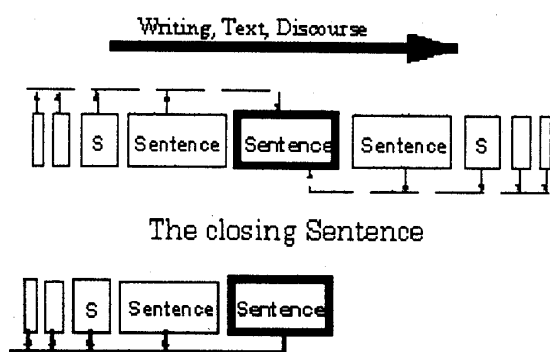


される。しかし、文章末では許されない。後に続く文がない以上、曖昧な言葉使いをすることは、文意を壊す危険さえある。文章末の文が先行する文の曖昧性を解消することはあっても、その文自体が曖昧性を残すことは慎まれる。文章末の文は不可避免的にこの性質を持つ。

解消すべき曖昧性は語義的なものばかりではない。多くの文章には「行間」が存在する。すなわち、文章を読む場面においては、節と節の間、文と文の間、段落と段落の間を結ぶ推論を必要とする。この推論は陽には現れないために、時として筆者の意図に反した誤解を読者に生じさせる。つまり、文章は語義以外の表面に現れない曖昧性を持ちうる。文間の推論から発する曖昧性を解消する必要がない文章も大いにあり得るが、必要とする文章も多く存在する。文間の推論が曖昧性を生じさせうる可能性を考慮するならば、後続する推論を持つことがより少ない文の方が全体の曖昧性を解消する機能を果たしやすい。後続する文も推論も持たない文章末の文が、曖昧性の解消には最も効率的であることになる。この性質は、文章末で「まとめ」を行う文章が決して少なくないことの状態証拠になる。この特異な性質を持つ文章末の文において、使われやすい概念を意味的に捉えることが本研究の当面の目標である（推論に関しては、例えば文献[4]第8章）



2 : 「大河」イメージ

「大河ターム」という呼称における「大河」という言葉は、文献[5]に由来する。現在市販されている文章作成のための指南書は様々なものがある。レポートの書き方などの報告書の書き方のものから、文章読本のよう

な文学的な文章の書き方を説くものがあり、文章論などの理論的背景を持つものから、筆者の経験に基づく方法論を展開するものまである。しかし、文章を如何に締めくくるか、文章末にどんな表現を用いるか、という具体的な指南はほとんど見あたらない。「文章の末尾は大事な部分である」「自分の意見を簡潔に述べるべき」「読者に好印象を残すことが必要だ」などの説明は少ないながらも存在する。しかし、大事だからどうすればよいのか、どういう単語が使いやすいのか、どんな表現を用いれば好印象を残すのか、という具体性のある説明はなく、必要性を説くのみである。

文献[5]は「小川と大河」というイメージの対比を用いて文章の終わり方について述べている。終わり方と意見のまとめ方にはっきりとした区別はしていない。大学生のための文章作成指南として書かれた本書では、「小川」とは個々の学生が書く文章のイメージであり、個別的でちっぽけな存在であることを含意している。一方で「大河」とは当該文章のテーマにおけるありがちなまとめ方のイメージであり、世間一般の態勢的な意見や評価を含意している。この2つのイメージを対比的に述べた後、一つ一つの文章はちっぽけな存在でありながらも流れがあるものであり、まさに小川のようなものであるが、大河に呑み込まれるだけで終わってはいけない、むしろ大河を貫通するぐらいの強い流れをもつ小川であれ、と説く。つまり、態勢的な意見に便乗して「ありがちな」まとめ方をすることはよろしくないが、それを避ければよいというものでもなく、ありがちなまとめ方を貫いて乗り越えるぐらいの力強いまとめ方をせよ、と述べている。

文献[5]の主張の中には大学生という若者を鼓舞する目的を含むであろうから、「大河に呑み込まれるな」というメッセージがなされるが、一般の文章ではありがちなまとめ方がなされることに問題はない。また、ありがちなまとめ方とは、それを乗り越えようとするが故に使用される頻度が高くなる可能性がある。先のイメージに乗っ取って表せば、小川が大河に呑み込まれる場合は当然のこと、流れに抗って貫く場合にでも大河と関与する終わり方をすることになる。つまり、ありがちなまとめ方をそのまま使わなくても、

