知 能 と 複 雑 系 116－4
（1999．5．24）

日本語新聞記事の末尾における名詞語彙

中村 隆志　小泉 明日美　本間 愛

{tks@human, asumi@info, a-honma@dolphin}.ge.niigata-u.ac.jp

新潟大学人文学部

新聞記事の文章末に着目し、文書の文字数が大きくなるほど、文章末に出現する頻度が高くなるような名詞の意味属性値を統計的に計量した。前回までの報告[1,2]において、文書数が 500 から 1500 までの文書群 495 個について、大河ターム（ある文書群に於いて、文書の文字数が増加するに連れ、文章末での出現頻度が増加する名詞を表す造語）の検出を行った。480 個(～ 97.0%)の文書群からのべ 4158 個の大河タームが取り出された。本報告では、各文書群の頻出名詞（TF{8}）、あるいは文章末２文だけから集計した頻出名詞（PF{8}）を比較対象として、得られた大河タームのシソーラス[7]上の意味属性値の計量を行った。その結果、「精神」「知的生産物（思考、学習）」「非暦日」「界」「様相」の５つの中間節点の下位範疇に属する名詞の使用頻度が大河タームに優位に見られ、各文書群との頻出名詞との間に明瞭な頻度の差が確認された。

キーワード：新聞記事、文章末表現、シソーラス

Specific noun usage in closing sentences in Japanese newspaper articles

Takashi Nakamura　Asumi Koizumi　Ai Homma

{tks@human, asumi@info, a-honma@dolphin}.ge.niigata-u.ac.jp

Department of Humanities, Niigata University

We focus on the closing sentences of newspaper articles and measure the frequency of semantic feature values of nouns whose usage frequency in the closing sentences increases as the article length increases. In the past report[1,2], data were compiled from 495 groups of newspaper articles which include between 500 and 1500 articles. A total of 4158 "taiga terms", nouns whose usage frequency in the closing sentences increases as the article length increases in a group of newspaper articles, were obtained from 480 groups(～97.0%). In this report, we measure the frequency of the semantic feature values of "taiga terms" in a thesaurus[7], compared to the most frequent nouns in the groups of articles(TF{8}), and to the most frequent nouns in only the last two sentences of each article in each group(PF{8}). The usage frequency of "taiga terms" is larger than both that of TF{8} and PF{8} for nouns which belong to subcategory of five nodes, 'seishin (spirit)', 'chiteki-seisanbutsu (intelligent products)', 'hi-rekibi (temporal expression except for calendar days)', 'kai (industry)', and 'yousou (phase)'.

keywords:Newspaper articles, closing sentences, thesaurus

1：はじめに

文章末の文には、後続する文がない。文章中、他の全ての文は後続する文を持ち、それらとの意味的結束を持つ可能性がある。しかし、文章末の文には、先行する文との結束しかない。筆者ら[1,2]は、この特殊性を明示的にするために、大河ターム（３章にて後述）という概念と抽出法を提案してきた。本稿は、その延長上にあるものであり、得られた大河タームの意味属性について考察する。

近年の電子化テキストの爆発的増加に伴い、自動要約の技術開発が急ピッチで進行しつつある。それらはいくつかの方法論に分類可能である[3]。抽出派と呼ばれる重要文抽出法において、文章末の扱い方は議論が分かれているが、現行の研究の多くは、文章末の文を重要視しないものとなっている[例えば 4]。この方法を定量的に支持する指摘[5]もあり、この方法は妥当であろう。

文章という構造をマクロに見れば、タイトルや冒頭文も特殊な位置づけにあるが、それと同様に、文章末の文も特殊な位置にある。タイトルや冒頭文では他の文との意味的結束を持つのが後方に、文章末の文では前方に限定されている。タイトルや冒頭文の重要性は疑いのないところであるが、同様に文章末の文も、文章内においても重要な文である可能性はあろう。[6]においては、逆に冒頭文と同様、文章末の文（[6]では最終文と表記）も筆者の主張が込められている文であるとして、重要文であるとしている。しかし、ここではその重要性をアプリオリに前提としており、その重要度は吟味されていない。

