人文科学とコンピュータ 55－4
（２００２．７．２６）

多国語新聞記事の大河ターム分析（その2）

中村　隆志　　**矢野　理恵**
tks@human.ge.niigata-u.ac.jp　　w01d407b@mail.cc.niigata-u.ac.jp
新潟大学人文学部

一連の研究及び先の報告をふまえて、大河タームを3つに分類して(IFW大河ターム、OFW大河ターム、NFW大河ターム)それらの頻度を分析した。頻度の分析は3年度分のCD-ROMでの使用頻度の比較という形で行った。その結果、各年度で類似した頻度分布をしていること、使用頻度の高い単語が一致することが確認された。さらにシソーラスを用いて意味属性の分布をカウントして、３年度間の傾向を比較した。その結果、3つの大河タームの内、最も特徴的であるNFW大河タームについて、単語間の比較では見られなかった傾向があることが確認された。

キーワード：新聞記事、文章末表現、シソーラス、大河ターム

Taigua Terms Analysis of Newspaper articles in Multi Language Ⅱ

Takashi Nakamura　　Rie Yano
tks@human.ge.niigata-u.ac.jp　　w01d407b@mail.cc.niigata-u.ac.jp
Facultyof Humanities, Niigata University

The present author and his colleagues have focused on the closing sentences of newspaper articles. Comparing the most frequent nouns intheclosing sentenses in the articles of all the groups, taigua terms were divided into 3 types, IFW-taigua terms, OFW-taigua terms and NFW-taigua terms. Most of the frequent words of the three types of taigua terms were common in each year's database. The frequency of the semantic feature values of the three types of taigua terms was measured using a thesaurus. The usage tendencies of NFW-taigua terms in the three years' database were more prominent in the order of the semanticfeature thanintheorder of thenouns.

keywords: Newspaper articles, closingsentences,thesaurus,TaiguaTerm

1:序

文章末を取り上げる論考では、最終文の形式、機能、心理的な効果について典型的な例が述べられることが多い（例えば[1-3])。先行研究[4-7]を含め、筆者らは異なる観点から文章末に注目している。それは、

1:文章末の文では、必要に応じて先行する文の曖昧性を補填しながら、自身には曖昧性の少ない表現が使われやすい。(図1)

2:先行する文脈や大意の曖昧性を解消する機能をもつ表現が文章末で使われる頻度は、長い文章の方が相対的に高い。(図2)

という2つの仮定から出発し、長い文章の文章末で出現しやすい単語とその意味属性を特定する、というものである。

第1の仮定は、最終文には後続する文がないことから導かれるものである。第2の仮定について、示唆的な傍証を述べておく。野本・松本[8]は新聞記事における主題の推定について、テキスト構造を利用することで、推定の精度が向上することを見いだしている。論考の中で、本文の冒頭から特定の単語数からなるブロックを切り出して推定に用いる方法(FLM方式)の有効性を示すとともに、推定精度そのものは本文の長さに応じて下がることが付せて述べられている。このことは、文章の長さが構成、文脈、内容の展開に少なからず影響していることを表すと考えられる。

先回の報告[7]の結果をふまえて、分節された大河イメージに即した集計と分析を行った。本稿では、大河タームを大きく３つに分類し、その出現頻度について3年度間の比較を報告する。資料は先回と同じく、日経新聞CD-ROMの97,9

8,99年度版[9-11]の 3年分を用いた。

2.大河タームの抽出と分類

大河タームの定義と抽出法を短く再掲する。

A:資料 :日経新聞CD-ROMの97,98,99年度に含まれる記事。但し、文章の形態でない記事を除いたため、使用した記事数は重複を除けば、97年度102610個、98年度104454個、99年度101450個である。

