Bank of English と British National Corpus における英国全国紙の POS タグ分布

高　見　敏　子

1．はじめに

イギリスの高級紙と大衆紙は，どちらも「新聞」というジャンルに分類される媒体でありながらそのスタイルが対照的であることから，しばしば異なる文体の例として取り上げられてきた。新聞の性質上，同じ日に同じ事柄に関する報道がなされることがしばしばあり，そうした記事は，例えば Crystal and Davy（1969）に見られるように，内容の類似性が高くしかも異なる文体で書かれた文例が随所に見られる興味深い実例になる。純粋に文体の比較を行いたい場合，「内容が同じで文体が異なる」複数のテクストが理想的なデータになるが，そのようなテクストを現実に探すのは実はなかなか難しい。高級紙と大衆紙の場合も内容が完全に同一ということはないが，比較的手軽に得られるテクストとして上記の理想にかなり近いものであるため，対照例としてよく取り上げられるのであろう。

一方，現在の代表的な大規模英語コーパスである Bank of English（BoE）と British National Corpus（BNC）はどちらもイギリスで構築されたコーパスで，ともにイギリスの新聞は重要な構成要素となっている。新聞にも経済紙・地方紙・夕刊紙などさまざまなものがあるが，イギリスの一般的な全国日刊紙に限っても各コーパスに大衆紙２紙，高級紙３紙の計５紙が含まれている[1]。本研究に用いた時点での BoE では各紙3,000万語前後，BoE に比べると規模が小さい BNC でも100万語前後のコーパスサイズがあり，個々の記事を比較する場合とは桁違いの量のテクストをデータとして利用することができる。

Crystal and Davy（1969）のように，同じ事柄を扱った高級紙と大衆紙の記事を２つ並べて丹念に内容を比較していく方法が文体の違いを精査する良い方法であることは確かである。しかし，１つの記事の比較だけでは，そこで見られた特徴がその記事にとどまらず他の多くの記事にも当てはまるその新聞の特徴と言えるかどうかにつ

[1] Bank of English については本稿で用いたデータを得た2001年３月時点での記述であり、2012年３月現在は大衆紙１紙、高級紙３紙となっている。

いてはよくわからないという弱点もある。そこで，そうした研究を補う方法として上述の大規模コーパスの英国全国紙サブコーパスを利用することが考えられる。

大規模コーパスに含まれる高級紙と大衆紙のサブコーパスを用いることにも問題がないわけではない。収録時期や収録期間が揃っておらず，同じ事柄を扱った記事が選ばれて集められているわけでもない。つまり，扱われている内容そのものが異なっているので，高級紙と大衆紙の比較を行っても，そこで見られる差は必ずしも文体上の違いとは言えず，内容自体の違いに因る面が無視できないのである。

しかし，実際に同じ日の高級紙と大衆紙を比べてみると，同じ事柄が多くの新聞で共通してある程度のまとまった記事として扱われるのはむしろ特に大きなニュースがある場合に限られ，実はさほど多いケースではない。そのようなニュースだけを集めて比較した場合，大事故や大事件等の報道に偏ることが予想され，例えば大衆紙ではあまり扱われない国際情勢の記事や，高級紙ではあまり扱われないセレブのスキャンダルは対象から外れてしまうことになる。しかし，こうした記事もそれぞれの新聞を特徴づける要素であるので，調査対象にまったく含めないというのも適切でない面がある。また，実社会においては文体の違いは扱う分野や内容とかなり密接に結びついており，内容と文体を完全に区別することは実態に即していないとも言える。既存の大規模コーパスは「内容を揃えた文体比較」にならないのは事実であり，その点には常に留意する必要があるが，別の視点から見れば実際の高級紙と大衆紙の有り様に近いデータと考えることもできるので，大規模コーパスの利用は高級紙と大衆紙の比較研究における有力な方法の１つと言えよう。

BoE と BNC に共通する重要な特徴の一つは，両者とも POS（part-of-speech）[2] タグが付与されているということである。コーパスに POS タグが付与されるようになったことで，同じ語形（例えば *smile*）が異なる品詞（例えば名詞と動詞）で用いられる場合についても区別して検索することができるようになった。また，個々の語に関する研究に加えて，POS タグによって語をグループ化して計量的に捉えることができるようになり，ジャンルと POS の分布の間に関連があることが知られるようになってきた。

例えば，Leech et al.（2001: 300）の対数尤度比を示した表から BNC の written English と spoken English との比較でそれぞれ POS タグの各上位５位までを示したものが表１・表２である[3]。表から，両ドメインを比較すると，書き言葉では話し言葉に比べて固有名詞，名詞，形容詞，冠詞が相対的に多く使われ，話し言葉では書

[2] BoE や BNC のドキュメントでは word class と part of speech の両方の用語が用いられている。本稿でも特に両者を区別しないが，主として POS の表記を用いることにする。

[3] 各タグの説明は同書 pp.20-23によるものである。同書の記述には本稿で用いた BNC に採用された UCREL C5 Tagset ではなく，より区分の細かい UCREL C6 Tagset が用いられている。このため表１～４のタグ表記は後の節で示す本稿の BNC の分析結果のタグ表記とは異なっている。

き言葉に比べて間投詞，分類外の語，代名詞（*I, you, it*）が相対的に多く用いられていることがわかる。

表１：BNC の written English に多く現れる POS タグ

Tag	Description
NP1	singular proper noun（e.g. *London, Jane*）
NN1	singular common noun（e.g. *book, girl*）
JJ	general adjective
NN2	plural common noun（e.g. *books, girls*）
AT	article（e.g. *the, no*）

表２：BNC の Spoken English に多く現れる POS タグ

Tag	Description
UH	interjection（e.g. *oh, yes, um*）
FU	unclassified word
PPIS1	1st person singular subjective personal pronoun（*I*）
PPY	2nd person personal pronoun（*you*）
PPH1	3rd person sing. neuter personal pronoun（*it*）

また，同書の informative writingとimaginative writing の表（p.304）から，それぞれ各上位４位までを示したものが表３・表４である。informative writing で imaginative writing に比べた場合に相対的に多い POS タグは名詞，基数詞，形容詞，前置詞の *of*，単位などであり，imaginative writingでinformative writing に比べた場合に相対的に多い POS タグは人称代名詞（*I, you, he / she*）とその所有格（*my, your, our* など）と動詞の過去形であることがわかる。

表３：BNC の informative writing に多く現れる POS タグ

Tag	Description
NN2	plural common noun（e.g. *books, girls*）
MC	cardinal number, neutral for number（*two, three, …*）
JJ	general adjective
IO	*of*（as preposition）
NNU	unit of measurement, neutral for number（e.g. *in, cc*）

表４：BNC の imaginative writing に多く現れる POS タグ

Tag	Description
PPIS1	1st person singular subjective personal pronoun（*I*）
PPY	2nd person personal pronoun（*you*）
VVD	past tense of lexical verb（e.g. *gave, worked*）
PPHS1	3rd person singular subjective personal pronoun（*he, she*）
APPGE	possessive pronoun, pre-nominal（e.g. *my, your, our*）

表１～４を比べてみると，書き言葉と話し言葉，informative writing と imaginative writing の対比で，一部に対応する関係があることが興味深い。書き言葉とinformative writing では名詞と形容詞が，話し言葉と imaginative writing では人称代名詞の使用が相対的に多いという点に類似性が見られる。

イギリスの高級紙と大衆紙はどちらも新聞という同じジャンルに 属しているが，一般に対照的な文体で書かれていると認識されている。そして用いられる言葉についても例えば Crystal and Davy（1969: 187-8）において高級紙は formality, technical terminology に特徴があり，大衆紙は informality, colloquialism, idiom に特徴があるとされ，Jucker（1992: 7）においても取り上げた例の語彙について，高級紙は“specialised and technical”，大衆紙は“colloquial and informal”であるとされている。このような対立は書き言葉・話し言葉や informative writing・imaginative writing の対立などに通ずる部分があり，したがって POS タグの頻度においてもその分布に特徴的な差が見られる可能性がある。Takami（2004）では高級紙と大衆紙に特徴的にみられる形容詞に絞って取り上げたが，本稿では２つの大規模コーパスBoE と BNC の英国全国紙サブコーパスを用いて，イギリスの高級紙と大衆紙における POS タグの分布を調べ，その特徴を明らかにしたい。

２．イギリスの全国日刊紙

はじめにイギリスの全国日刊紙に関する基本的な事柄を簡単にまとめておきたい。現在（2012年３月）イギリスの全国日刊紙は合わせて10紙あり，このうち高級紙 は *The Times, The Independent, The Guardian, The Daily Telegraph, Financial Times* の５紙，大衆紙は *The Sun, Daily Mirror, Daily Star, Daily Mail, Daily Express* の５紙である。この他の大衆紙として1995年に廃刊になった *Today* があり，同紙の廃刊以前は大衆紙が６紙あった。