筆者ら[1,2]は、文章末の文での語彙使用の特徴を取り出すため、大河タームと云う概念を提示した。その狙いを述べよう。文章末の文も、ある一つの文でしかない。総文字数が 150 前後、文の数が４つの文章と、総文字数 4000 前後、総文数が 100 の文章について、文章末の文を比べた場合、文法上は当然のこと、文の大きさや構造にも見かけ上の差は見られない。しかし、そこで使用される単語の語彙的つながりの累積は、明らかに長い文章の方が大きい。ならば、逆に、文章が長くなった場合、語彙的つながりの累積を引き受けやすい単語が使用される頻度は高くなると予想できる。この観点から、あるテーマの文書のグループについて、文字数が大きくなるに連れて、文章末での使用頻度が高くなる名詞を大河タームとした。

本稿では得られた大河タームのシソーラス[7]上の分布について報告する。

2．文章作成指南書の文章末表現の扱い方

文章作成を指南する書物はハードカバー、新書、文庫、受験参考書にいたるまで膨大な数がある。このうち、受験参考書をのぞけば、文章の結び方に直接言及しているものは極めて稀である。間接的に、結びは重要である、読み手に読後の良い印象を与えるよう心がけるべきである、と説くものは多少あるが、結び方を指南していると言えるものではない。

逆に、小論文対策の高校、大学受験参考書には、結びの文の主題とその題述の仕方まで直接指定しているものさえある。これは、小論文問題に出やすいテーマ（最近では、ゴミ問題、福祉、ネットワーク社会、など）をそのまま取り上げ、採点者に失敗なく好印象を与えるような結び方の実例を、著者が選出しているからできることである（敢えて引用は控えるが、福祉がテーマであれば、「予算の充実と人員整備を自治体に要求していくことが必要である」などの記述で締め括れ、とまで述べられている）。一般の文章で、文章末の文で使う単語を直接指定することはおそらく不可能であろう。結局、現行の受験参考書は、小論文を暗記問題の拡張へと変成させており、もはや文章作成を指南するものではない。これらは文章作成指南書のカテゴリーから除外されるべきものとなっている。

本稿では、心構えの域を出て、文章の結び方に直接言及している２つの指南書について述べる。両者は極めて稀な例である。

１つ目は文献[8]である。これは筆者にとって好印象であった結び方の文章をいくつか取り上げ、そのまま例文として転載している。これらを分類することを試みているが、分類基準などが示されることもなく、もとより、少数の文章しかないため、具体的基準をしめすべくもない。結びの文ほど重要な部分はない、書き出しが立派でも結びが悪ければ、却って読み手に悪印象を与えるものである、と述べながらも、具体的指南には至らない。

２つ目は文献[9]である。これも良い結び方の例、悪い結び方の例をそのまま転載して、その良し悪しを論じている。具体的指南には至らない。但し、この書では、一歩踏み出して、文章の結びのあり方を「大河と小川」の対比という抽象的描像で表現した。本稿で用いている造語「大河ターム」はこの描像に由来している。小川とは一つ一つの文章の論旨の流れを表しており、それらは元来、トリビアルなものであり、決して大それたものにはなり得ないことの比喩として用いられている。大河とは、一般化できるものではないが、文章の論旨を包括するような上位概念、あるいは締め括りに用いられやすい定型的表現を比喩的に表している。筆者は、これら上位概念や定型的表現を用いることで、体裁を整えることのみに終始することを批判する。この描像は、上述の受験参考書にあるような定型的締め括り方（上述の福祉の例であれば、予算の充実と人員整備の要求などを用いること）に対する批判としても有効である。

さらに、小川の例が示唆するところは、文章には、それぞれ論旨の流れがあり、文章末の文はその流れを一旦止める役割を持っていることである。無論、流れを一旦止めるだけでなく、後発する同テーマの文章へと流れを引き継ぐ役割を兼ねる。これは１章で述べた、「語彙的つながりの累積」より一歩進んだ捉え方と言えよう。この見方を用いるならば、文章末の文において他の文と違った語彙の用いられ方が為されることに不思議さはない。この違いを大河タームとして具体的に表現し、文献[9]の抽象的描像を乗り越えることが一連の本研究の目標である。