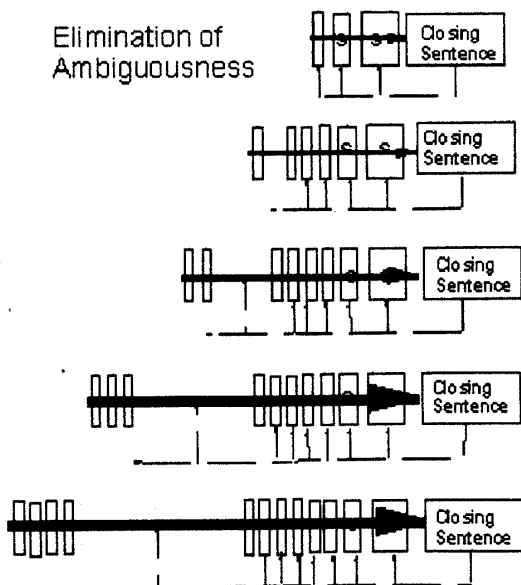
それに関連する表現が為されるということであり、それ故にありがちなまとめ方を構成する表現の一部はますます使われる傾向が高くなる可能性がある。さらに、ありがちなまとめ方が文章末でなされるならば、そのまとめ方は曖昧性に対して強い解消力を持つ表現を含んでいることを示唆する。曖昧性の解消を迫られる文章末においては、解消力の強い表現が好んで使われる可能性が高く、ありがちはますます助長され、強固なものになる可能性がある。

文献[5]で示された抽象的イメージを援用して、本研究で用いる「大河」イメージをまとめておこう。それは文章末においてまとめに使われやすく、先行する曖昧性を解消しやすい単語や表現のイメージである。世間一般の態勢的な意見や考え方とは直接関係がない。あくまでもありがちで強力なまとめ方を構成するような単語や表現のことである。本研究の目的はこれらイメージの抽象性から出発し、具体的モデルを導くことにある。

3：抽出アルゴリズム

アルゴリズムの説明のために、3点確認しておこう。

- A:文章末の文は後続文を持たず、先行する曖昧性を解消できる最後の文である。
 B:文章の曖昧性は語義的なものもあるが、推論に由来するものもある。推論を多く要する文章は曖昧性を持ちやすい。



C:資料として新聞記事を用いる。新聞記事は厳しい字数制限があるため、不要な表現は割愛され、必要最小限に近い形での表現がなされやすい。

A、B、Cの3点を総合して推測できることは、新聞記事において文章の長さが長くなればなるほど、文章末に使われやすい表現が存在しうる、ということである。つまり、語義的なものを含め、推論に由来する曖昧性は、文章が長くなればなるほど発生しやすくなり、文章末でそれらを解消する表現が使われやすくなるのが推測できる。新聞記事は不特定多数の読者に向けて書かれており、不要な表現が少ない。無駄に長い文章は寡少であり、伝えるべき内容に応じて簡潔に書かれているものがほとんどである。文章末でも同様に簡潔に書かれているとすると、文章を終わらせやすく、曖昧性を解消する力が強く、まとめに用いやすい単語や表現が使われる頻度は高くなるだろう。

対照例として3つの場合を挙げる。

- (1) 文の数が5個で文字数が250文字程度の文章
- (2) 文の数が10個で文字数が500文字強程度の文章
- (3) 文の数が20個で文字数が1000文字強程度の文章

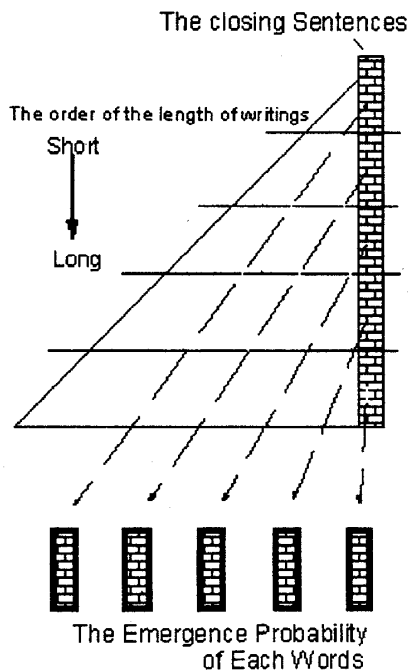
3つの例の違いは文の数と文字数の違いだけである。3つとも同じテーマの文章としよう。(1)において、先行する文の曖昧性を強力に解消するような単語が使われることは考えにくい。あり得ないものではないが、一般に出現頻度は低いだろう。では、(2)の文章ではどうだろうか？少なくとも、(1)のものよりは曖昧性が大きくなる可能性は高くなるだろう。さらに(3)の文章の長さになれば、複数の段落を持つものもかなり増えてくる。推論は文間だけでなく、段落間にも及び得るため、(2)で生じる以上の曖昧性が発生しうる。よって、文章末だけを比べた場合、先行する文の曖昧性を強力に解消するような単語の使用頻度は、相対的に見て(1)、(2)、(3)の順に高くなると推測される。上の推測を利用して、これらの単語を抽出する手続きが以下に述べる抽出アルゴリズムである。

