・村上征勝氏の協力を得た。記して謝意を表す次第です。