状態空間表現を用いた文章の特徴付け

状態空間表現を用いた文章の特徴付け

馬場康維（統計数理研究所）
小森　理（統計数理研究所）

Feature Extraction of Sentence Structure based on State Space Representation Model

Yasumasa Baba (The Institute of Statistical Mathematics)
Osamu Komori (The Institute of Statistical Mathematics)

１．はじめに

様々な目的で文章をデータとした解析が行われている。文章の類型化、著者推定、ウエブ上の文章からのトピックの抽出など分野も目的も用いられる手法も様々である。たとえば、著者推定の問題では、品詞の出現比率や特定の単語の出現比率がしばしば用いられる。この方法は品詞や特定の単語の静的な分布を用いて文章の特徴を抽出することによる解析である。ところで、文章は文の連なりから成っており、文章の要素である文は一連の語や記号の系列で成り立っている。即ち、文や文章は形態素の系列で成り立っている。したがって文章の特徴を用いて何らかの解析をするには、単に形態素の静的な分布のみならず形態素の出現順序を考慮した動的な解析が有効であろう。そこで、一連の語や記号の連なりである文章を“品詞”という状態を推移する系列とみなし、文章の構造をこの系列の構造としてとらえることで文章構造の解析ができるのではないかと考えたのがこの研究の発端である。

“状態”の定義は分析の対象・目的によって変わる。形態素解析を利用してテキストデータを品詞の系列で表現する場合には、名詞、動詞、助詞などの品詞が状態に対応する。文の構造を抽出する際には名詞句、動詞句といった品詞の結合した状態を用いた方が形態素のままの状態で文を表現するよりは文の構造が把握しやすく、構造の分析には適している。一方、より詳細な構造を分析の対象にするならば、名詞を名詞の種別に分割した状態を考えるというように状態の分割も必要である。さらに文章全体を構造化してとらえるには段落の状態を考慮する必要がある。このように“状態”は分析の場面、場面に応じてフレキシブルに定義されるものである。

この報告では、文章構造のモデル化の基礎的な研究として文を状態空間で表現し時系列としてとらえる試みについて述べる。ここで用いたデータは国立国語研究所共同研究プロジェクト「文章における語彙の分布と文章構造」により作成されたテキストデータの一部である。文章あるいは文の解析にはまず文法的なモデルを用意し単語の意味を考慮するというような方法があるが、ここでは、データから得られる情報をもとに文の構造的な把握をするというプロセスによって文章構造のモデル化を図る。具体例として上記の名詞句，動詞句等の推移確率を計算し，それにもとづいた主成分分析の結果を紹介する。これは名詞，動詞，助詞等の頻度による解析，つまり静的な解析とは異なり，文章のつながりを考慮している点で，既存の解析法とは異なっている。また文章の特徴抽出をより細密化する試みとして，名詞句，動詞句等をさらにいくつかの「パター

ン」にまとめることができる例も紹介する。このパターンの抽出は文章を一つ一つ読んで探していく作業であるためかなりの労力を費やすが，基本的なパターンはそれほど多くはないと予想されるため，一旦辞書のようなデータベースを構築できれば，さまざまな分野の作品またその著者たちの特徴をより鮮明に捉えることができると思われる。

ここで述べるモデル化の方法はまだ完成されたものではないが、機械学習を利用したデータの自動的な収集、文章の類型化による文章の分類等様々な応用が考えられる。

２．品詞による表現－形態素解析の利用

最も基礎的で素朴な品詞状態による表現の例を示す。ここで例示に用いたデータは、近藤和敬、“ヒルベルトの数学における公理的方法からカヴァイエスの概念の哲学へ”（以下、近藤論文と呼ぶ)をテキスト化し形態素解析を行って得られた品詞データである。テキストデータには、段落のタグがついており、“論文のタイトル＋著者名＋所属”は一つの段落として扱われている。表１は形態素解析の結果を示している。最も基礎的なこのタイプのデータを時系列的に表現しただけでも文の特徴が見いだせる。形態素解析の品詞のカテゴリーが異なった状態になるように（句読点０，名詞句１０，動詞句２０，形容詞句３０，副詞４０，連体詞５０，接続詞６０，その他-１０）数値を対応させた。この数値は便宜上割り振ったもので何らかの最適化をしたものではない。この数値を割り振られた品詞の状態空間を用いて文章の一部を表現してみると図１、図２のようになる。図１は、近藤論文の“タイトル＋著者名＋所属部分”である。図２は最初の文の時系列である。図１には句読点がないこと、動詞＋句点で終わっていないこと等、タイトルであることが類推できる特徴が存在する。一方、図２では文末が動詞＋句点という典型的な連結で終わっている。このことからも、状態空間表示により時系列を表現することで、文の特徴抽出が可能になることが推察される。