文献[1][2][3]における抽出アルゴリズムとここで述べるアルゴリズム[4]にはいくつかの修正点が存在する。違いをはっきりさせるため、今回のアルゴリズムを ver.3 とナンバリングする。

大河ターム抽出プログラム version 3

- 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

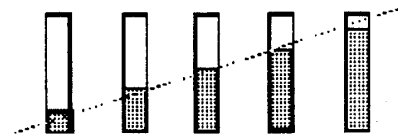


- 3.得られた文書グループ内の全ての記事を文字数順に並べる。
- 4.記事数が同数になるように X 個のブロックに分割する。平均文字数最小のブロックを第0ブロックとし、昇順に順序付けを行う。
- 5.文書群の全ての文書の末尾 k センテンスを取り出し、これらに対して、形態素解析を行う。
- 6.各単語毎にブロック別の集計を行う。
- 7.第 j 文書グループにおける第 x ブロック内の全ての文書の末尾 k センテンスにおける単語 n の出現回数を $F(j,n,x)$ とし、出現回数最大のブロックの出現回数を $F_{max}(j,n)$ とする。このとき単語 n は

$$F_{max}(j,n) \geq A(j)/X/P$$
 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。

上の手続きを経て抽出される単語の各を「第 j 文書グループの大河ターム」と呼ぶ。あるいは第 j グループのキーワードが例えば「2000年問題」だった場合、「グループ「2000年問題」の大河ターム」と呼ぶ。

記事の Filetering に関しては、文章の定義にも関わる問題であり、最適なものを同定することは難しい。今回用いた条件もヒューリスティックなものである。また、記事の筆者は既発表の同一テーマの文章を読んでいることが多いため、例え文章末に使いやすい単語があるとしても、乱用することは慎まれる。読者を飽きさせ、文章の質を落とすからである。このことから逆算して、文章末に使われやすい単語を取り出すにはかなりの数の記事が必要になる。この境界値を含め、 (X, k, t, P) の各のパラメータ値はヒューリスティックな判断をもとに定めた。



4 : 3つの分析視点

3章のアルゴリズムにおいて、最も考察すべき点は手続き8の頻度分布の直線近似(傾きが t 以上)である。先述したように「文章の長さが長くなればなるほど、文章末に使われやすい表現が存在しうる」点を考慮して上のアルゴリズムは立案されている。文章の長さをパラメータにする限り、最も単純な線形性をあてはめるのが妥当であるため、上記の単純回帰分析を用いることになる。しかし、文章の長さが長くなればなるほど相対的に頻度が上がることと、頻度分布が直線近似できることは直接は結びつかない。もし文章の長さが長くなればなるほど文章末での頻度が上がる単語があり、かつ文書グループを多く用意すれば、その内のいくつかはこのアルゴリズムを通過するだろう、という見込みのもとに抽出はなされる。その意味で、2章で述べた大河イメージと抽出結果の融合は分析の仕方に懸かっている。