高級紙と大衆紙の区分は以前は紙面の大きさとも一致しており，高級紙は broadsheet，大衆紙はtabloidとも呼ばれるのが一般的であった。しかし2003年に *The Independent* と *The Times* が broadsheet 判に加えて compact 判という tabloid に近い判型でも発行するようになり，翌年には broadsheet 判を廃止して compact 判のみの発行となった。*The Guardian* もこの動きに倣い，2005年から broadsheet 判をやめて tabloid に近い Berliner 判での発行となった。このため現在でも broadsheet であるのは *The Daily Telegraph* と *Financial Times* の２紙だけとなっている。

イギリスの新聞の区分の仕方には，高級紙と大衆紙というよく知られている２区分の他に，例えば Jucker（1992）にも用いられた up-market, mid-market, down-market という３区分がある。これは読者の社会階層分布に基づく区分で，高級紙・大衆紙という区分との対応で言えば，up-market は高級紙にあたり，mid-market

と down-market は大衆紙の下位分類にあたる[4]。具体的には *The Sun, Daily Mirror, Daily Star* の 3 紙が down-market，*Daily Mail, Daily Express, Today* の 3 紙が mid-market に分類される。大衆紙という言葉でひとくくりにされてきた新聞の中にも読者層という観点でみると違いがあり，Jucker（1992）でその違いと名詞句の構造という言語的特徴との関連が示されたように，他の言語的特徴にもその違いが見られる可能性があるので，高級紙と大衆紙という視点に加えて，up-market, mid-market, down-market という視点も併せて持っておきたい。

3．使用コーパス

本節では本稿で用いたイギリスの 2 つの大規模コーパスである Bank of English（BoE）と British National Corpus（BNC）のイギリス全国日刊紙に関する基本的な情報をまとめておきたい。なお，経済紙である *Financial Times* については残りの一般紙とは性格が異なると考えられるため本稿の研究対象から除外した。

3．1　Bank of English（BoE）

3．1．1　Bank of English の英国紙サブコーパス

Bank of English（BoE）はオンライン・コーパスであり，さらに時々で更新されてきたため，アクセスした時期によって内容が異なる。本稿の分析は2001年 3 月に得たデータに基づく。当時の BoE 英国全国紙サブコーパスは表 5 の 5 つであった[5]。

表 5 ：Bank of English の英国全国紙サブコーパス

サブコーパス	含まれる新聞名	コーパスサイズ（テクスト数）	発行年
sunnow	*The Sun* *The News of the World*	31,786,908 （597）	1997-2000
today	*Today*	26,606,537 （794）	1992-1995
indy	*The Independent*	30,386,339 （260）	1990, 1995, 1998, 1999
guard	*The Guardian*	32,339,864 （332）	1995, 1999
times	*The Times* *Sunday Times*	31,110,198 （208）	1995, 1996, 1999, 2000

[4] 詳細は Jucker（1992: 50）または Jucker（1992: 273）に基づいて作成したイギリス全国日刊紙の読者の社会階層構成のグラフ（高見 2003: 76）を参照されたい。

[5] 「コーパスサイズ、テキスト数、発行年に関する情報は Jeremy Clear 氏による英国 Birmingham 大学内の英語研究に関するメーリングリストへの投稿記事 ‘Bank of English update’（2000年11月29日付）に拠る。

５つのサブコーパスのうち，sunnow と times には日曜紙が含まれている[6]。*Today* は1995年に廃刊になっていたが，2001年３月時点ではサブコーパスの１つとしてまだ残されていた[7]。

表５のコーパスサイズは，当時の Bank of English のサブコーパス一覧画面で表示されていた数字で，語ではないテクストタグなども含まれるため，後に示す総語数よりもいずれも大きい値になっている。コーパスサイズに関しては，*Today* は他のサブコーパスに比べて若干小さいものの，残りの４つはサイズがほぼ同じ大きさとなっており，このことは語彙頻度を比べる際にコーパスサイズの影響が小さくなる良い条件と言える。

発行年については，廃刊になった *Today* 以外は1995-2000年の間となっており，ばらつきはあるものの，ある程度近い時期のものが集められていたと言える。[8]

３．１．２　Bank of English の POS タグセット

Bank of English の POS タグは，その下位セットである WordbanksOnline のオンライン・マニュアル[9] に説明があり，全部で46種類の POS タグが掲載されている。本稿ではこの表に記載がないタグ（$）をつけ加え，表５に挙げた５つの英国全国紙サブコーパスに現れた POS タグのみ（計40）をアルファベット順に並び換えて表６に示す。

表６：Bank of English の英国紙サブコーパスに現れた POS タグ

Tag	Description
BE	verb 'to be' base form: *be*
BED	verb 'to be' past tense: *were*
BEDZ	verb 'to be' 3rd past tense: *was*
BEM	verb 'to be' 1st pers pres sing: *am*
BEN	verb 'to be' past participle: *been*
BER	verb 'to be' 3rd pers pres plural: *are*
BEZ	verb 'to be' 3rd pers, pres sing: *is*
CC	co-ordinating conjunction（*and, or*）
CD	number
CS	subordinating conjunction（*unless, although*）

[6] *The News of the World* は2011年に廃刊になった。
[7] 現在も WordbanksOnline（5,600万語版）には元の約５分の１のコーパスサイズではあるが today がサブコーパスとして残っており、検索できるようになっている。
[8] ただし同じ年の発行であっても発行月は必ずしも一致していない。
[9] http://www.titania.bham.ac.uk/docs/direct_reference.html（アクセス日2012年３月23日）

DEM	demonstrative pronoun: (*this, that*)
DT	determiner
DTG	determiner/pronoun: (*these, those, both, either*)
DTP	possesive determiner: (*my, our*)
EX	existential '*there*'
HV	verb 'to have' base form
HVD	verb 'to have' past tense: *had*
HVZ	verb 'to have' 3rd person pres sing: *has*
IN	preposition (*in, up*)
JJ	adjective
MD	modal verb
NN	common singular noun
NNS	common plural noun
NP	proper noun
PN	general non-personal pronoun (*anyone, everything, none*)
PPL	reflexive pronoun singular: *herself, myself*
PPLS	reflexive pronoun plural: *themselves, yourselves*
PPO	personal pronoun object case: (*me, her*)
PPP	possessive pronoun: (*mine, yours, hers*)
PPS	personal pronoun subject case: (*I, she*)
RB	adverb
TO	'*to*' infinitive marker
UH	formulaic interactive expression: *yes, ugh, um*
VB	verb base form
VBD	verb past tense form
VBG	verb -ING form
VBN	verb past participle form
VBZ	verb 3rd pers pres sing
WH	WH- word
$	possessive *'s*: (*BBC's, Britain's*)

3．2　British National Corpus (BNC)

3．2．1　British National Corpus の英国全国紙ファイル

British National Corpus（BNC）はこれまで（2012年３月現在）に３つのバージョンが公開されているが，本研究では2007年にリリースされた XML Edition を用いた。BNC は個々のファイルの集合体になっていて，あらかじめ特定の種類のファイルを集めた「サブコーパス」を用意するという形式はとっていない。しかし，各ファイルのヘッダー部分にその内容に関する詳細な情報が記されているので，ユーザーはその情報を基に各自の研究の目的に合ったファイルを集めてサブコーパスのように用いる

ことができる。

各ファイルの出典はディスクに収録されている 'BNC User Reference Guide' の中の 'List of Sources' の項（ファイル名 bibliog.html）に記されている。このリストで本稿の研究対象となるファイルを検索したところ，表７に示すように合計356のファイルが該当した。

表７：British National Corpus の英国全国日刊紙のファイル

新聞名	語数の合計	ファイル数	発行年
The Daily Mirror	719,051	6	1992
Today	899,266	10	1992
The Daily Telegraph	1,154,625	93	1992
The Independent	992,594	145	1989
The Guardian	863,192	102	1989

表７は各ファイルのtoken数の合計である。BoE の１紙あたりのコーパスサイズ3,000万前後に比べると BNC は100万語前後とかなり小さく感じられるが，オンライン・アクセスのみの BoE と異なり，BNC はディスク（XML 版は DVD-ROM）に収録された形で入手できるうえ，コーパスを構成するファイルがテキストファイルで提供されていて，その全文をデータとして利用できるので研究上の自由度が大きいという利点がある。

３．２．２　British National Corpus の POS タグセット

表８に BNC XML Edition に適用された，57のタグ[10] からなる C5 と呼ばれる POS タグセットの一覧を示す[11]。

表８：British National Corpus XML Edition の POS タグ

Tag	Description
AJ0	Adjective (general or positive) (e.g. *good, old, beautiful*)
AJC	Comparative adjective (e.g. *better, older*)
AJS	Superlative adjective (e.g. *best, oldest*)
AT0	Article (e.g. *the, a, an, no*)
AV0	General adverb: an adverb not subclassified as AVP or AVQ (see below) (e.g. *often, well, longer* (adv.), *furthest*)
AVP	Adverb particle (e.g. *up, off, out*)
AVQ	Wh-adverb (e.g. *when, where, how, why, wherever*)