B:定義 :y年度の新聞記事データベースから、テーマAの文章を集めたものを、「y年度におけるテーマAの文書グループ」と呼ぶ。各テーマの文書グループに順序をつけて、第i番目のグループを「y年度における第iテーマの文書グループ」と呼ぶ。大河タームは以下のアルゴリズムによってグループ毎に抽出される。

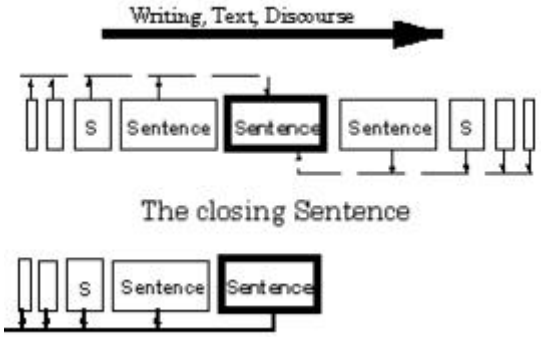

図1:最終文の他の文とのつながり

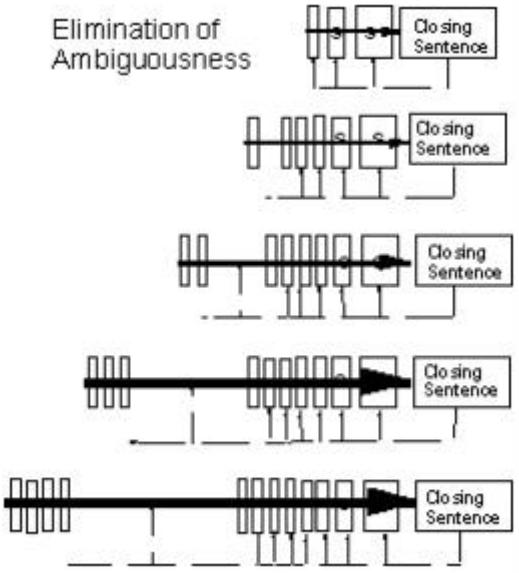

図2:文章の長さと曖昧性の増大

C:大河ターム抽出プログラム version 4

1. 新聞記事CD-ROM内の記事の内、以下の条件を満たすものは文章の体裁をとっていないものとして除外する。
   1-1:文の数が 3以下
   1-2:文字数が200以下
   1-3:箇条書き
   1-4:図表
   1-5:スポーツの結果、書籍などの売り上げランキング、イベント告知
   1-6:円相場、先物取引の相場、日銀概況
   1-7:賞与、会社人事、死亡記事、家屋移転
   1-8:インタビュー記事、首相の所信表明演説など口述録音の書きおこし
2. キーワード検索を用いて、同一テーマの記事を集め、それらを一つの文書グループとする。以下の条件を満たす文書グループだけを抽出に用いる。
   200 =< 記事数 < 2000
3. 得られた文書グループ内の全ての記事を文字数順に並べる。

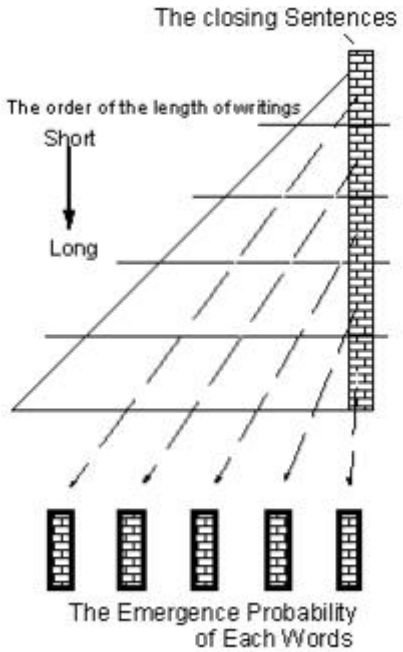

図3:文書グループのブロック分けと最終文の取り出し

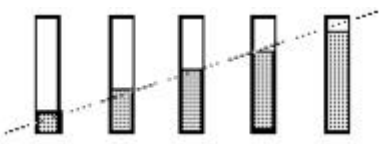

図4:ブロック別使用頻度の例

4.記事数が同数になるように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う。
5.文書群の全ての文書の末尾kセンテンスを取り出して、形態素解析を行う。[12]
6.各単語毎にブロック別の集計を行う。
7.第j文書グループにおける第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単語nの出現回数を$F(j,n,x)$とし、出現回数最大のブロックの出現回数を$Fmax(j,n)$とする。このとき単語nは