分析の観点は大きく3つに分かれる。

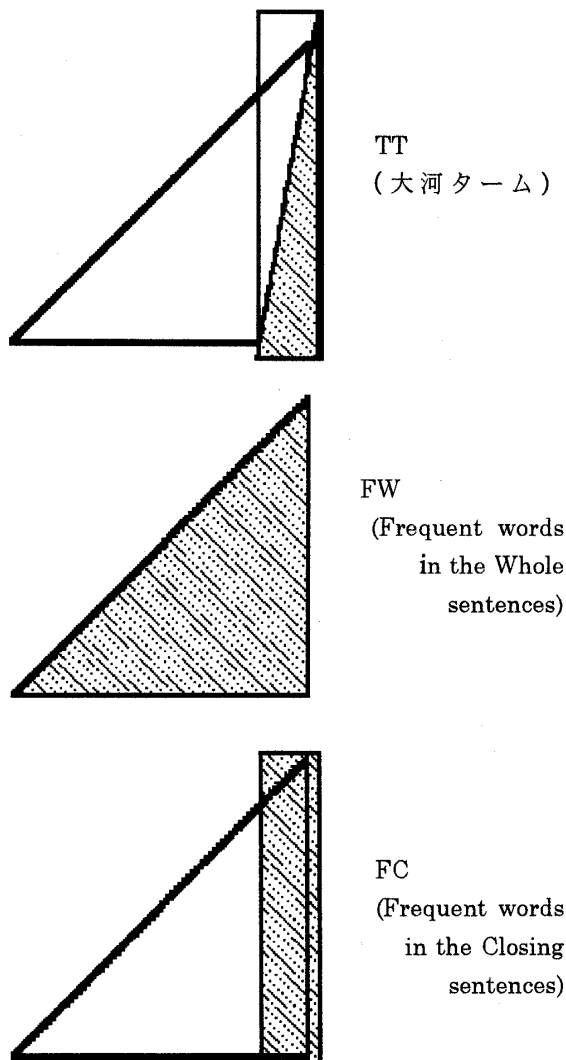
- (1) グループ別大河タームの抽出率、述べ大河ターム数とグループ平均の大河ターム数、全大河タームのグループ内頻出単語($FW\{w\}, FC\{w\}$)含有率
- (2) 最頻 NFW 大河ターム
- (3) シソーラスを用いた意味属性比較

文献[4]の報告に基づき、日経新聞 CD-ROM の97,98,99年度版[6,7,8]の3年分を資料とした(抽出アルゴリズムは version 3、パラメータは $(X, k, t, P) = (5, 2, 1.0, 10)$ 、形態素解析ツールは茶筌 ver.1.0[9]、単語は数詞を除く名詞に限定)結果を用いて各分析を解説する。

- (1) 3年分の CD-ROM から文書グループの数は、97年度で1112個、98年度で1149個、99年度で934個である。これらの文書グループから、グループ別大河ターム抽出率は97年度で99.2%、98年度で99.6%、99年度で99.3%が算出される。つまり、各年度ともほとんどの文書グループにおい

て大河タームが抽出されることがわかる。97年度のグループ平均の大河ターム数は97年度で7.11個、98年度で7.19個、99年度で6.95個が算出される。

比較のため、2つの指標 $FW\{w\}$ 、 $FC\{w\}$ を導入する*1。「第 j 文書グループの $FW\{w\}$ 」とは第 j 文書グループ内の全ての単語(ここでは数詞を除く全名詞)の上位 w 位までの頻出単語のことである。「第 j 文書グループの $FC\{w\}$ 」とは第 j 文書グループの全ての文書の末尾 k 文(ここでは $k=2$) だけから算出した頻出単語であり、同じく上位 w 位までのものを示す。模式的図を以下に表そう。



*1 頻出単語については文献[1][2][3]での表記を変更している。変更点は(旧)TF → (新)FW、(旧)PF → (新)FCとしている。本研究では今後も新表記を用いていくことにする。

	97年度	98年度	99年度
Intersection of TT and FW	1562(19.9%)	1597(19.4%)	1218(18.9%)
Intersection of TT and FC	1592(20.3%)	1605(19.5%)	1231(19.1%)
Intersection of TT,FW and FC	1245(15.8%)	1264(15.4%)	960(14.9%)

表1：大河タームとFW、FCとのIntersection

	影響	業界	資金	通信
FW{7}とするグループ数	0	2	23	8
TTとする他グループ数	4	30	32	6

表2：99年度グループ「2000年問題」の大河タームの他テーマでの使用状況

順位	97年度		98年度		99年度	
	TT	グループ数	TT	グループ数	TT	グループ数
1	今後	220	今後	183	今後	182
2	可能性	82	可能性	101	可能性	106
3	各社	38	回復	30	各社	22
4	声	20	収益	24	見方	18
5	導入	19	各社	23	国内	17
6	検討	18	特捜	20	声	16
7	見方	18	見方	17	考え	16
8	特捜	18	声	16	期待	13