表１　文の形態素解析

文字	品詞	品詞（詳細）	文節ID
数学	名詞	一般	1
基礎	名詞	一般	1
論	名詞	接尾	1
の	助詞	連体化	1
論争	名詞	サ変接続	2
の	助詞	連体化	2
結果	名詞	副詞可能	3
、	記号	読点	0
数学	名詞	一般	1
的	名詞	接尾	1
認識	名詞	サ変接続	1
の	助詞	連体化	1
確実	名詞	形容動詞語幹	2
性	名詞	接尾	2
の	助詞	連体化	2
アプリオリ	名詞	一般	3
な	助動詞	*	3
基礎	名詞	一般	4
付け	名詞	接尾	4
が	助詞	格助詞	4
不可能	名詞	形容動詞語幹	5
で	助動詞	*	5
ある	助動詞	*	5
こと	名詞	非自立	6
から	助詞	格助詞	6
、	記号	読点	0
合理	名詞	一般	1
論	名詞	接尾	1
的	名詞	接尾	1
な	助動詞	*	1
認識	名詞	サ変接続	2
論	名詞	接尾	2
は	助詞	係助詞	2
その	連体詞	*	3
説得	名詞	サ変接続	4
力	名詞	接尾	4
を	助詞	格助詞	4
半減	名詞	サ変接続	5
さ	動詞	自立	5
せ	動詞	接尾	5
た	助動詞	*	5
よう	名詞	非自立	5
に	助詞	副詞化	5
思わ	動詞	自立	6
れる	動詞	接尾	6
。	記号	句点	-1

図１　タイトルの品詞による時系列表現

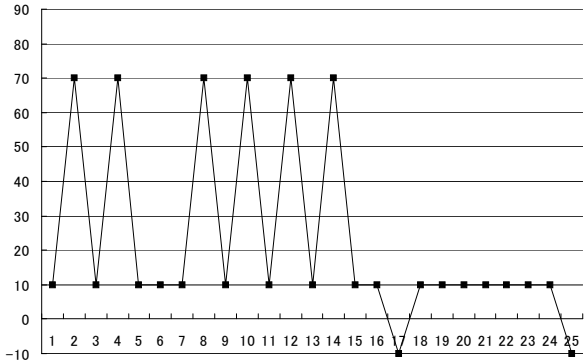
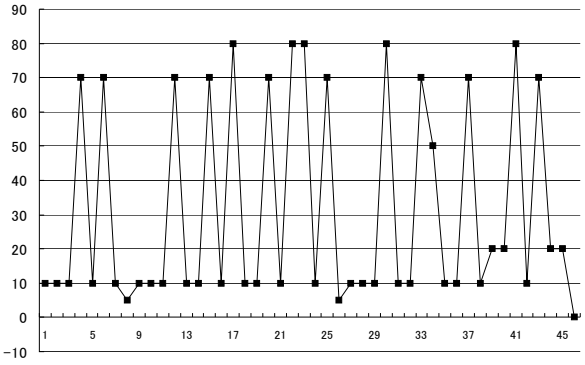