$$Fmax(j,n) \geqq A(j)/X/P$$

を満たさねばならない。
8.頻度分布$(x, F(n,x))$に対して、単純回帰分析を行い、回帰係数の推定値がtより大で、かつ、回帰係数の検定において、回帰係数の値が0である帰無仮説を有意水準0.05で棄却できる単語のみを抽出する。
9.上の手続きで抽出される単語の各を「y年度の第j文書グループの大河ターム」と呼ぶ。

D:比較指標FW[x]:各文書グループから得られる上位x位までの頻出名詞の集合である。大河タームのグループ別平均抽出個数が7であるため、全グループについて一律FW[7]を用いる。

先行研究において大河タームは頻出名詞との比較の上で理念的に3つに分類されることを示した。

1: ある文書グループにおいて大河タームとして抽出されるが、同時にその当該グループの頻出名詞であるもの。文書グループのテーマに依存して頻出する非常に重要な名詞である。
2: ある文書グループにおいて大河タームとして抽出され、かつ当該グループの頻出名詞ではないが、同一年度の他の文書グループの頻出名詞であるもの。他の文書グループで頻出しており、他の文書グループの内容や関係を間接的に含むと考えられる。
3: ある文書グループにおいて大河タームとして抽出され、かつ同一年度の全ての文書グループにおいて頻出名詞とならないもの。使用されるテーマに依存せず、他のグループでの使用状況にも依存せずに長い文章の文章末に使用されやすい名詞である。

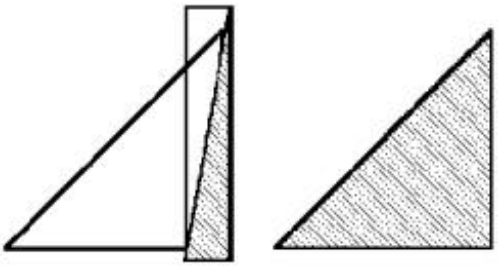

図5: 文書グループ内で抽出に用いる範囲。
左が大河ターム、右がFW

これら3パターンの頻度を集計し、3年度間の共通した傾向を見いだすために以下の定義を行う。先行研究では比較の指標としてFC[7]を併用したが、本稿ではFW[7]のみで行う。各大河タームは年度ごとに集計される。

***IFW大河ターム：** 同一年度の大河タームの内、抽出された当該グループのFW[7]の要素であるもの。
***0FW大河ターム：** 同一年度の大河タームの内、抽出された当該グループと別の文書グループのFW[7]の要素であるもの。
***NFW大河ターム：** 同一年度の大河タームの内、その年度の全文書グループのFW[7]の要素にならないもの。

同じ名詞が別々の文書グループから大河タームとして重複抽出されることは頻繁に起こる。重複の頻度は大河タームによって異なる。よって、大河タームは重複するグループ数によって、頻度を比較可能である。以下に用語と集合を定義して、上記３分類の大河タームの出現頻度を比較する指標、及び全体との比較のための指標を導入する。

***最頻** IFW-TT[x]：重複回数の多いもの上位x位までのIFW大河タームの集合。
***最頻** OFW-TT[x]：重複回数の多いもの上位x位までのOFW大河タームの集合。
***最頻** NFW-TT[x]：重複回数の多いもの上位x位までのNFW大河タームの集合。
*最頻 TT[x]:同一年度の大河タームの内、、重複回数上位x位までの大河タームの集合。