３．抽出手順と既知の結果

３－１．手順

以下の手順０から手順８を適用して抽出された名詞を"大河ターム"と呼ぶ。抽出法は前掲の論文[1,2]と同様である。各手順の解説も[1,2]を参照されたい。

０．調査対象は新聞記事データベースを選び、本報告では日本経済新聞 CD-ROM、97年版[10]を使用した。

１．キーワード検索を行い、検索結果の文書数が 500 個から 1500 個までの文書群を全て選び、各、得られた文書をまとめて、１つの文書群とする。

２．検索して得られた文書群のうち、センテンス数が４未満のもの、あるいは文章でなく箇条書きに終始するものを除外する。

３．声明で記事が終わる場合は、発言内外を問わず、句点から句点まで、あるいは鉤括弧までを１つのセンテンスとする。

４．文書群内の全ての文書を文字数順に並べ、ほぼ同数になるようにX個のブロックに分割する。平均文字数最小のブロックを第０ブロックとし、昇順に順序付けを行い、平均文字数最大のブロックを第$(X-1)$ブロックとする。

５．各文書群において、全ての文書の末尾のkセンテンスを取り出し、これらに対して、日本語形態素解析プログラム「茶筌 1.0」を用いて形態素解析を行う[11]。この中から普通名詞、固有名詞、サ変名詞、時相名詞を取り出し、各名詞毎に集計を行う。

６．ある文書群jにおいて、第xブロック内の全ての文書の末尾kセンテンスにおける、ある名詞nの出現回数を$F(j,n,x)$とする。$F(j,n,x)$の最大値を$Fmax(j,n)$とする。文書群 j 内の全文書数をＡ(j)とする。この時、$Fmax(j,n') \geqq A(j)/X/P$ を満たさない場合、名詞n'を手順7,8 で行う解析から除外する。

7．頻度分布 $(x, F(j,n,x))$に対して、単純回帰分析を行い、回帰係数を求める。さらに、回帰係数の検定を行い、回帰係数の値が 0 である帰無仮説を有意水準 0.05 で棄却できる名詞のみを抽出する。また、得られた回帰係数 E が $E > t$ を満たすことを抽出の条件とする。

8．頻度分布 $(x, F(j,n,x))$に対して、歪み度 $V(n)$を求める。手順7で抽出された名詞の内、$V(n) < 0$ を満たすもののみを抽出する。

以上の手順により、文書群から抽出された名詞の集合をその文書群の大河タームと呼ぶ。抽出された大河タームとの比較のために各の文書群から２つのテーブルを抽出する。各の文書群全体の最頻出名詞の上位 w個を取り出して得られるテーブルを TF{w}（前掲の論文[1,2]では OTF と表記)、各の文書群の文書の末尾 k 文のみから名詞の頻度を求め、その最頻出名詞の上位 w 個を取り出して得られるテーブルを PF{w}（同じく、OPF と表記）と呼ぶ。

３－２．これまでの抽出結果

3-1 節で述べた手順の中のパラメータは、X=5, k=2, t=1.0,P=10 を用いた。各の値はヒューリスティックな設定である。

結果１：上記の方法の手順１により、495 個の文書群が得られた。この 495 個の文書群に対して、手順２から手順８までを実行した結果、480 個(～ 97.0%)の文書群から大河タームが得られた。

結果２：480 個の文書群から得られた大河タームはのべ 4158 個であった。他の文書群との重複を除けば、927 個であった。文書群毎の大河タームの個数の平均値は～ 8.45 (4158/480)であった。