[10] この他に punctuation のタグが４つある。
[11] 出典：BNC XML Edition 収録の posguide.html。

CJC	Coordinating conjunction (e.g. *and, or, but*)
CJS	Subordinating conjunction (e.g. *although, when*)
CJT	The subordinating conjunction that
CRD	Cardinal number (e.g. *one, 3, fifty－five, 3609*)
DPS	Possessive determiner-pronoun (e.g. *your, their, his*)
DT0	General determiner-pronoun: i.e. a determiner-pronoun which is not a DTQ or an AT0.
DTQ	Wh-determiner-pronoun (e.g. *which, what, whose, whichever*)
EX0	Existential there, i.e. *there* occurring in the *there is* ... or *there are* ... construction
ITJ	Interjection or other isolate (e.g. *oh, yes, mhm, wow*)
NN0	Common noun, neutral for number (e.g. *aircraft, data, committee*)
NN1	Singular common noun (e.g. *pencil, goose, time, revelation*)
NN2	Plural common noun (e.g. *pencils, geese, times, revelations*)
NP0	Proper noun (e.g. *London, Michael, Mars, IBM*)
ORD	Ordinal numeral (e.g. *first, sixth, 77th, last*).
PNI	Indefinite pronoun (e.g. *none, everything, one* [as pronoun], *nobody*)
PNP	Personal pronoun (e.g. *I, you, them, ours*)
PNQ	Wh-pronoun (e.g. *who, whoever, whom*)
PNX	Reflexive pronoun (e.g. *myself, yourself, itself, ourselves*)
POS	The possessive or genitive marker *'s* or *'*
PRF	The preposition *of*
PRP	Preposition (except for *of*) (e.g. *about, at, in, on, on behalf of, with*)
TO0	Infinitive marker *to*
UNC	Unclassified items which are not appropriately considered as items of the English lexicon.
VBB	The present tense forms of the verb BE, except for *is, 's*: i.e. *am, are, 'm, 're* and *be* [subjunctive or imperative]
VBD	The past tense forms of the verb BE: *was* and *were*
VBG	The -ing form of the verb BE: *being*
VBI	The infinitive form of the verb BE: *be*
VBN	The past participle form of the verb BE: *been*
VBZ	The -s form of the verb BE: *is, 's*
VDB	The finite base form of the verb BE: *do*
VDD	The past tense form of the verb DO: *did*
VDG	The -ing form of the verb DO: *doing*
VDI	The infinitive form of the verb DO: *do*
VDN	The past participle form of the verb DO: *done*
VDZ	The -s form of the verb DO: *does, 's*
VHB	The finite base form of the verb HAVE: *have, 've*
VHD	The past tense form of the verb HAVE: *had, 'd*
VHG	The -ing form of the verb HAVE: *having*
VHI	The infinitive form of the verb HAVE: *have*
VHN	The past participle form of the verb HAVE: *had*
VHZ	The -s form of the verb HAVE: *has, 's*
VM0	Modal auxiliary verb (e.g. *will, would, can, could, 'll, 'd*)

VVB	The finite base form of lexical verbs (e.g. *forget, send, live, return*) [Including the imperative and present subjunctive]
VVD	The past tense form of lexical verbs (e.g. *forgot, sent, lived, returned*)
VVG	The -ing form of lexical verbs (e.g. *forgetting, sending, living, returning*)
VVI	The infinitive form of lexical verbs (e.g. *forget, send, live, return*)
VVN	The past participle form of lexical verbs (e.g. *forgotten, sent, lived, returned*)
VVZ	The -s form of lexical verbs (e.g. *forgets, sends, lives, returns*)
XX0	The negative particle *not* or *n't*
ZZ0	Alphabetical symbols (e.g. *A, a, B, b, c, d*)

4．POS タグの頻度データの作成

本稿で着目しているのは POS タグの頻度であるが，POS タグはコーパスの中で各word unit に付与されている情報であるので，POS タグ付の語彙頻度表を入手し，そこから POS タグのみの頻度表を作成した。本節にその手続きを記述する。

4．1　Bank of English（BoE）の語彙頻度データ

Bank of English で一般ユーザーが利用できる機能はオンラインのアクセスによる検索語を指定したコンコーダンスライン作成や共起語の表示などで，通常はサブコーパスの語彙頻度表を得ることはできない。本稿のデータとして用いた BoE の語彙頻度表は2001年 3 月に英国 Birmingham 大学内で Jeremy Clear 氏の協力により得られたものである。

4．2　British National Corpus（BNC）の語彙頻度データ

British National Corpus は既に述べたように全文にアクセスできるのでユーザーが語彙頻度データを作成することができる。しかし BNC XLM Edition では POS タグは XLM タグの中に記述されており，一般のコンコーダンサーで対応することが難しかったので，本研究のために英国全国紙の各ファイルについて POS タグ付きの語彙頻度データを作成するにあたっては園田勝英氏作成の Python プログラム[12] を利用させていただいた。このプログラムは BNC の各ファイルについて w タグを付与された word classとheadword の頻度表を出力するものである。このプログラムを表 7 の356ファイルに適用し，得られた頻度表を出典の新聞ごとに合計した。

4．3　POS タグに関する問題点と本研究における対処

大規模コーパスはそのサイズの大きさから，人手によってすべての POS タグを付

[12] 特に公開されているわけではなく，2011年 2 月に個人的に使わせていただいたものである。

与することは非現実的であり，BoE，BNC のどちらのコーパスもタグ付けプログラムによって自動的に付与されている。

問題となるのはその精度である。BoE については95％程度以上とされていたようであるが，不定期に更新されることもあって残念ながら詳細な検証は行われていない模様である。

BNC については添付のドキュメント・ファイル（posguide.html[13]）に POS タグに関する詳細な解説がある。表 8 に BNC で用いられている POS タグの一覧を載せたが，表 8 は実は基本となる single tag のリストであり，実際の BNC のファイルには ambiguity tag と呼ばれる，２つの POS が併記されたタグ（例：AJ0-NN1，NP0-NN1，VVN-VVD など）が少なからず付与されている。これはタグ付けプログラムが一方の POS に決定できないときに，より確率の高い POS を第 1 タグ，次の候補となる POS を第 2 タグとして付与するものである。前節で得た BNC の語彙頻度データには計30の ambiguity tag が含まれていた。

上記の添付ドキュメントファイルによれば，BNC の書き言葉テクストから45,000個の POS タグを標本抽出して調べたところ，ambiguity tag の割合は3.83％であったとのことであるが，本研究で得た英国全国紙の頻度データではそれよりも高く，最高は *The Daily Mirror* の4.74％，最低でも *The Guardian* の4.03％であった。（残りの 3 紙の値は *Today* 4.26，*The Daily Telegraph* 4.26，*The Independent* 4.15。いずれも小数点第 2 位で四捨五入。）可能性としては見出しを始めとする新聞特有の表現が標本よりも高い ambiguity tag の割合に関係しているのかもしれない。

この ambiguity tag については，BNC の書き言葉45,000と話し言葉5,000の合わせた計50,000の標本タグについてのかなり詳しい検証結果が添付のドキュメントファイル posguide.html に記されているが，ambiguity tag についてすべて第 1 タグを採用した場合のコーパス全体としての誤付与率（原文では error rate）は書き言葉で2.01％（同ファイル Table 28）と推定されている。

4. 2 節で得られた BNC の 5 つの語彙頻度データには表 8 の57の single tag に加えて，既に述べたように30の ambiguity tag が含まれていた。これをこのまま別個のものとして扱うと87種類となって煩雑すぎることになるし，POS の重複が生じてしまうことも都合が悪い。そこで本稿では便宜上，本稿においては ambiguity tag の頻度は第 1 タグの single tag の頻度と合算して取り扱うこととした。

実は，posguide.html で示されたデータを利用してもう少し細かくタグの頻度の補正を行う方法も考えられる。しかし，上述した誤付与率の推定値から，単純に第 1 タグを採用するという方法でもおよそ98％についてはほぼ正しい POS が付与されてい

[13] このファイルは冒頭に Geoffrey Leech と Nicholas Smith による BNC World Edition の HTML 版マニュアルの改訂版であると記されている。なお，このファイルでは主に word class という用語が用いられているが，本稿では表記の統一上，ここでも POS と表記している。

ると推定できること，30の ambiguity tag の総頻度は本研究の語彙頻度データの５％未満で１つあたりの tag で考えるとさほど大きな割合を占めないことなどの理由から，ambiguity tag について第１タグと看做しても分析結果にさほど大きな影響を及ぼさないと判断した。

５．POS タグ分布

本節では前節までの手続きによって得られた Bank of English と British National Corpus における POS タグ頻度の分布を概観し，階層的クラスター分析（ウォード法）を用いて POS タグの頻度に基づく新聞間の関係を示すとともに，どのような POS タグが各クラスターを特徴づけているかを見る。