この定義から、各グループから抽出される全ての大河タームはIFW大河ターム、OFW大河ターム、NFW大河タームのいずれかとなる。

OFW、NFW大河タームについて具体例を挙げよう。「影響」「業界」「資金」「通信」の4つの名詞は99年度の文書グループ「2000年問題」の大河タームである。これらは99年度の文書グループ「2000年問題」におけるFW[7]の要素ではなく、従ってIFW大河タームではない。これら4つの名詞が99年度の全文書グループから抽出される大河ターム、及びFW[7]の要素と、どれぐらい重複するグループを持つかを表1に示す。各の数値は重複するグループ数を示す。

4つの大河タームとも、複数の文書グループの大河タームとして重複して抽出される。1つ目の「影響」と残りの3つ「業界」「資金」「通信」は性質が異なることがわかる。「影響」は99年度の全文書グループから抽出されるFW[7]の要素になることがない。一方、「業界」「資金」「通信」は99年度の「2000年問題」以外の多くの文書グループのFW[7]の要素となる。つまり、99年度におけるテーマ「2000年問題」については「影響」がNFW大河ターム「業界」「資金」「通信」がOFW大河タームとなる。

但し、99年度で抽出される全ての「業界」「資金」「通信」がOFW大河タームとは限らない。別のグループでこれらが大河タームとして抽出された場合、そのグループでのFW[7]の要素であることは起こり得る。つまり、別々のグループで重複して抽出された、同じ名詞の大河タームは、あるグループではIFW大河タームとして、別のグループではOFW大河タームとして分類されることが起こり得る。一方、NFW大河タームとなる名詞は、定義上、同一年度の全ての文書グループのFW[7]の要素にならないことから、他文書グループのIFW大河ターム、OFW大河タームのいずれにもなることはない。

	影響	業界	資金	通信
TTの重複回数	4	30	32	6
FW[7]との重複回数	0	2	23	8

表1 :大河タームの重複例

3. 名詞別集計結果

各年度の全ての大河ターム、及びIFW大河ターム、OFW大河ターム、NFW大河タームについて、その総数を求めた。年度別の数値を表1に示す。TTは大河ターム全体の総数、IFW-TTはIFW大河タームの総数を示す(OFW-TT、NFW-TTも同様)。IFW大河ターム、OFW大河ターム、NFW大河タームの総数のそれぞれに付く括弧内の数値は当該年度の大河タームの総数との割合である。

	97年度	98年度	99年度
TT	7843	8231	6445
IFW-TT	1562(19.9%)	1597(19.4%)	1218(18.9%)
OFW-TT	5007(63.8%)	5363(65.2%)	4090(63.5%)
NFW-TT	1274(16.2%)	1271(15.4%)	1137(17.6%)

表2 :年度別、各大河タームの総数

表2において、大河タームの総数、及び各IFW、OFW、NFW大河タームにおいて使用頻度の割合が3年度間でほぼ一定の傾向にあることが見て取れる。各年度においては、話題として取り上げられるトピックは年々変化しており(住民投票、沖縄基地問題、統一地方選挙、毒物カレー事件、2000年問題など)、大河タームも入れ替わる一方で、年度単位でマクロ的に見てやれば、文章の末尾に使用されやすい名詞は一定の割合で存在していることが示される。

各年度の最頻TT[15]を表3に、最頻IFW-TT[10]を表4に、最頻OFW-TT[10]を表5に、最頻NFW-TT[10]を表6に示す。

表3の結果から、上位の重複数の大きな大河タームの出現頻度は、3年間で近い傾向にあることが見て取れる。出現頻度の傾向の近さを測るため、以下の指標z値を定義する。

$$z(i) = i - |\text{最頻 TT}[i,97] \cap \text{最頻 TT}[i,98] \cap \text{最頻 TT}[i,99]|.$$