結果３：同じ大河タームが複数の文書群から重複して抽出されることが起こる。ある名詞の重複抽出の回数が n 回なら、その名詞は n 個の文書群で大河タームとして抽出されたことになる。480 個の文書群から抽出された全ての大河タームについて、重複抽出の回数をカウントした。最大個数のものから上位 20 個は、企業(168)、日本(161)、今後(132)、市場(121)、問題(67)、経済(66)、改革(57)、経営(57)、会社(55)、事業(54)、競争(49)、米(49)、可能性(44)、銀行(36)、金融(35)、アジア(34)、開発(34)、技術(31)、情報(31)、地域(30)、となる。括弧内は文書群数を表す。これらを「最頻大河ターム{20}」と呼称する。こちらの括弧内は上位 20 個までという表示である。

順位	意味属性(５段目以上)	個数	順位	意味属性(５段目以上)	個数
1	人間	694	11	人(職業、地位、役割)	124
2	行為	615	12	様相	85
3	制度	395	13	機関	78
4	団体・党派	353	14	実質	41
5	精神	243	15	公共施設	36
6	非暦日	209	16	自然物	35
7	知的生産物(思考、学習)	202	17	伝承・情報・評判	35
8	人工物	176	18	地域(範囲)	34
9	界	155	19	行政区画	25
10	変動	155	20	内外	25

表１：大河タームが意味属性（６段以下の下位範疇に対応するものは５段目の親ノードに属するものとする）に対応する回数を集計したもの

結果４：これより、w=8 として、TF{8}を作成し、大河タームと比較する。文書群毎の大河タームは他のキーワード検索で得られる文書群においては頻出名詞となって TF{8}に含まれることは大いにあり得る。逆に全ての文書群の TF{8}に含まれないが、大河タームとして複数の文書群から重複して抽出されるようなものも存在する。全ての文書群の TF{8}に含まれず、かつ重複抽出の回数の多い大河タームの上位 20 個は以下の通りである。今後(132)、可能性(44)、業界(24)、政治(19)、見方(17)、考え(16)、コスト(15)、大手(14)、国内(13)、社会(13)、検討(12)、声(12)、自治体(11)、金融機関(10)、今回(10)、手(10)、各社(9)、関係(8)、程度(8)、影響(7)。括弧内は文書群数を表す。これらを「NTF 大河ターム{20}」と呼称する。こちらの括弧も上位 20 個の意である。

４．大河タームの意味属性

４－１：大河タームのシソーラス上の分布

927 種 4158 個全ての大河ターム及び全ての文書群の TF{8}と PF{8}について、シソーラス上の分布を比較した。TF{8}は 936 種、PF{8}は 880 種、両者とも総数はのべ 3960 個(=8*495)にのぼる。以下に方法について述べる。

１：使用するシソーラスは翻訳システム ALT-J/E の翻訳辞書でもある日本語語彙大系の意味属性体系を使用する。

２：大河ターム及び、TF{8}、PF{8}の全名詞について、シソーラス上で属する意味属性の段数を集計し、その平均値を比較した。その結果、大河タームの意味属性体系中の平均段数は 7.369、TF{8}の意味属性体系中の平均段数は 7.471、PF{8}の意味属性体系中の平均段数は 7.522 であった。

３：日本語語彙大系の意味属性体系を最下層まで用いれば、約 3000 の意味属性が使用可能であるが、927 種の大河タームを分類する基準としては大きすぎる。よって、シソーラスの５段目のノードを分類基準として設定した。各の名詞の単語意味属性値は、５段目以上のノードが代表するものとして、大河ターム、TF{8}、PF{8}の全名詞が、シソーラス上のどの属性に対応しやすいかをカウントした。６段目以下の下位範疇に属する名詞は５段目の親ノードに属するものとして集計し、分布を調べた。なお、今回は一般名詞意味属性体系に該当するものだけを扱った。５段目までのノードの数は 136 個である。集計の結果、上位 20 位までを表１から表３に列挙した。表１が大河ターム、表２が TF{8}、表３が PF{8}に対応する。

順位	意味属性(5段目以上)	個数	順位	意味属性(5段目以上)	個数
1	行為	716	11	知的生産物(思考、学習)	86
2	人間	582	12	機関	79
3	団体・党派	368	13	界	66
4	制度	340	14	公共施設	60
5	人工物	295	15	習俗	60
6	人(職業、地位、役割)	223	16	出来事	39
7	行政区画	220	17	創作物	35
8	精神	155	18	自然物	34
9	変動	115	19	仕事場	32
10	非暦日	89	20	実質	30