図２　文の品詞による時系列表現



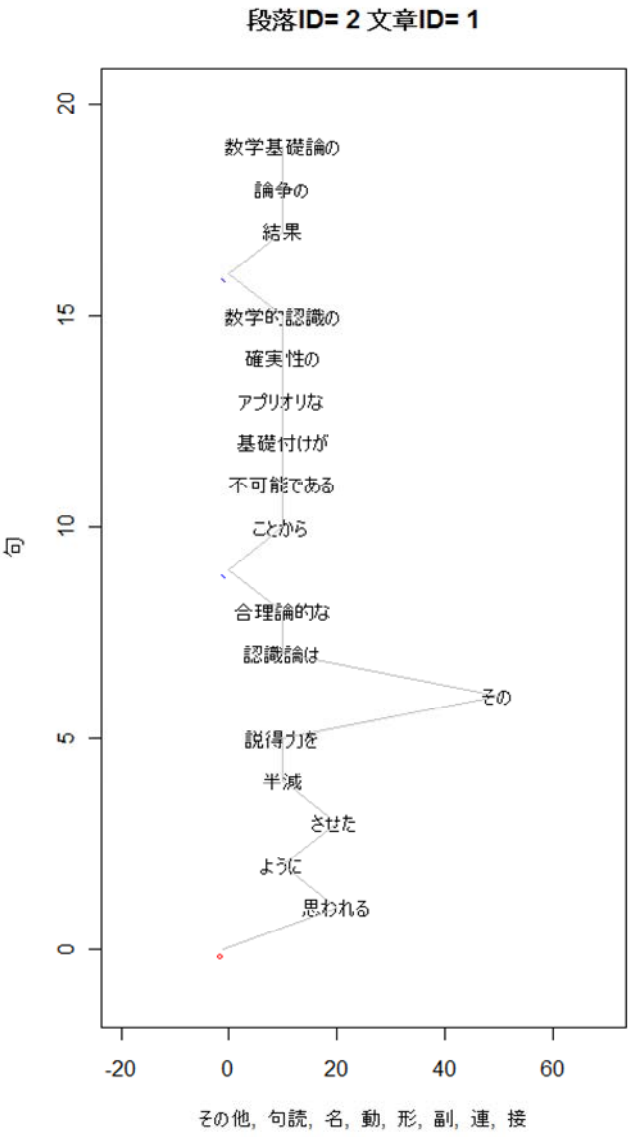
３．状態の縮約

一連の文章に形態素解析を行うと、例えば、“数学基礎論”は“数学”、“基礎”、“論”というように３つの名詞に分解され、名詞状態が連続して出現する。上述のグラフ表現では“名詞”状態が３時点続くことになり、時系列を観察すると、状態“名詞”のフラットな直線が現れる。品詞の状態による表現空間での時系列を観測すると、数学＋基礎＋論のように名詞のみが連続する場合はそれらを一つの単語（便宜上名詞と呼ぶ）として扱うことが可能な場合がほとんどであることが分かる。これを踏まえ、第１ステップとして、連続した“名詞”状態を一つの“名詞”状態として縮約を行った。

さらに、例えば、“数学基礎論＋の”が次の“論争”を修飾している。名詞＋助詞で一つのかたまりと考える方が文の構造の表現には便利である。即ち、品詞による状態表現を特定の結びつきを示す品詞と品詞の状態に縮約した方が構造解析には都合がよい。そこで、品詞で表現した状態を縮約し助詞を中心にまとめた状態で文を表現したものが図３である。

このように助詞を中心にして状態をまとめ

図３　助詞の違う要素表現（句）による文の時系列表



てみると助詞の種類によりそれぞれの役割があることが分かる。そこで、名詞＋助詞を一つの状態とみなし、近藤論文のデータから推移確率を作ったものが表４である。表中、“名詞の”や“名詞を”はそれぞれ、名詞＋助詞（の）や名詞＋助詞（を）を表している。つまり名詞と助詞の連結ごとに状態を割り振ったことになる。これが第２段階の状態空間の縮約である。表５は表４の項目に関する出現頻度をまとめたものである.

表４　助詞による状態表現の推移確率(%)（近藤論文）

	句点	名詞を	名詞に	名詞が	名詞な	名詞の	名詞は	動詞	読点	その他
句点	0	3.6	3.6	8.0	4.4	11.1	12.9	0	0	55.7
名詞を	0	0	10.1	0.5	3.5	3.5	0.5	41.4	7.1	33.1
名詞に	0	1.6	3.2	1.6	6.4	4.3	0	41.7	7.5	33.2
名詞が	0	4.0	9.0	0	4.5	5.6	0	26.0	6.8	44.8
名詞な	0	20.7	3.7	9.8	2.4	18.3	7.3	0	1.8	35.5
名詞の	0	13.6	7.1	11.3	7.3	18.6	6.8	0.5	0	35.5
名詞は	0	7.0	4.2	0.7	5.6	9.9	0.7	3.5	47.2	21.0
動詞	26.2	5.5	9.4	7.5	2.8	9.7	7.2	0	7.2	25.1
読点	0	7.2	6.3	6.6	9.6	26.2	3.9	1.5	0	38.4

表５　句等の出現頻度（近藤論文）

	度数	パーセント
句点	226	6.9
名詞を	198	6.1
名詞に	187	5.7
名詞が	177	5.4
名詞な	164	5.0
名詞の	382	11.7
名詞は	142	4.3
動詞	362	11.1
読点	332	10.2
その他	1096	31.4

さらに“名詞の”＋“名詞”は合わせて“名詞”とみなしても良い状態である。名詞についてはこれが第２段階の縮約になる。この外に、すぐに目につく状態の縮約には、“名詞”＋“する”がある。この状態は、その機能上から、“名詞”と“する”を合わせた状態を“動詞”状態とみなすことができる。これらのプロセスを簡単に示すと下記のようになる。