表3：NFW大河タームの年度別重複グループ数（上位8位まで）

大河タームの中にはそれが抽出されたグループ内の頻出単語であったり、文章末だけで算出した頻出単語である可能性がある。大河タームのグループ別平均個数が3年分を通して約7個であるので、比較の指標としてFW{7}、FC{7}を用いる。大河タームが各グループのFW{7}、及びFC{7}に含まれる割合を求めた結果が上の表1である。

- (2) 同じグループのFW{w}に入らない大河タームの特徴を捉える。ある大河タームはその抽出されたグループ以外の文書グループからでも大河タームとして抽出されることがあり得る。また、抽出されたグループのFW{w}に入らなくても、他のグループのFW{w}に入ることがあり得る。この状況を理解しやすくするため、99年度グループ「2000年問題」でのFW{7}でない4つの大河ターム「影響」「業界」「資金」「通信」について表2において説明しよう。

表2の数字は上段は4つの大河タームの各をFW{7}とする99年度の他グループ、下

段は各を大河タームとする99年度の他グループの数を表している。つまり、「影響」のように99年度における他のグループで頻出単語にならないもの、「業界」「資金」「通信」のように他グループで頻出単語になるものがある。この「影響」のように全てのグループにおいてFW{7}に入らない大河タームをNFW大河タームと呼ぶ。大河タームは他のグループでも大河タームになり得るため、重複するグループ数によって、使用頻度を比べることができる。NFW大河タームも同様である。表3はNFW大河タームの重複グループ数が多いものを表している。

表3の上位を見れば明らかのように、3年分で1位「今後」と2位「可能性」は定常的であり、3位以下との重複グループ数に大きな違いがある。3位以下では多少の変動はあるが同じような単語が使われている。NFW大河タームには3年間で一定の傾向があることがわかる。

- (3) 全大河タームと全グループのFW{7}と全グループのFC{w}の全ての単語について

シソーラス（日本語ならば例えば日本語彙大系の意味属性体系 [10]）上の意味属性を調べ、属性の使用頻度を比較する [2][3]。また、大河ターム内比較として、グループ内で FW{7}、FC{7}に入る大河タームの全てと NFW 大河タームについて意味属性を調べ、属性の使用頻度を比較する [4]。これについては特徴ある意味属性を抽出ができるが精度の問題がある。よって、これについては稿を改めたい。

5. 考察

(1)の分析において、まず重要なことは、ほとんどの文書グループには大河タームが存在することである。これが3年分とも平均約7個ずつ存在している。ただし、これらは文字数をパラメータとした文章末での頻度の増加傾向だけから抽出されているため、例えば、

- a: そのグループのテーマの文章にとって頻繁に使われるべき単語が存在し、その出現回数がチャンスレベルで文章末で増加する頻度分布になった、
- b: そのグループのテーマの文章にとって、文章の長さに関係なく、文章を終わらせるのに使われる単語が存在し、その出現回数がチャンスレベルで文章の長さに応じて増加する頻度分布になった、

場合も考えられる。よって、上記の場合に当てはまる可能性のある大河タームを取り出すため、全大河タームが同一グループの FW{7}、FC{7}に含まれる割合を求めた。両者とも3年分を通じて20%程度であり、さらにFW{7}、FC{7}の両方に含まれる大河タームの割合が15%程度であるならば、全大河タームの内、約75%がFW{7}でもFC{7}でもない、つまり、そのテーマの頻出単語でもなく、長さに関係なく文章末に使われやすい単語でもないような大河タームである、といえるだろう。特に大河タームとして抽出される単語の中で、短い文章の文章末でも使われるような単語は文章末での頻度そのものが高くなり、FC{7}に入ってしまう。よって、大河タームであり、かつFC{7}でないことは短い文章の文章末でほとんど使われておらず、長い文章の文章末でよく使われていることを

示している。このことはこれらが「大河」イメージにより近いものであることを示す。

(2)の分析において、例えば、上述の99年度グループ「2000年問題」における「業界」「資金」「通信」などは、「2000年問題」を語り続ける上で、おそらく非常に重要な単語である。しかし、「業界」「資金」「通信」は、「2000年問題」以外の多くのテーマでも重要な単語である。それらは他の記事グループにおいては文章末以外でも積極的に使用されうるし、あるいはその単語自体があるグループのテーマになることさえある。このことは少なくとも2つのことを示唆する。

- 1: 「2000年問題」での「業界」「資金」「通信」などが文章末での使用頻度が高いことは、そのグループのテーマそのものに依存する。
- 2: 「業界」「資金」「通信」は、他グループの多くの記事で語られる内容を意味的に含む。