５．１　Bank of English の英国全国紙サブコーパスにおける POS タグ分布

BoE の５つの英国全国紙サブコーパスにおける POS タグの出現度数（実頻度）と出現率（％）をそれぞれ表９・表10に示す。

表９：BoE の５つの英国全国紙サブコーパスにおける POS タグの出現度数

Tag	sunnow	today	indy	guard	times	５紙計
BE	160,692	130,261	160,487	175,508	169,358	796,306
BED	80,321	65,943	74,723	81,511	72,989	375,487
BEDZ	278,225	219,890	224,481	233,118	229,207	1,184,921
BEM	47,378	27,365	17,325	17,084	15,668	124,820
BEN	83,200	66,596	77,049	85,747	78,273	390,865
BER	135,527	113,861	140,299	153,484	139,209	682,380
BEZ	360,962	305,523	365,799	394,525	363,364	1,790,173
CC	932,642	733,198	878,491	946,613	889,708	4,380,652
CD	339,263	282,835	276,700	305,212	302,042	1,506,052
CS	507,269	442,500	629,998	672,819	641,059	2,893,645
DEM	76,245	67,083	81,447	86,290	80,579	391,644
DT	2,393,693	2,060,967	2,728,451	2,914,664	2,773,695	12,871,470
DTG	145,813	116,871	168,255	178,795	162,593	772,327
DTP	530,375	426,269	420,815	445,996	437,675	2,261,130
EX	57,458	45,880	65,221	68,742	60,365	297,666
HV	208,699	146,569	153,100	163,023	150,689	822,080
HVD	110,807	80,117	96,590	103,682	98,209	489,405
HVZ	133,617	110,522	119,962	134,536	130,917	629,554
IN	3,216,319	2,715,019	3,409,434	3,644,278	3,462,979	16,448,029
JJ	1,619,822	1,447,773	1,970,410	2,138,512	1,977,374	9,153,891
MD	410,869	314,047	348,833	378,886	366,694	1,819,329

NN	4,219,954	3,785,814	4,539,879	4,920,388	4,676,123	22,142,158
NNS	1,316,967	1,206,254	1,596,731	1,791,455	1,634,021	7,545,428
NP	3,172,826	2,436,008	2,741,026	2,709,744	2,842,442	13,902,046
PN	96,607	74,462	89,818	92,722	81,373	434,982
PPL	19,345	16,278	21,119	21,740	19,571	98,053
PPLS	6,052	3,717	7,018	7,776	6,186	30,749
PPO	423,209	291,740	270,146	285,970	260,185	1,531,250
PPP	2,179	2,174	2,839	1,974	5,479	14,645
PPS	1,247,702	867,816	790,320	815,548	765,688	4,487,074
RB	1,425,526	1,173,795	1,415,165	1,487,655	1,369,289	6,871,430
TO	523,928	416,168	475,566	518,464	488,615	2,422,741
UH	14,563	11,185	14,298	15,953	13,598	69,597
VB	1,200,250	915,428	992,226	1,078,439	1,002,009	5,188,352
VBD	957,474	705,709	631,844	675,736	619,735	3,590,498
VBG	611,870	524,504	619,927	676,381	624,553	3,057,235
VBN	722,302	667,242	769,284	865,695	810,804	3,835,327
VBZ	247,606	230,648	280,398	309,343	291,963	1,359,958
WH	367,115	317,529	396,933	433,356	393,253	1,908,186
$	221,639	188,433	218,727	243,698	214,448	1,086,945
合計	28,404,671	23,565,560	28,062,407	30,031,364	28,507,533	138,571,535

表10：BoE の５つの英国全国紙サブコーパスにおける POS タグの出現率（%）

Tag	sunnow	today	indy	guard	times	総平均
BE	0.57	0.55	0.57	0.58	0.59	0.57
BED	0.28	0.28	0.27	0.27	0.26	0.27
BEDZ	0.98	0.93	0.80	0.78	0.80	0.86
BEM	0.17	0.12	0.06	0.06	0.05	0.09
BEN	0.29	0.28	0.27	0.29	0.27	0.28
BER	0.48	0.48	0.50	0.51	0.49	0.49
BEZ	1.27	1.30	1.30	1.31	1.27	1.29
CC	3.28	3.11	3.13	3.15	3.12	3.16
CD	1.19	1.20	0.99	1.02	1.06	1.09
CS	1.79	1.88	2.24	2.24	2.25	2.09
DEM	0.27	0.28	0.29	0.29	0.28	0.28
DT	8.43	8.75	9.72	9.71	9.73	9.29
DTG	0.51	0.50	0.60	0.60	0.57	0.56
DTP	1.87	1.81	1.50	1.49	1.54	1.63
EX	0.20	0.19	0.23	0.23	0.21	0.21
HV	0.73	0.62	0.55	0.54	0.53	0.59

HVD	0.39	0.34	0.34	0.35	0.34	0.35
HVZ	0.47	0.47	0.43	0.45	0.46	0.45
IN	11.32	11.52	12.15	12.13	12.15	11.87
JJ	5.70	6.14	7.02	7.12	6.94	6.61
MD	1.45	1.33	1.24	1.26	1.29	1.31
NN	14.86	16.07	16.18	16.38	16.40	15.98
NNS	4.64	5.12	5.69	5.97	5.73	5.45
NP	11.17	10.34	9.77	9.02	9.97	10.03
PN	0.34	0.32	0.32	0.31	0.29	0.31
PPL	0.07	0.07	0.08	0.07	0.07	0.07
PPLS	0.02	0.02	0.03	0.03	0.02	0.02
PPO	1.49	1.24	0.96	0.95	0.91	1.11
PPP	0.01	0.01	0.01	0.01	0.02	0.01
PPS	4.39	3.68	2.82	2.72	2.69	3.24
RB	5.02	4.98	5.04	4.95	4.80	4.96
TO	1.84	1.77	1.69	1.73	1.71	1.75
UH	0.05	0.05	0.05	0.05	0.05	0.05
VB	4.23	3.88	3.54	3.59	3.51	3.74
VBD	3.37	2.99	2.25	2.25	2.17	2.59
VBG	2.15	2.23	2.21	2.25	2.19	2.21
VBN	2.54	2.83	2.74	2.88	2.84	2.77
VBZ	0.87	0.98	1.00	1.03	1.02	0.98
WH	1.29	1.35	1.41	1.44	1.38	1.38
$	0.78	0.80	0.78	0.81	0.75	0.78
合計	100.00	100.00	100.00	100.00	100.00	100.00

コーパスサイズに差があるので表10の出現率で比べてみると，直観的な印象としてはいずれのサブコーパスにおいても，各 POS タグの頻度にはさほど大きな差はないように見える。57種類のタグがあるため必然的に個別のタグの頻度自体が小さくなるので，出現率の差をとってもさほど大きな値にはならないからである。

しかし表10のデータを使ってクラスター分析（ユークリッド平方距離, ウォード法）を行ったところ，図１[14] のように５つのサブコーパスの関係が示された。

[14] 図１および表11は Seagull-Stat 2010にて作成。図２・３および表16・19も同様。

図１：POS タグの出現率による BoE のクラスター分析

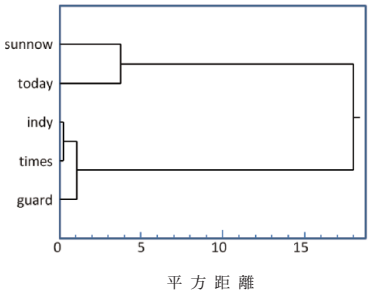

図１から，表10では大きな差がないように見えた BoE の５つの英国全国紙サブコーパスであったが，POS タグの分布で分類すると，大衆紙サブコーパス（sunnow, today）と高級紙サブコーパス（indy, times, guard）とにまず大きく分けられることが示された。さらに細かく見ていくと，高級紙サブコーパスの中で POS タグの分布がもっとも近いのは indy と times で，guard はこの２紙に比べると幾分 POS タグの分布状況が異なっているようである。一方，２つの大衆紙サブコーパス間の距離（＝非類似度）は，高級紙サブコーパス間の距離に比べてかなり大きい。図１の描画の元データである，BoE の５つのサブコーパス間の POS タグ出現率による非類似度行列を表11に示す。

表11：BoE の５つの英国全国紙サブコーパスの POS タグ出現率による非類似度行列

	sunnow	today	indy	guard	times
sunnow	0	3.671	13.844	18.198	14.584
today	3.671	0	4.566	6.687	4.781
indy	13.844	4.566	0	0.723	0.212
guard	18.198	6.687	0.723	0	1.039
times	14.584	4.781	0.212	1.039	0