上式の最頻TT[i,97]、最頻TT[i,98]、最頻TT[i,99]は、それぞれ97年度、98年度、99年度の最頻TT[i]を示す。$|x|$は集合xの要素の数を表す。各項の最頻TTを入れ替えて最頻IFW-TT[10]、最頻OFW-TT[10]、最頻NFW-TT[10]についても同様の計量を行う。表3を例に取ると、i=1の時、つまり各年度の1位の大河タームについて、それぞれ「の」「こと」「企業」となるため、積集合は空集合となり、z(1)=1。i=2の時、つまり各年度の2位までの大河タームについては、97年度2位、98年度1位、99年度2位の「こと」が共通なので、積集合の要素の数は1となり、z(1)=1。i=3の時は「の」「こと」が共通なのでz(3)=1。つまり、z値は3年間の共通分からの各年度の誤差を示す。i=12の時、

積集合の要素数が 11 であり、$z(12)=1$ であるが、$i=13$ の時、$z(13)=2$ となり、$z(14)=3$ となる。$i>12$ 以降、z 値が大きく上昇していくことから、本稿では $z(i)<2$ を満たしている順位内では 3 年度間の出現頻度に近い傾向があるものとする。

同様に最頻 IFW-TT[10]、最頻 OFW-TT[10]、最頻 NFW-TT[10]についても同様の z 値の計量を行った結果(表4、表5、表6の最右列)、IFW 大河タームについては第 6 位まで、OFW 大河タームについては第 8 位まで、NFW 大河タームについては第 3 位までが $z<2$ を満たしていることから、3年度間で同様の傾向を示していることが見て取れる。特に NFW 大河タームについては、先行研究でも指摘したとおり、「今後」と「可能性」が突出して重複グループ数が多い。

	97年度		98年度		99年度		
Or	TT	No	TT	No	TT	No	z
1	の	502	こと	478	企業	329	1
2	こと	476	の	446	こと	326	1
3	日本	329	企業	372	の	323	1
4	企業	321	日本	349	日本	197	0
5	市場	237	市場	208	今後	182	1
6	今後	220	経済	186	市場	172	1
7	ため	160	今後	183	ため	156	1
8	経済	112	ため	175	事業	113	1
9	事業	103	事業	121	可能性	106	1
10	米	102	金融	120	経済	100	1
11	改革	100	米	118	米	80	1
12	経営	100	経営	107	経営	78	1
13	競争	90	可能性	101	地域	60	2
14	会社	86	競争	98	問題	58	3
15	問題	86	銀行	88	会社	56	4

表3 :年度別最頻TT[15]

	97年度		98年度		99年度		
Or	IFW-TT	No	IFW-TT	No	IFW-TT	No	z
1	こと	305	こと	311	こと	225	0
2	の	187	の	145	企業	126	1
3	企業	100	企業	109	の	98	0
4	日本	91	日本	82	日本	44	0
5	米	54	経済	63	米	43	1
6	市場	40	米	59	市場	34	1
7	事業	32	銀行	40	銀行	27	2
8	会社	23	市場	36	経済	22	2
9	銀行	22	事業	35	事業	18	1
10	開発	18	市	25	会社	16	2

表4 :年度別最頻IFW-TT[10]

それら以下の単語に関しては重複数も小さく、重複数の比較だけから年度間の傾向を見いだすことは難しい。

4. 意味属性による集計結果

表3を見る限り、出現頻度の高い大河タームの使用状況は3年度間で一定の傾向を持つと考えられる。表3から得られるのは大河タームが抽出される重複グループ数であり、これは名詞の使用頻度を反映しているものの、文章末において、どんな概念が使われやすいか、どの意味属性を持つものが使われやすいかまでを推定することはできない。よって、全ての大河タームについて、意味属性値の分布を集計した。意味属性を与える指標として、先行研究と同様に日本語語彙大系[13]の中の単語意味辞書と単語意味属性体系を使用した。単語意味辞書から