名詞＋名詞　⇒　名詞
名詞＋助詞（の）⇒　名詞（の）
名詞（の）＋名詞　⇒　名詞
名詞＋する　⇒　動詞

このように、次々に状態をまとめていくことにより、階層構造を持った状態空間が構成できるこ

とになる。

４．論文の特徴の比較

比較のために他の論文データを用いて集計を行った。比較に用いたデータは、横地徳広“認識論的転回の地平を求めてーハイデガーとカント『純粋理性批判』”（以下、横地論文と呼ぶ)、である。推移確率を表６に、各句等の状態の出現頻度を表７に示してある。この推移確率の状態（項目）も表４と同じものを用いている。

表６　助詞による状態表現の推移確率(%)（横地論文）

	句点	名 詞を	名詞に	名詞が	名詞な	名詞の	名詞は	動詞	読点	その他
句点	0	6.1	3.0	0.8	2.3	5.3	23.5	0	0	59.5
名詞を	0	0	13.3	0.6	0	1.9	3.2	32.9	0.6	47.1
名詞に	0	3.4	1.7	3.4	0.8	1.7	0.8	35.6	7.6	44.3
名詞が	0	6.4	7.7	0	0	9.0	0	23.1	5.1	48.9
名詞な	0	6.1	3.0	9.1	3.0	21.2	12.1	0	0	45.3
名詞の	0	22.0	11.0	10.4	4.4	4.4	6.6	11.0	0	39.0
名詞は	0	8.6	5.2	3.4	1.7	15.5	0	1.7	31.9	32.4
動詞	28.8	8.0	8.4	1.8	2.2	6.6	8.4	0	10.2	25.0
読点	0	8.2	6.5	8.2	2.2	22.9	9.5	0.4	0	41.7

表７　句等の出現頻度（横地論文）

	度数	パーセント
句点	133	6.6
名詞を	158	7.8
名詞に	118	5.8
名詞が	78	3.9
名詞な	33	1.6
名詞の	182	9.0
名詞は	116	5.7
動詞	226	11.2
読点	231	11.4
その他	749	33.8

表４、表５、表６、表７から、２つの論文の表現の比較が可能になる。いずれの場合でも、動詞の次に続くのは句点である確率が高く、句点の次には文章のはじまりである名詞が続く確率が高いというような傾向があることが分かる。これらの中で主語の役割を担うものは主に“名詞は”と“名詞が”であるが，句点からの推移確率が２つの論文で大きく異なることも分かる。文章のスタイルの違いが推移確率に反映されていると言える。ここでは，２つの論文のみを比較したが、多くの論文について推移確率の比較あるいは頻度の比較を行うことにより文献間の距離が算出で

きる。したがってそこから文献の類型化ができるであろう。

５．主成分分析の適用例

前節で二つの論文の比較をした。もう少し論文を増やして論文間の比較をしてみる。ここでは上記の二つに以下の３論文を追加して分析を試みた。手塚博「ミシェル・フーコーの権力分析における真理の概念‥ 権力の行使としての反省」(手塚論文)，小島優子「ヘーゲルにおける「罪責」と「犯罪」―『精神現象学』を中心に」(小島論文)，山田圭一「最晩年ウィトゲンシュタインの連続性テーゼが意味するもの」(山田論文)。

まずそれぞれの論文で前節の表５のように句等の出現頻度を計算し，少なくとも一つの論文中に５パーセント以上出現する句等を選び出した。横地論文では「名詞が」が 3.9 パーセントであるが，近藤論文では 5.4 パーセントであるため，推移確率の項目の一つに選ばれている。 このようにして抽出した状態をベースにして、状態の頻度の分布と推移確率を求めたものが表８，表９，表１０，表１１，表１２，表１３である。それぞれの特徴が読み取れる。

表８　助詞による状態表現の推移確率(%)（手塚論文）

	句点	名 詞を	名詞に	名詞が	名詞な	名詞の	名詞は	動詞	読点	その他
句点	0	2.9	3.4	5.8	2.4	12.5	13.5	0	0	59.9
名詞を	0	0	12.0	0.5	0	0.5	0.5	39.3	2.6	44
名詞に	0	4.0	1.1	0	2.9	1.1	0	44.8	10.3	35.7
名詞が	0	4.6	13.2	0	2.6	10.5	0	15.8	11.8	42
名詞な	0	15.3	0.9	10.8	2.7	10.8	3.6	0	1.8	54
名詞の	0	16.5	8.3	8.9	6.7	10.2	7.9	0.6	0.3	40.5
名詞は	0	3.1	5.5	2.3	4.7	8.6	0	2.3	42.2	31.7
動詞	21.7	4.5	6.1	8.0	2.6	6.7	3.5	0	7.3	39.2
読点	0	7.2	4.1	7.8	3.8	23.5	6.6	0.3	0	46.1