表11から非類似度がもっとも小さいのは高級紙サブコーパス間で，その値が0.212～1.039であるのに対して，２つの大衆紙サブコーパス間の非類似度は相対的に大きく，3.671と3倍以上であることがわかる。つまりこの大衆紙２紙の間には高級紙３紙の間よりも大きな違いがあるということになる。逆に非類似度がもっとも大きいのは sunnow と guard の間の18.198で，sunnow は高級紙のいずれとも13を超える大きな非類似度を示しており，５つの新聞サブコーパスの中でもっとも異なる POS タグ分布を持っていることがわかる。一方，もう一つの大衆紙である todayは高級紙との非類似度が比較的低く，いずれも一桁に留まっている（4.566～6.687）。高級紙と

大衆紙の間の非類似度は，大衆紙どうしの非類似度の3.671をいずれも上回っていて，大衆紙と高級紙の間の差が大きいことが確認できる。

次に BoE を２つのクラスターに分けた場合に，それぞれのクラスターを特徴づけている POS タグが何かを特定する。ここでは第１クラスターを sunnow と today からなる大衆紙クラスター，第２クラスターを indy, guard, times からなる高級紙クラスターとする。特徴的な POS タグを特定する１つの方法は，各 POS タグのクラスター毎の平均出現率の差をとり，その差の絶対値の大きなものとするやり方である。例えば第１クラスターの平均出現率から第２クラスターの平均出現率を引くと，第１クラスターに多く出現する POS タグはより大きな値を示す。逆に，第２クラスターの平均出現率から第１クラスターの平均出現率を引くと，第２クラスターに多い POS タグがより大きな値を示すことになる。

上記の方法の一つの欠点として考えられるのは，平均出現率の差の大きさだけが判断の尺度となり，その差がもとの出現率に占める相対的な割合について考慮されていないということである。例えば，第１クラスターでの平均出現率が11％，第２クラスターでの平均出現率が10％であるタグＡと第１クラスターでの平均出現率が６％，第２クラスターでの平均出現率が５％であるタグＢがあると仮定した場合，ＡＢどちらのタグについても２つのクラスター間平均出現率の差は１％であるが，10％のうちの１％と，５％の１％ではその割合が異なっており，後者の場合の方がその差が持つ相対的な重要性が高いとする考え方もあるということである。

しかし逆に，平均出現率の差が平均出現率に占める割合のみを考えると，ほとんど出現率のない POS タグが，絶対値としてはわずかな差に過ぎないのに重要性があると過大に評価されてしまう恐れもある。

上記の２点を考慮して，本稿では，平均出現率のデータから各クラスターを特徴づける POS タグを特定する指標として，単純な平均出現率（％）の差の他に，５つの新聞サブコーパス全体における各 POS タグの出現率 p を求め，第１クラスターの平均出現率と第２クラスターの各 POS タグの平均出現率の差を $p(1\text{-}p)$ の平方根で除した値を計算し，この２つの指標による結果を参照することとした[15]。この方法は確立した対処法というわけではなく，本研究の目的に適う補正手段として採用したに過ぎないが，この換算を行ったクラスター間の平均値の差を本稿では便宜上「標準化した（クラスター間）平均差」と呼ぶこととする。

---

[15] POS タグの頻度の（母集団）分布は二項分布と考えられる。二項分布の分散は $np(1\text{-}p)$ であることから，総数 n である語の母集団における出現割合が p であるような POS タグの頻度の標準誤差は $\sqrt{np(1\text{-}p)}$ とおくことができる。表12では n はどの POS タグにも共通の値なので省いても順位自体は変わらない。そこで本稿では単純に $\sqrt{p(1\text{-}p)}$ で除している。この対処法は前田忠彦氏の御教示による。なお，表12では「クラスター平均の差」「全体平均」をどちらも％で示しているが，「標準化した平均差」の計算には本来の値（すなわち表の数値の1/100）を用いた。表13・17・18についても同様。

結論を言えば，以下に示すように上の２つの方法による結果は，若干の順位の変動は見られたものの，それぞれのクラスターをもっとも特徴づけている POS タグの上位の組み合わせにはそれほど大きな違いは生じなかった。

表12：BoE の大衆紙クラスターを特徴づける POS タグ

Tag	第１クラスターの平均（%）	第２クラスターの平均（%）	全体平均（%）	第１クラスター－第２クラスター	左欄の順位	標準化した平均差	左欄の順位
PPS	4.006	2.718	3.213	1.288	1	0.073	1
NP	10.669	9.513	9.954	1.156	2	0.039	4
VBD	3.158	2.208	2.571	0.950	3	0.060	2
VB	4.023	3.520	3.715	0.504	4	0.027	6
PPO	1.353	0.935	1.096	0.418	5	0.040	3
BEM	0.140	0.057	0.089	0.083	12	0.028	5

BoE の大衆紙クラスターを特徴づける POS タグとして，表12ではそれぞれの計算結果の上位５位までを挙げた。実際には「第１クラスター－第２クラスター」（＝第１クラスター平均と第２クラスター平均の差）の５位の PPO と12位の BEM の間に DTP，CD，BEDZ，HV，MD，TO の６つの POS タグがランクされている。なお順位は異なるものの，12位までの POS タグは２つの方法で一致した。

表12から，BoE の大衆紙クラスターを特徴づける主な POS タグは，クラスター間平均差の絶対値によれば主格人称代名詞，固有名詞，動詞の過去形，動詞の原形，目的格人称代名詞などであり，出現率を考慮した平均差を考えた場合は主格人称代名詞，動詞の過去形，目的格人称代名詞，固有名詞，１人称 be 動詞（*am*）という結果であった。

次に BoE の高級紙クラスターを特徴づける POS タグを表13に示す。

表13：BoE の高級紙クラスターを特徴づける POS タグ

Tag	第１クラスターの平均（%）	第２クラスターの平均（%）	全体平均（%）	第２クラスター－第１クラスター	左欄の順位	標準化した平均差	左欄の順位
DT	8.519	9.644	9.216	1.125	1	0.039	3
JJ	5.877	6.972	6.554	1.095	2	0.044	1
NNS	4.839	5.751	5.403	0.911	3	0.040	2
NN	15.340	16.195	15.855	0.856	4	0.023	5
IN	11.333	12.050	11.777	0.717	5	0.022	6
CS	1.817	2.227	2.072	0.410	6	0.029	4

表13では２つの方法で６位までの POS タグの組み合わせが一致しており，大衆紙クラスターを特徴づける POS タグよりも順位の変動幅は小さい。順位に若干の差はあるものの，どちらの方法でも上位３位までが決定詞，形容詞，名詞の複数形の組み合わせであり，以下に名詞の単数形，前置詞，従位接続詞という組み合わせが続いている。

5．2　British National Corpus の英国全国紙ファイルにおける POS タグ分布

BNC XML Edition の英国全国紙５紙のファイルにおける POS タグの出現度数（実頻度）と出現率（％）をそれぞれ表14・表15に示す。表中，各紙の名称は適宜略記している。

表14：BNC の英国全国紙ファイルにおける POS タグの出現度数

Tag	Mirror	Today	Telegraph	Independent	Guardian	５紙計
AJ0	44,282	77,069	85,973	75,578	64,865	347,767
AJC	830	1,821	2,352	1,926	1,747	8,676
AJS	1,240	2,145	1,949	1,389	1,155	7,878
AT0	58,896	102,641	107,763	96,303	82,894	448,497
AV0	28,257	50,039	47,296	40,126	34,340	200,058
AVP	8,560	13,435	8,936	6,817	6,274	44,022
AVQ	1,631	2,818	2,277	1,898	1,578	10,202
CJC	20,897	36,977	35,307	29,447	26,034	148,662
CJS	9,892	17,270	15,807	12,926	11,220	67,115
CJT	3,221	5,955	7,629	7,693	6,527	31,025
CRD	15,615	27,634	25,558	18,290	17,305	104,402
DPS	13,385	22,148	16,794	12,719	10,940	75,986
DT0	11,147	20,755	21,298	18,711	15,799	87,710
DTQ	2,355	4,627	5,860	5,727	4,583	23,152
EX0	1,184	2,237	2,418	2,086	1,717	9,642
ITJ	422	421	365	254	302	1,764
NN0	5,126	10,274	8,265	6,842	5,947	36,454
NN1	118,230	199,624	191,050	168,530	144,582	822,016
NN2	35,617	61,833	65,090	57,293	51,144	270,977
NP0	67,806	112,464	92,170	75,553	69,516	417,509
ORD	4,676	8,446	7,398	5,607	5,037	31,164
PNI	2,004	3,505	2,751	2,160	1,971	12,391
PNP	36,082	57,186	36,366	27,592	23,991	181,217
PNQ	2,684	4,469	3,951	2,978	2,707	16,789
PNX	628	1,078	1,064	886	759	4,415
POS	7,020	11,592	10,403	9,250	7,908	46,173
PRF	13,955	26,330	33,195	31,113	26,074	130,667
PRP	61,551	107,842	105,545	90,727	77,719	443,384
TO0	11,921	21,437	18,627	17,089	14,844	83,918
UNC	1,473	2,917	3,802	4,340	3,563	16,095
VBB	4,227	7,380	5,921	5,005	4,019	26,552
VBD	9,693	15,499	12,013	10,053	8,965	56,223