	97年度		98年度		99年度		
Or	OFW-TT	No	OFW-TT	No	OFW-TT	No	z
1	の	315	の	302	の	225	0
2	日本	238	日本	267	企業	203	1
3	企業	221	企業	263	日本	153	0
4	市場	197	市場	172	ため	145	1
5	こと	171	こと	167	市場	138	1
6	ため	150	ため	157	こと	101	0
7	経済	99	経済	123	事業	95	1
8	改革	92	競争	98	経済	78	1
9	競争	90	金融	97	経営	63	2
10	経営	90	経営	93	問題	57	2

表5 :年度別最頻OFW-TT[10]

	97年度		98年度		99年度		
Or	NFW-TT	No	NFW-TT	No	NFW-TT	No	
1	今後	220	今後	183	今後	182	0
2	可能性	82	可能性	101	可能性	106	0
3	声	20	回復	30	見方	18	1
4	導入	19	収益	24	国内	17	2
5	検討	18	特捜	20	声	16	3
6	見方	18	見方	17	考え	16	3
7	特捜	18	声	16	期待	13	3
8	考え	15	期待	15	リストラ	13	3
9	予想	15	地元	15	指摘	12	4
10	銘柄	13	分野	15	国	12	5

表6 :年度別最頻NFW-TT[10]

各大河タームの意味属性値を得た後、単語意味属性体系の各ノード上での個数分布を集計した。単語意味属性体系の最深12段、約2700のノードのうち、5段目までのノードを用いて、大局的に分類する。5段目以下のノードの意味属性を持つ単語については、その5段目の親ノードが意味を代表するものとして頻度を集計した。表3から表6までは、大河タームについての名詞単位の重複グループ数を示すが、以下の表7から表10では、意味属性ノード単位の重複グループ数を示している。

	97年度		98年度		99年度		
	意味属性(5段目以上)	計	意味属性(5段目以上)	計	意味属性(5段目以上)	計	z
1	行為	1031	行為	985	行為	918	0
2	団体 ･党派	650	制度	792	団体 ･党派	580	1
3	制度	637	団体 ･党派	683	制度	551	0
4	類	514	事	479	類	329	1
5	事	476	類	456	事	326	0
6	精神	365	非暦日	328	精神	288	1
7	知的生産物(思考 ･学習)	319	精神	289	非暦日	285	1
8	非暦日	312	人工物	285	変動	260	1
9	界	308	界	255	人工物	244	2
10	人工物	304	変動	245	知的生産物(思考 ･学習)	218	2

表 7:年度別大河ターム全体についての意味属性値の集計結果

	97年度		98年度		99年度		
	意味属性(5段目以上)	計	意味属性(5段目以上)	計	意味属性(5段目以上)	計	z
1	事	305	事	311	事	225	0
2	類	187	団体 ･党派	196	団体 ･党派	189	1
3	団体 ･党派	179	制度	153	行為	155	1
4	行為	148	行為	149	類	98	1
5	人工物	105	類	145	制度	87	1
6	制度	73	人工物	92	人工物	75	0
7	人 〚職業 ･地位 ･役割〛	42	人 〚職業 ･地位 ･役割〛	47	人 〚職業 ･地位 ･役割〛	38	0
8	界	40	行政区画	38	界	34	1
9	知的生産物(思考 ･学習)	32	界	37	変動	23	1
10	精神	32	公共施設	28	精神	21	2

表8 :年度別IFW大河タームについての意味属性値の集計結果

	97年度		98年度		99年度		
	意味属性(5段目以上)	計	意味属性(5段目以上)	計	意味属性(5段目以上)	計	z
1	行為	783	行為	736	行為	686	0
2	制度	532	制度	594	制度	430	0
3	団体 ･党派	459	団体 ･党派	472	団体 ･党派	376	0
4	類	318	類	304	類	226	0
5	界	266	界	218	変動	193	1
6	知的生産物(思考 ･学習)	251	事	167	知的生産物(思考 ･学習)	175	2
7	精神	177	理由 ･目的等	160	界	170	2
8	事	171	人工物	157	人工物	150	3
9	人工物	169	知的生産物(思考 ･学習)	154	精神	150	2
10	理由 ･目的等	150	変動	148	理由 ･目的等	147	2