表２：TF{8}が意味属性（６段以下の下位範疇に対応するものは５段目の親ノードに属するものとする）に対応する回数を集計したもの

４：５段目以上のある意味属性 s について、属する名詞に大河タームが対応する個数を T-T[s]、TF{8}が対応する個数を TF[s]、PF{8}が対応する個数を PF[s]とする。このとき、下位範疇を含む意味属性 s において、大河タームが対応する個数と TF{8}が対応する個数の比を表すために、以下の指標、Ctf[s]を用いる。

$$Ctf[s] = T\text{-}T[s] \,/\, (T\text{-}T[s]+TF[s]).$$

同様に、PF{8}について以下の指標、Cpf[s]を用いる。

$$Cpf[s] = T\text{-}T[s] \,/\, (T\text{-}T[s]+PF[s]).$$

両指標とも、$0 \leqq Ctf[s] \leqq 1$、$0 \leqq Cpf[s] \leqq 1$ を満たす。５段目以上の意味属性の全てに対して、Ctf 値、Cpf 値を求めた。表４において、表１と同じ順序で意味属性を配して、Ctf 値、Cpf 値を列挙した。

４－２：結果

得られた表１から表３を比較する。３つの表において、１位から４位までの人間、行為、制度、団体・党派は、共通している。ところが、５位以降から特徴が現れる。表１（大河ターム）での５位と７位の「精神」「知的生産物（思考、学習）」は、筆者が意見を述べる時に用いる名詞が多い。「精神」が TF{8}では８位、PF{8}では６位となっており、「知的生産物（思考、学習）」では、TF{8}が 11 位、PF{8}では 10 位となっており、使用回数も少ない。大河タームで意味属性「精神」に区分けされた主な名詞として「改革」「考え」「声」があり、「知的生産物（思考、学習）」に区分けされた主な名詞として「問題」「技術」「政策」「計画」があげられる。

表１で６位の意味属性「非暦日」は、表２での 10 位、表３での 11 位と比べて、使用回数も約２倍である。「非暦日」に区分けされる大河タームは、「今後」が突出して多く使われており、「今回」「期」「時代」が続く。「今後」は 3-2 節で述べた NTF 大河タームの最頻出名詞でもある。また、大河タームの表１で９位の「界」は、「市場」「業界」が多く使用されている。TF{8}、PF{8}では、それぞれ 13 位で使用回数にも明瞭な差が見られる。

順位	意味属性(5段目以上)	個数	順位	意味属性(5段目以上)	個数
1	行為	728	11	非暦日	112
2	人間	567	12	機関	81
3	団体・党派	345	13	界	78
4	制度	336	14	習俗	66
5	人工物	251	15	公共施設	61
6	精神	203	16	出来事	41
7	人(職業、地位、役割)	202	17	創作物	33
8	行政区画	172	18	自然物	30
9	変動	138	19	伝承・情報・評判	29
10	知的生産物(思考、学習)	124	20	実質	27

表３：PF{8}が意味属性（６段以下の下位範疇に対応するものは５段目の親ノードに属するものとする）に対応する回数を集計したもの

さらに、大河タームの表１での12位での「様相」であるが、TF{8}で33位、PF{8}で37位と表に現れないながらも、明瞭な使用回数の差がある。この差もNTF 大河ターム「可能性」に依るところが大きい。

表４におけるCtf値とCpf値を比べると、「知的生産物（思考、学習)」「非暦日」「界」は、表１の上位10位の意味属性では、最も値が大きく、「精神」がそれらに継ぐ。また、12位の「様相」はCtf値が約0.88、Cpf値が約0.89と高い値を取っており、上述の結果を数量的に示している。

５．考察

上記の結果から、いくつかの特徴的な傾向が得られた。

１：大河タームの多くは、頻出名詞、あるいは文章末の２文のみから集計した頻出名詞と、意味属性が一致する。使用回数を比較すれば、上位４つの意味属性までは明瞭な差はない。