VBG	791	1,294	995	922	829	4,831
VBI	3,507	6,771	6,428	6,261	5,382	28,349
VBN	2,016	3,478	3,292	2,781	2,477	14,044
VBZ	8,420	15,128	13,406	11,666	9,216	57,836
VDB	891	1,440	911	734	651	4,627
VDD	835	1,241	857	723	654	4,310
VDG	138	295	166	127	105	831
VDI	489	819	555	463	358	2,684
VDN	217	368	277	210	175	1,247
VDZ	368	602	580	513	411	2,474
VHB	2,771	4,855	3,770	2,906	2,506	16,808
VHD	2,457	4,183	3,602	3,163	2,776	16,181
VHG	235	401	446	320	239	1,641
VHI	1,442	2,632	2,236	1,870	1,575	9,755
VHN	263	479	276	232	174	1,424
VHZ	3,192	5,891	5,292	4,554	3,755	22,684
VM0	9,091	16,828	14,253	12,712	11,390	64,274
VVB	9,652	14,713	11,833	9,285	8,121	53,604
VVD	23,822	36,207	25,215	20,082	18,699	124,025
VVG	11,464	19,619	17,169	14,593	12,759	75,604
VVI	17,572	31,256	25,632	22,864	20,404	117,728
VVN	16,882	28,326	26,296	24,135	20,706	116,345
VVZ	6,763	11,760	11,191	9,539	8,215	47,468
XX0	4,778	7,899	5,919	5,553	4,716	28,865
ZZ0	503	821	827	402	550	3,103
合計	733,066	1,257,174	1,166,617	1,003,518	872,869	5,033,244

表15：BNC の英国全国紙ファイルにおける POS タグの出現率（%）

Tag	Mirror	Today	Telegraph	Independent	Guardian	総平均
AJ0	6.04	6.13	7.37	7.53	7.43	6.91
AJC	0.11	0.14	0.20	0.19	0.20	0.17
AJS	0.17	0.17	0.17	0.14	0.13	0.16
AT0	8.03	8.16	9.24	9.60	9.50	8.91
AV0	3.85	3.98	4.05	4.00	3.93	3.97
AVP	1.17	1.07	0.77	0.68	0.72	0.87
AVQ	0.22	0.22	0.20	0.19	0.18	0.20
CJC	2.85	2.94	3.03	2.93	2.98	2.95
CJS	1.35	1.37	1.35	1.29	1.29	1.33
CJT	0.44	0.47	0.65	0.77	0.75	0.62
CRD	2.13	2.20	2.19	1.82	1.98	2.07
DPS	1.83	1.76	1.44	1.27	1.25	1.51
DT0	1.52	1.65	1.83	1.86	1.81	1.74
DTQ	0.32	0.37	0.50	0.57	0.53	0.46
EX0	0.16	0.18	0.21	0.21	0.20	0.19
ITJ	0.06	0.03	0.03	0.03	0.03	0.04
NN0	0.70	0.82	0.71	0.68	0.68	0.72
NN1	16.13	15.88	16.38	16.79	16.56	16.33
NN2	4.86	4.92	5.58	5.71	5.86	5.38
NP0	9.25	8.95	7.90	7.53	7.96	8.30
ORD	0.64	0.67	0.63	0.56	0.58	0.62
PNI	0.27	0.28	0.24	0.22	0.23	0.25
PNP	4.92	4.55	3.12	2.75	2.75	3.60
PNQ	0.37	0.36	0.34	0.30	0.31	0.33
PNX	0.09	0.09	0.09	0.09	0.09	0.09
POS	0.96	0.92	0.89	0.92	0.91	0.92
PRF	1.90	2.09	2.85	3.10	2.99	2.60
PRP	8.40	8.58	9.05	9.04	8.90	8.81
TO0	1.63	1.71	1.60	1.70	1.70	1.67
UNC	0.20	0.23	0.33	0.43	0.41	0.32
VBB	0.58	0.59	0.51	0.50	0.46	0.53
VBD	1.32	1.23	1.03	1.00	1.03	1.12
VBG	0.11	0.10	0.09	0.09	0.09	0.10
VBI	0.48	0.54	0.55	0.62	0.62	0.56
VBN	0.28	0.28	0.28	0.28	0.28	0.28
VBZ	1.15	1.20	1.15	1.16	1.06	1.15
VDB	0.12	0.11	0.08	0.07	0.07	0.09

VDD	0.11	0.10	0.07	0.07	0.07	0.09
VDG	0.02	0.02	0.01	0.01	0.01	0.02
VDI	0.07	0.07	0.05	0.05	0.04	0.05
VDN	0.03	0.03	0.02	0.02	0.02	0.02
VDZ	0.05	0.05	0.05	0.05	0.05	0.05
VHB	0.38	0.39	0.32	0.29	0.29	0.33
VHD	0.34	0.33	0.31	0.32	0.32	0.32
VHG	0.03	0.03	0.04	0.03	0.03	0.03
VHI	0.20	0.21	0.19	0.19	0.18	0.19
VHN	0.04	0.04	0.02	0.02	0.02	0.03
VHZ	0.44	0.47	0.45	0.45	0.43	0.45
VM0	1.24	1.34	1.22	1.27	1.30	1.28
VVB	1.32	1.17	1.01	0.93	0.93	1.06
VVD	3.25	2.88	2.16	2.00	2.14	2.46
VVG	1.56	1.56	1.47	1.45	1.46	1.50
VVI	2.40	2.49	2.20	2.28	2.34	2.34
VVN	2.30	2.25	2.25	2.41	2.37	2.31
VVZ	0.92	0.94	0.96	0.95	0.94	0.94
XX0	0.65	0.63	0.51	0.55	0.54	0.57
ZZ0	0.07	0.07	0.07	0.04	0.06	0.06
合計	100.00	100.00	100.00	100.00	100.00	100.00

BoE の場合と同様に，表15を見る限りでは各 POS タグの出現率に若干の差はあるものの，全体としての分布状況にはさほど相違が無い様に思われる。しかし表15のデータに階層的クラスター分析（平方距離，ウォード法）を適用すると，やはり BoE の場合と同様に，図２に示すように大衆紙と高級紙という２つのクラスターに明確に分けられた。

図２：POS タグの出現率による BNC のクラスター分析

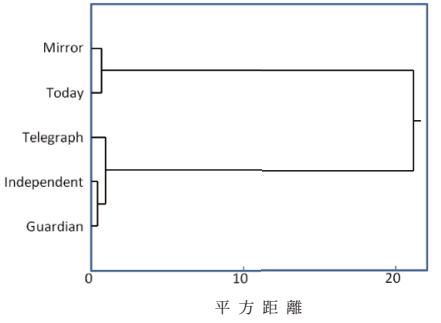

図２は，５つの英国全国紙が，まず大衆紙クラスター（*Daily Mirror*，*Today*）と高級紙クラスター（*The Daily Telegraph*，*The Independent*，*The Guardian*）とに大きく２つに分かれる点で図１と似ているが，各新聞間の距離（非類似度）の関係はやや異なる様相を見せている。図１でかなり離れていた大衆紙クラスター内の距離は，図２の *Daily Mirror* と *Today* の間では小さくなり，高級紙クラスター内の距離とほぼ同じ程度になっている。高級紙クラスター内を見てみると，図１においては高級紙３紙の中ではわずかな差ながら一番離れた距離にあった *The Guardian* が，図２においては *The Independent* と一番類似性が高く，３つの高級紙の中でもっとも離れているのは *The Daily Telegraph* になっている。図２の描画の元データである，BNC の５つの英国全国紙ファイルの POS タグの出現率による非類似度行列を表16に示す。

表16：BNC の５つの英国全国紙ファイルの POS タグ出現率による非類似度行列

	Mirror	Today	Telegraph	Independent	Guardian
Mirror	0	0.676	12.227	18.314	15.498
Today	0.676	0	8.348	13.755	11.323
Telegraph	12.227	8.348	0	1.002	0.568
Independent	18.314	13.755	1.002	0	0.399
Guardian	15.498	11.323	0.568	0.399	0

表16でもっとも非類似度がもっとも小さいのは *The Independent* と *The Guardian* の0.399であり，次に *The Daily Telegraph* と *The Guardian* の0.568が続く。しかし，BoE の結果とは異なり，BNC では２つの大衆紙，つまり *Daily Mirror* と *Today* の非類似度が0.676と小さく，この値は *The Daily Telegraph* と *The Independent* の1.002を下回っている。つまり BNC では大衆紙どうし，高級紙どうしの非類似度がほぼ同程度になっているということになる。一方，表16で非類似度がもっとも大きいのは *Daily Mirror* と *The Independent* の18.314で，*Daily Mirror* は他の高級紙２紙とも12を超える非類似度を示している。もう一つの大衆紙 *Today* も高級紙３紙といずれも比較的大きな非類似度を示しているが，*The Daily Telegraph* だけは非類似度が一桁に留まっており，POS タグの出現率において，*The Daily Telegraph* が高級紙の中で大衆紙にもっとも近いことが読み取れる。