表9 :年度別OFW大河タームについての意味属性値の集計結果

	97年度		98年度		99年度		
	意味属性(5段目以上)	計	意味属性(5段目以上)	計	意味属性(5段目以上)	計	z
1	非暦日	264	非暦日	249	非暦日	223	0
2	精神	156	様相	134	様相	118	1
3	様相	110	精神	118	精神	117	0
4	行為	100	行為	99	行為	77	0
5	変動	80	変動	73	変動	44	0
6	知的生産物(思考・学習)	36	制度	44	制度	34	1
7	制度	32	人工物	36	人〈職業・地位・役割〉	30	1
8	人工物	30	知的生産物(思考・学習)	30	知的生産物(思考・学習)	28	1
9	言語	21	人〈職業・地位・役割〉	16	因果	27	2
10	人間	18	郷里	15	機関	21	2

表10：年度別NFW大河タームについての意味属性値の集計結果

表 7 から表 10 まで、上位の意味属性の出現傾向は共通分が大きい。前章の z 値を表 7 の意味属性の順位に転用するために以下の定義を行う。

*最頻 SFTT[*x*,*y*]：第 *y* 年度の大河タームについて意味属性ノード単位で重複回数の上位 *x* 位までの意味属性の集合。

前章同様、

$$z(i) = i - |\text{最頻 SFTT}[i,97] \cap \text{最頻 SFTT}[i,98] \cap \text{最頻 SFTT}[i,99]|$$

に従って z 値を求める。IFW 大河ターム、OFW 大河ターム、NFW 大河タームの場合も同様に行う。$z(i)<2$ を満たす最大の順位は表 7 で第 8 位、表 8 で第 9 位、表 9 で第 5 位、表 10 で第 8 位となる。各大河タームの z 値の単語単位の集計と意味属性ノード単位の集計の違いを表 11 に示す。

表11において、意味属性ノード単位の集計を単語単位の集計と比べると、大河ターム全体、OFW大河タームでは、共通する順位を下げている。その一方で、IFW大河ターム、NFW大河タームは共通と見なせる順位を上げている（太字）。

	TT	IFW	OFW	NFW
単語単位	12	6	8	3
意味属性単位	8	**9**	5	**8**

表 11：各大河タームの集計単位と出現傾向が近いと見なせる順位

特にNFW大河タームについて、表6では上位2位までを除いては顕著な共通性は見いだせなかったが、表10の集計結果を見れば、上位8位程度まではほぼ共通した傾向を見いだすことができる。NFW大河ターム「今後」「可能性」をそれぞれ含む「非暦日」「様相」が上位であるのは表6からの帰結といえるが、「精神」「行為」「変動」「知的生産物(思考・学習)」「制度」「人工物」が上位に共通してあることは注目に値する。

5. 考察

表7から表11の要点を述べる。

1: 意味属性ノード単位で大河タームの重複グループ数を集計した結果、この場合も出現頻度の年度間の共通性が確認された。

2: 単語単位の場合と同様、重複グループ数の大きい意味属性の中には、IFW大河タームでの頻度が高いものとOFW大河タームでの頻度が高いものがある。

3: 単語単位の重複グループ数では一定の傾向が見いだしにくかったIFW大河ターム、NFW大河タームについて、意味属性ノード単位の集計にすることで、それぞれ上位9位程度、8位程度までで共通の出現傾向が見られた。