２：上位４つの意味属性を除けば、大河タームの特徴が顕現する。特に、筆者の意見を述べるための「精神」「知的生産物（思考、学習)」、未来について述べる「今後」を含む「非暦日」、状況について語るための「界」「様相」など、特徴的な意味属性が得られた。

３：大河タームとTF{8}、PF{8}の意味属性の段数の平均を比較した結果、明瞭な差は見られなかった。大河タームは、意味的な上位概念、あるいは包括概念という位置づけではその特徴を表せない。「精神」「知的生産物（思考、学習)」「非暦日」「界」「様相」などの意味属性が、特徴を表す一つの指標となる。

６．展望

今後の調査予定、展望を述べておく。

１：文章末の文において、「精神」「知的生産物（思考、学習)」「非暦日」「界」「様相」などの意味属性を持つ名詞の使用頻度を改めて求める。あるテーマでは大河タームとして検出されなくても、これらの意味属性を持つ名詞が使用されている文章は相当数存在すると見込まれる。

２：これらの下位範疇に絞って、大河タームとしての使用度を抽出する。

３：文章末の文の文体を調査する。特に「非暦日」に属する時間表現の使用の仕方の特徴を捉える。

これらについては、別稿にて報告を予定中である。

	意味属性(5段目以上)	Ctf	Cpf		意味属性(5段目以上)	Ctf	Cpf
1	人間	0.54	0.55	11	人(職業、地位、役割)	0.36	0.38
2	行為	0.46	0.46	12	様相	0.88	0.89
3	制度	0.54	0.54	13	機関	0.50	0.49
4	団体、党派	0.49	0.51	14	実質	0.58	0.60
5	精神	0.61	0.54	15	公共施設	0.38	0.37
6	非暦日	0.70	0.65	16	自然物	0.51	0.54
7	知的生産物(思考、学習)	0.70	0.62	17	伝承、情報、評判	0.56	0.55
8	人工物	0.37	0.41	18	地域(範囲)	0.67	0.60
9	界	0.70	0.67	19	行政区画	0.10	0.13
10	変動	0.57	0.53	20	内外	0.89	0.81

表４：Ctf値、Cpf値（表１の順序の沿って表示）

謝辞

日本語形態素解析システム茶筌を開発し、精度向上に取り組み続けている奈良先端科学技術大学院大学松本研究室の皆様に謝意を申し上げます。また、調査を許可下さった日本経済新聞社殿に感謝いたします。

参考文献

[1] 中村隆志、本間愛：新聞記事における文章末表現での名詞語彙の出現特性、情報処理学会報告、NL128-13,1998

[2] 本間愛、中村隆志：新聞記事における文章末表現における特異的名詞語彙の出現特性、情報処理学会報告、CH41-2,1999

[3] 任福継、定永靖史、統計情報と文章構造特徴に基づく重要文の抽出、情報処理学会研究会報告、NL,125-7、1998

[4] 吉見毅彦、奥西稔幸、山路孝浩、福持陽士、表題へのつながりに基づく文の重要度評価、自然言語処理、6-1、43-57、1999

[5] 野本忠司、松本裕治：テキスト構造を利用した主題の推定について、情報処理学会報告、NL114-8,1996.

[6] 山本和英、増山繁、内藤昭三：文章内構造を複合的に利用した論説文要約システムGREEN、自然言語処理, vol. 2(1), pp. 39-55,1995.

[7] 池原悟、他編：日本語語彙大系、岩波書店、1997

[8] 井上敏夫、「良文に学ぶ一文章鑑賞と文章作法」、明治図書、1988

[9] 加藤典洋：言語表現法講義、岩波書店、1996.

[10] 日本経済新聞社、日本経済新聞97年CD-ROM版、日本経済新聞社、1997.

[11] 松本裕治、北内啓、山下達雄、平野喜隆、今一修、今村友明：日本語形態素解析システム『茶筌』version 1.0 使用説明書、Information Science Technical Report, NAIST-IS-TR97007, 奈良先端科学技術大学、1997.