次に図２で明確に分かれた２つのクラスターについて，第１クラスターを *Daily Mirror* と *Today* からなる大衆紙クラスター，第２クラスターを *The Daily Telegraph, The Independent, The Guardian* からなる高級紙クラスターとして，それぞれのクラスターを特徴づける POS を見る。BoE の表12・13と同様に，それぞれのクラスター内での出現率平均の差のほかに，本稿の方法で「標準化」した平均差についても調べる。

まず，大衆紙クラスターを特徴づける POS タグを表17に示す。

表17：BNC の大衆紙クラスターを特徴づける POS タグ

Tag	第１クラスターの平均（%）	第２クラスターの平均（%）	全体平均（%）	第１クラスター－第２クラスター	左欄の順位	標準化した平均差	左欄の順位
PNP	4.735	2.872	3.603	1.864	1	0.100	1
NP0	9.098	7.798	8.300	1.300	2	0.047	3
VVD	3.065	2.102	2.466	0.963	3	0.062	2
DPS	1.794	1.320	1.511	0.474	4	0.039	5
AVP	1.118	0.721	0.875	0.397	5	0.043	4

若干の順位の変動はあるものの，２つの方法で算出した指標で上位５までの組み合わせは一致しており，BNC の大衆紙クラスターをもっとも特徴づけている POS タグは人称代名詞（主格，目的格，所有格），固有名詞，動詞の過去形，所有代名詞，副詞辞という結果になった。

次に，高級紙クラスターを特徴づける POS タグを表18に示す。

表18：BNC の高級紙クラスターを特徴づける POS タグ

Tag	第１クラスターの平均（%）	第２クラスターの平均（%）	全体平均（%）	第２クラスター－第１クラスター	左欄の順位	標準化した平均差	左欄の順位
AJ0	6.085	7.444	6.914	1.358	1	0.054	2
AT0	8.099	9.443	8.916	1.344	2	0.047	3
PRF	1.999	2.978	2.598	0.979	3	0.062	1
NN2	4.889	5.716	5.387	0.827	4	0.037	4
NN1	16.003	16.578	16.342	0.575	5	0.016	9
CJT	0.457	0.723	0.617	0.266	7	0.034	5

高級紙クラスターを特徴づける POS タグは４位までの組み合わせが２つの指標で一致し，形容詞，冠詞，前置詞，名詞の複数形という結果であった。５位についてはやや大きな順位の差があり，クラスター平均の差でみると名詞の単数形，標準化した平均差でみると従位接続詞 *that* という結果であった。

ところで，4. 3節で述べたように本研究では ambiguity tag についてすべて第1タグであるものとして頻度の集計を行った。実際には第１タグが正しい割合はほぼすべての ambiguity tag について50～90％の範囲にあり，本稿の仮定は全体の割合から考えて結果に大きな違いを生じない選択であると考えられるものの，第１タグの割合を水増ししていることは間違いない。そこで，その影響の有無を検証するために，今度は第１タグと第２タグの正しい割合が半々であると仮定した場合のデータを作成し，同様にクラスター分析を行った。データの詳細は本稿では割愛するが，図３に分析結果の樹形図を，表19に非類似度行列表を示す。結果を比較すると，図２と図３はほぼ同じである。第１タグが正しい割合は実際にはこの２つの仮定データの間にあるはずなので，図２と図３から，ambiguity tag の扱いは本稿で行った分析方法ではそ

の結果に及ぼす影響は非常に小さいことが確認できた。

図３：第１・第２タグを各50％と仮定したデータによる BNC のクラスター分析

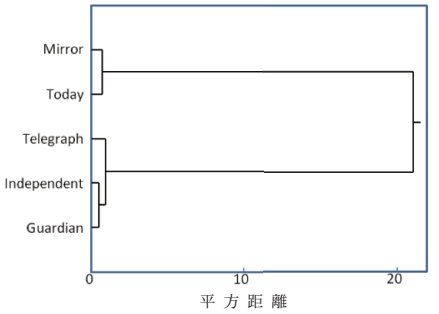

表19：第１・第２タグを各50％と仮定したデータによる BNC の非類似度行列

	Mirror	Today	Telegraph	Independent	Guardian
Mirror	0	0.671	12.097	18.247	15.268
Today	0.671	0	8.364	13.870	11.218
Telegraph	12.097	8.364	0	1.000	0.558
Independent	18.247	13.870	1.000	0	0.483
Guardian	15.268	11.218	0.558	0.483	0

６．考察

本節では前節で得られた BoE と BNC の結果を総合して考察を行う。

まず，POS タグの出現率自体の数字はさほど大きく異なっているわけではないが，クラスター分析を行うと，どちらのコーパスにおいても大衆紙２紙と高級紙３紙の間で明確に区分された。２つのコーパスに含まれている新聞に若干の違いがあり，用いられた POS タグセットは数や分類がかなり異なっているが，この点で一致する結果が出たことは，大衆紙と高級紙で使用される品詞の割合に差があることを示していると言えるだろう。「新聞」という同じジャンルに属している大衆紙と高級紙であるが，一般に認識されているこの２種類の新聞の文体の差の一端は，用いられる品詞の差という抽象的なレベルにおいても表れるということになる。

各クラスターについてみると，BoE の分析では高級紙の *The Independent, The Times, The Guardian* の３紙が非常に近いという結果であった。一方，BNC の分析では，*The Independent* と *The Guardian* の２紙は非常に近い関係にあるが，*The Daily Telegraph* はそれに比べると若干離れた関係にあることが示された。

大衆紙については，BNC の結果では *Daily Mirror* と *Today* は非常に近い関係にあるのに比べ，BoE の結果では *The Sun*（＋*The News of the World*）と *Today* はかなり離れている関係にあることが示された。

BoE と BNC は適用されているタグセット自体も，また現われたタグセットの数も異なるのでその点を忘れてはならないが，ともに POS タグの出現率（百分率）をデータとして計算した非類似度行列で単位は揃っているので，仮に表11と表16の表を合わせて7紙の関係を大まかに捉えてみると，例えば，もっとも離れているのは *The Sun*（＋*The News of the World*）で，やや離れて *Daily Mirror*，もう少し離れたところに *Today* があり，ここに大きな距離があって *The Daily Telegraph* があり，やや離れたところに *The Independent*，*The Times* が位置し，この二紙から若干離れたところに *The Guardian* がある，とまとめることができる[16]。図４に，BoE と BNC の二つの非類似度行列を総合して考えた英国全国紙７紙の位置関係を仮に直線的に解釈した場合例を示してみる。紙面の都合上，紙名は適宜略記している。

図４：BoE と BNC の非類似度行列を総合した英国紙７紙の位置関係の例

Sun　Mirror　Today　　Telegraph　　Independent　Times　Guardian

図４に示したように，４つの高級紙の中では，POS タグの出現率という点で *The Daily Telegraph* は他の３紙とやや異なることが示唆されている。大衆紙との非類似度も比較的小さく，このことは高級紙の中で *The Daily Telegraph* がもっとも発行部数が多いことと関係があるかもしれない。例えば，他の３紙に比べて親しみやすい内容を多く扱ったり，読みやすい文章になっているという特徴があり，それが POS タグの出現率に表れている可能性が考えられる。具体的にどのような文体上の特徴が POS タグの出現率と結びついているのかについてはより詳細にコーパスを調査することが必要である。

３つの大衆紙については，BoE の結果では mid-market に分類される *Today* と down-market に分類される *The Sun* の間にはかなり大きな違いがあるが，BNC の結果では，*Today* とやはり down-market に分類される *Daily Mirror* の間の差は非常に小さかった。このことは講読者の社会階層比率に基づく大衆紙の区分と POS タグの出現率から見た大衆紙の区分に若干の違いがある可能性を示している。

それぞれのクラスターを特徴づける POS タグについてまとめてみると，大衆紙に

[16] 表11では today は guard との距離（＝非類似度）よりも indy との距離が近いという結果になっているが，表16では逆に Today と Independent の非類似度の方が，Today と Guardian よりも離れていて，ここに矛盾が生じる。このように２つの表には明らかな不整合があるため７紙の相対的な遠近関係を矛盾なく直線上に並べることはできない。本節の記述および図４はあくまでも仮に表11の結果を優先させた解釈のイメージ図に過ぎないことに注意されたい。

ついては，BoE では人称代名詞（主格，目的格）・一般動詞の過去形と原形・固有名詞・1 人称 be 動詞（*am*）が，BNC の結果では人称代名詞（主格，目的格，所有格）・所有代名詞・固有名詞・一般動詞の過去形・副詞辞が，高級紙に比べて出現率の高い POS タグと特定された。人称代名詞（主格，目的格）・一般動詞の過去形・固有名詞については 2 つのコーパスで一致した結果となり，高い類似性が見られた。このことは，人物に関する記述が多く，また狭い紙面[17] で最低限の情報を伝えるために主語と動詞という基本的な文の組立てが中心になる大衆紙の特徴と符合していると考えられる。また，副詞辞については大衆紙で句動詞が用いられることが多いためであることが，1 人称 be 動詞（*am*）が多いのは，取材対象である人物の発言を直接話法で表記することが多いためであることが推測できる。