逆に、NFW大河タームはこの性質をもたないと言える。「2000年問題」の「影響」は上の表と定義からしてNFW大河タームである。直感的に考えても、「影響」という単語が「2000年問題」というテーマだからこそ使用されたとは考えにくいし、「影響」という単語をもちいる時に、他の文章で使用された意味や前後の文脈を考慮する必要はない。つまり、使用される文章のテーマに直接依存せず、他の文章でなされる使用状況にも依存せず、かつ長い文章の文章末において多く使用される単語である、ということが導ける。

6: 「大河」イメージの分節とまとめ

3年分の日経新聞 CD-ROM を通じて定期的に得られる大河タームについて、大きく分けると、少なくとも3つに分類できる。1つ目は同一グループ内の FW{w}、FC{w}に含まれるものである。2つ目は同一グループ内では FW{w}、FC{w}に入らないが、他のグループの FW{w}に入るもの、3つ目は NFW 大河タームである。

- 1: 1つ目は2章で述べた大河イメージと違うものになるが、文章末に使われる頻出単語であることは考察の対象とされるべきであり、今後の検討課題の一つとする。

2: 2つ目は他のテーマで頻出する重要な単語でありながら、あるテーマの文章末に用いられやすい単語である。この単語は他のテーマの文章で語られ続けており、その意味内容を不可避免的に含んでいる。よって、これらの単語が使われる形で終わる文章は、別のテーマの内容を含んでしまう反面、当該テーマだけでは語れなかった豊かな意味を最後に文章に持ち込んでから終わることになる。また、あるグループの大河タームが別グループの文章の $FW\{w\}$ になるならば、そのグループのテーマとの内容上の関連がつくことが予想される。この大河タームを用いることで、グループのテーマ間の関係づけをもたらすことも予想される。

3: 3つ目は他のテーマの頻出単語にならないにも関わらず、文章の長さが長くなればなるほど使用頻度が増す大河タームである。これは当該テーマの内容にも他のテーマの内容にも依存せずに使われていると見なせるため、逆に任意のテーマの文章(ここでは新聞記事)の終わり方に頻繁に使われやすい単語であるといえる。

考察すべきは2と3の違いである。2章で述べた大河イメージは少なくとも2つに分節されねばならない。1つには文章のテーマに依存せず、一般の文章(ここでは新聞記事)の文章末においてまとめに使われやすいもの、もう一つは文章のテーマに依存しながら、他の文章の内容を含意するものである。

本研究の特徴は、新しいメディアとツールを用いることもさることながら、あるテーマの文章末を分析するための材料として同一年度の他テーマの文章の頻出単語を用いる点にある。これは意味論的、あるいは語用論的にも重要な視点であると考えられる。また、記事の筆者が当該テーマの他の文章だけでなく、他テーマの関連する文章も読んだ上で記事を執筆していることが予想されることから、さらには想定されている読者の多くも当該テーマの記事と関連する記事のいくつかを読んでいるのが予想されることから、この方法論は擁護されるだろう。

本研究の展望は2通りである。1つには、様

々な他言語の新聞記事について同様の抽出法を用いてNFW大河ターム、あるいはその意味属性を比較することである。これは、言語に依存しない一般性が新聞記事の文章にあるかどうかを調査すると同時に、各言語の特徴を再発見する契機を提供しうる。もう一つは大河タームと文章の接続構造との関係である。大河タームは、文脈上いかなる機能を果たしているのか、統括概念との関係はつくのか、を評価実験を交えながら考察していくことである。これらは、特徴ある単語や意味属性の絞り込みのための論理を提供する可能性がある。その結果は将来的には言語処理のための参照情報となることも期待される。

参考文献

- [1] 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
- [2] 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
- [3] Takashi Nakamura, Tatsuo Hemmi and Asumi Koizumi: Semantic Features of specific words in the closing sentences of newspaper articles, Proceedings of The second Annual Conference of The Japanese Society of Language Sciences, 2000
- [4] 中村隆志、本間愛：日本語新聞記事の文章末における特異的単語の意味属性 ～経時比較～(準備中)
- [5] 加藤典洋：言語表現法講義、岩波書店、1996.
- [6] 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.
- [7] 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、1999.
- [8] 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、2000.
- [9] 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『茶釜』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.
- [10] 池原悟、他編：日本語語彙大系、岩波書店、1997