表8におけるIFW大河タームの、表10におけるNFW大河タームの意味属性での重複グループ数の集計結果は、出現頻度の低い名詞を集めたことが原因で年度間での共通性を得ることができたことを示している。最も顕著な例として、表10で3年度とも出現頻度が5位になった意味属性

変動」を取り上げる。意味属性ノード「変動」あるいはその下位範疇のノードに属するNFW大河タームの各年度で頻度の高いものを以下に挙げる。

97年度：「導入(19)」「再編(8)」「破たん(8)」「調整(7)」「抜本(5)」他15個。

98年度：「回復(30)」「介入(5)」「破たん(5)」「控除(3)」「連合(3)」他19個。

99年度：「導入(8)」「普及(8)」「再開(4)」「安定(3)」「追及(3)」他19個。

括弧内は大河タームとしての重複グループ数である。上位5つのNFW大河タームを見る限り、3年度間で共通する傾向を見いだしにくく、また、各の大河タームの重複グループ数も小さい。これら出現頻度の小さな大河タームの重複グループ数を意味属性単位で集計した結果、共通した傾向を見いだすことができた。

6. まとめ

大河タームの出現頻度について、重複して現れるグループ数を用いて3年度間で比較した結果、上位12位程度まで同様の傾向を得た。大河タームをIFW大河ターム、OFW大河ターム、NFW大河タームに分類した。これらの重複グループ数を3年度間で比較した結果、IFW大河タームとOFW大河タームでそれぞれ上位6位と8位程度までで同様の傾向を得た。NFW大河タームでは上位2位の「今後」「可能性」に顕著な結果が出たが、下位の名詞では共通した傾向を見いだせなかった。

大河タームを日本語語彙大系の意味属性体系を用いて意味属性ノード単位で重複グループ数を集計した結果、単語単位での比較結果と同様、上位の意味属性の出現頻度に共通した傾向を見いだすことができた。IFW大河タームとNFW大河タームでは、単語単位で見いだせなかった3年間での共通性が、それぞれ高い順位にまで見いだされた。特にNFW大河タームは定義から長い文章の文章末での使われやすさに一般性が高いと考えられるため、これら上位の意味属性を特定できたことは、文章末の表現を分析していく上で意義深いと考えられる。

短い文章には出現しにくく、長い文章の文章末に出現しやすい名詞を取り出した結果、いくつかの特徴的な名詞と意味属性を抽出することができた。新聞記事という特殊な文章だけを用いており、また抽出アルゴリズムや分析方法にも何段階かのヒューリスティックな判断が入っているものの、3年度間でほぼ共通した傾向を得ることができた。ここで得られた結果は、文章の結尾、書き納めのタイプを再考する上での手がかりとなること、ひいては文脈解析のための理論的寄与をすることが期待される。

参考文献

[1] 市川孝：『国語教育のための文章論概説』、教育出版、1978.

[2] 市川孝：『改訂文章表現法』、明治書院.

[3] 進藤咲子：「書き終わりのタイプ」、『国文学：解釈と鑑賞』臨時増刊号、1974-6.

[4] 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.

[5] 中村隆志、小泉明日美、本間愛：「日本語新聞記事の文章末における特異的名詞」、情報処理学会報告、ICS116-4, 1999.

[6] Takashi Nakamura, Tatsuo Hemmi & Asumi Koizumi: Semantic Features of specific words in the closing sentences of news paper articles, Proceedings of The second Annual Conference of The Japanese Society of Language Sciences, 2000.

[7] 中村隆志、廣木真理：「多国語新聞記事の大河ターム分析（その1）」、情報処理学会報告、CH48-5

[8] 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.

[9] 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、1998.

[10] 日本経済新聞社、日本経済新聞98年CD-ROM版、日本経済新聞社、1999.

[11] 日本経済新聞社、日本経済新聞99年CD-ROM版、日本経済新聞社、2000.

[12] 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.

[13] 池原悟、他編：「日本語語彙大系」、岩波書店、1997.