一方，高級紙を特徴づける POS タグとしては，BoE の結果では決定詞・形容詞・名詞（単数形，複数形）・前置詞・従位接続詞が，BNC の結果では形容詞・冠詞・前置詞 *of*・名詞（単数形，複数形），従位接続詞 *that* が特定された。BoE の決定詞の大部分は冠詞である。高級紙を特徴づける POS タグは 2 つのコーパスにおいて冠詞・形容詞・名詞（単数形，複数形），前置詞，従位接続詞という結果でほぼ一致した結果となり，やはり高い類似性が見られた。冠詞・形容詞・前置詞・名詞はいずれも名詞句の構成要素になる品詞であり，平均文長が長く，構文的にも大衆紙より複雑である高級紙の特徴と符合する。同様に従位接続詞も文がより長くなる品詞であるが，従位接続詞 *that* に関しては，以前，新聞でもっとも頻度の高い一般動詞の語形である *said* の構文について *Daily Mirror* と *The Times* のテクストを調べた際（高見 1996:102）に，*Daily Mirror* では直接話法を用いることが多く，間接話法をとる場合も *that* 節はほとんど用いられない（1％程度）が，*The Times* では16％程度で *that* 節をとっていた結果が出たこととも符合する。

ここで本稿の表 1 ～表 4 で示したジャンルによる BNC の POS タグの分布（Leech et al. 2001）と比較してみたい。まず表1・2の書き言葉と話し言葉との比較では，書き言葉で名詞・形容詞・冠詞が多い点は高級紙の特徴と，話し言葉で代名詞が多い点は大衆紙の特徴との対応が見られる。高級紙も大衆紙も当然ながら BNC で「書き言葉」の範疇に分類されているが，1 節で触れたように大衆紙は “colloquial and informal”（Jucker 1992:7）であるとされているとおり，高級紙との比較ではどちらかと言えば話し言葉に近い POS タグ分布になっていることがわかる。しかし，書き言葉でもっとも高い対数尤度比を示した固有名詞は，本稿の結果では高級紙ではなく大衆紙を特徴づける POS タグになっており，これは書き言葉である大衆紙が話し言

[17] 本稿で使用したコーパスはいずれも高級紙が小型化する前のデータであるが，高級紙が小型化した現在，その影響が言語にも表れている可能性がある。その可能性を示す一つの例として，2004年 2 月27日付の The Times は，同じ日の broadsheet 版と compact 版とで紙幅に合わせて見出しに若干の違いが施されていたことが挙げられる（高見 1996:40）。

葉と大きく異なる点の１つであることがわかる。

表３・４の informative writing と imaginative writing との比較では，どちらも writing である点で共通性が高いためか，書き言葉と話し言葉以上に高級紙と大衆紙との対応が見られる。informative writing に多い名詞の複数形・形容詞・前置詞 *of* は高級紙の特徴と，imaginative writing に多い人称代名詞（主格，所有格）・動詞の過去形は大衆紙の特徴と一致する。高級紙は informative writing に近く，大衆紙はimaginative writing に近いということになるのは興味深い。大衆紙は現実社会について報じるもので，もちろん imaginative writing ではないのであるが，娯楽性が高いという点で imaginative writing に近い特徴を持っているようである[18]。

８．おわりに

本稿ではイギリスの大規模コーパスである Bank of English と British National Corpus に含まれているそれぞれ５つのイギリスの全国日刊一般紙について，POS タグの分布からクラスター分析による分類と，クラスター間・クラスター内での比較を行った。新聞の組み合わせや収録時期，使用されたタグセットに違いがあっても，クラスター分析の結果，２つのコーパスで大衆紙クラスターと高級紙クラスターとに明確に分類され，各クラスターに特徴的な POS タグを特定することができた。またクラスター内においても，クラスター分析結果の非類似度行列から各新聞のおよその相対位置をとらえることができた。また，書き言葉・話し言葉および informative writing・imaginative writing のジャンルで多い POS タグと高級紙・大衆紙クラスターを特徴づける POS タグにいくつかの対応が見られることがわかった。

POS タグ付きの大規模コーパスや POS タグを付与するタグ付けプログラム（POS タガー）の普及にともない，かつては実施が困難だった POS 出現率の研究が増えて来ている。一方，POS 出現率の研究成果の有用性については現在のところ広く認識されるまでには至っていないが，今後こうしたジャンルと POS 出現率の関係に関する研究成果を積み重ねることによって，文体の違いを構成する要因を明らかにするデータの一部になりうるものと考える。

本稿では個別の POS タグの出現率のみを取り上げ，品詞ごとの集計については割愛した。というのも，同じ品詞に属する POS タグでも，特に動詞はその語形によってその振る舞いが異なっていたからである。一般動詞の過去形と原形は大衆紙を特徴づける POS タグであったが，不定形 *be* や一般動詞の過去分詞形，三人称単数現在などは相対的には高級紙の出現率の方が高かった。同じ動詞でも語形によって文体との

---

[18] おそらく Leech et al.（2001）の分析自体において BNC の高級紙・大衆紙の両方が，書き言葉と informative writing のデータとして用いられたものと考えられるので，本節の記述は自己包含的な比較になっていると思われる。書き言葉や informative writing と共通する結果はその影響が考えられるが，その逆の結果になっている話し言葉や imaginative writing と共通する大衆紙の結果は注目に値すると言えるだろう。

関連があるらしいことも興味深い結果である。

文体の違いを詳細に検討するには，やはり個々の語について調べることが必要になってくる。Takami（2004）で形容詞について調べたが，本稿の結果を参照しながら今後他の品詞についても調査し，大衆紙と高級紙の文体の違いが具体的にどのような点にあるのかについてさらに研究を進めていきたい。

謝辞

本稿は平成23年度統計数理研究所公募型共同利用研究・一般研究２「イギリスの巨大コーパスにおける新聞サブコーパスの統計学的比較研究」（課題番号23-共研-2026）の成果の一部であり，同研究所で2012年３月７- ８日に開催された合同研究発表会『言語研究と統計2012』で行った口頭発表をもとに，大幅な修正を加えたものである。統計数理研究所と数年来，研究の遂行に有益な助言をいただいている同研究所の前田忠彦氏，『言語研究と統計』のすべての関係者に心より感謝申し上げる。また，本研究の実施にあたって極めて重要なオリジナルデータの入手や作成に大きな御協力をいただいた，COBUILD スタッフ（2001年当時）の Jeremy Clear 氏と，北海道大学の園田勝英氏に厚く御礼申し上げる。両氏の御協力がなければ本研究はなりたたなかった。なお言うまでもなく本稿に関する誤りがあればすべて筆者の責によるものである。（論文発表後に訂正箇所が判明した場合は http://www.hucc.hokudai.ac.jp/ ~p16537/index3.html で公表するので確認されたい。）

参考文献

Crystal, D. and D. Davy.（1969）*Investigating English Style*. London: Longman.

Jucker, A.（1992）*Social Stylistics: Syntactic Variation in British Newspapers*. Berlin/New York: Mouton de Gruyter.

Leech, J., P. Rayson, & A. Wilson.（2001）*Word Frequencies in Written and Spoken English*. Harlow, UK: Pearson Education.

Nakamura, J., N. Inoue & T. Tabata.（2004）*English Corpora under Japanese Eyes: JAECS Anthology Commemorating its 10th Anniversary*. Amsterdam: Rodopi.

Takami, S.（2004）“A Corpus-Driven Identification of Distinctive Words: ‘Tabloid Adjectives’ and ‘Broadsheet Adjectives’ in the Bank of English,” in Nakamura, Inoue & Tabata（eds.）, 115-35.

高見敏子（1996）「イギリスの高級紙と大衆紙―語彙の基礎的観察」（研究ノート）『英語コーパス研究』3, 95-104.

高見敏子（2003）「『高級紙』と『大衆紙』のcorpus-drivenな特定法」『北海道大学大学院国際広報メディア研究科・言語文化部紀要』44, 73-105.［http://www.hucc.hokudai.ac.jp/~p16537/index5.html から入手可。http://www.hucc.hokudai.ac.jp/~p16537/index3.htmlに掲載している補足と訂正についても併せて参照されたい。］

高見敏子（2005）「変わり行く英国の新聞」簗田憲之・橋本尚江（編著）『言語文化部公開講座　変わり行く英国』北海道大学言語文化部研究報告叢書58, 27-50.

使用統計パッケージ

Excelアドイン工房「Seagull-Stat 2010」(http://www.jomon.ne.jp/~hayakari/index.html)