表９　句等の出現頻度（手塚論文）

	度数	パーセント
句点	209	6.6
名詞を	191	6.1
名詞に	174	5.5
名詞が	152	4.8
名詞な	111	3.5
名詞の	315	10.0
名詞は	128	4.1
動詞	313	10.0
読点	319	10.1
その他	1232	36.8

表１０　助詞による状態表現の推移確率(%)（小島論文）

	句点	名詞を	名詞に	名詞が	名詞な	名詞の	名詞は	動詞	読点	その他
句点	0	2.6	15.1	5.7	1.0	8.3	12.5	0	0	54.6
名詞を	0	0	5.6	2.2	0	5.6	2.2	43.0	6.1	35.7
名詞に	0	9.2	1.5	0	1.0	2.4	1.5	40.8	18.0	25.8
名詞が	0	6.1	11.4	0	0.8	12.1	0	21.2	6.1	43.1
名詞な	0	18.6	9.3	7.0	0	7.0	7.0	0	0	51.2
名詞の	0	11.8	11.8	6.7	2.0	10.2	7.1	1.6	0	49.2
名詞は	0	6.0	6.0	0.6	0	6.0	0	9.6	46.4	25.2
動詞	27.2	5.7	6.5	3.4	0.8	5.1	9.6	0	4.2	37.5
読点	0	7.3	8.1	10.2	3.1	20.9	9.2	2.9	0	38.5

表１１　句等の出現頻度（小島論文）

	度数	パーセント
句点	193	6.5
名詞を	179	6.0
名詞に	206	6.9
名詞が	132	4.4
名詞な	43	1.4
名詞の	254	8.6
名詞は	166	5.6
動詞	353	11.9
読点	382	12.9
その他	1060	34.4

表１２　助詞による状態表現の推移確率(%)（山田論文）

	句点	名詞を	名詞に	名詞が	名詞な	名詞の	名詞は	動詞	読点	その他
句点	0	0.5	7.9	2.6	2.1	11.5	6.8	0	0	68.2
名詞を	0	0	9.3	0.5	0.5	8.2	1.6	27.5	1.1	50.3
名詞に	0	1.9	1.3	1.3	1.3	4.4	2.5	32.1	6.9	48.2
名詞が	0	2.0	6.5	0	5.2	13.7	0	14.4	3.9	54.9
名詞な	0	13.6	5.8	10.7	7.8	20.4	1.9	0	0	40.1
名詞の	0	16.9	7.1	10.3	3.4	14.9	8.1	1.0	0	37.7
名詞は	0	3.9	0.6	2.6	0.6	12.9	0	4.5	27.7	46.2
動詞	26.6	2.7	4.4	5.1	4.8	9.9	6.1	0	9.6	30.0
読点	0	6.2	4.7	8.0	4.7	27.9	9.8	0	0	39.5

表１３　句等の出現頻度（山田論文）

	度数	パーセント
句点	192	5.9
名詞を	182	5.6
名詞に	159	4.9
名詞が	153	4.7
名詞な	103	3.2
名詞の	409	12.6
名詞は	155	4.8
動詞	293	9.0
読点	276	8.5
その他	1335	39.9

この５つの論文の状態空間での推移確率をデータとして主成分分析を行った。まず９×１０の推移確率の行列を行ごとにベクトル化し要素数９０のベクトルとした。そのベクトルに対し主成分分析をし，第１主成分と第２主成分を表示すると図４の散布図が得られる。（山田論文，横地論文），（手塚論文，小島論文）と（近藤論文）に大きく３つに分かれることが分かる。第１主成分の寄与率は0.51, 第２主成分の寄与率は0.21 となり第３主成分までの累積寄与率は0.88 となる。第１主成分、第２主成分、第３主成分の因子負荷量を計算し，図示したものが図５，図６，図７である。なお因子負荷量の図では、

句点⇒句，読点⇒読，名詞が⇒名が，名詞の⇒名の

といった省略表記をしている。たとえば、状態間の推移を表す「句点名詞の」は「句名の」というように省略した表現が用いられている。

図４　推移確率から求めた５つの論文の特徴づけ

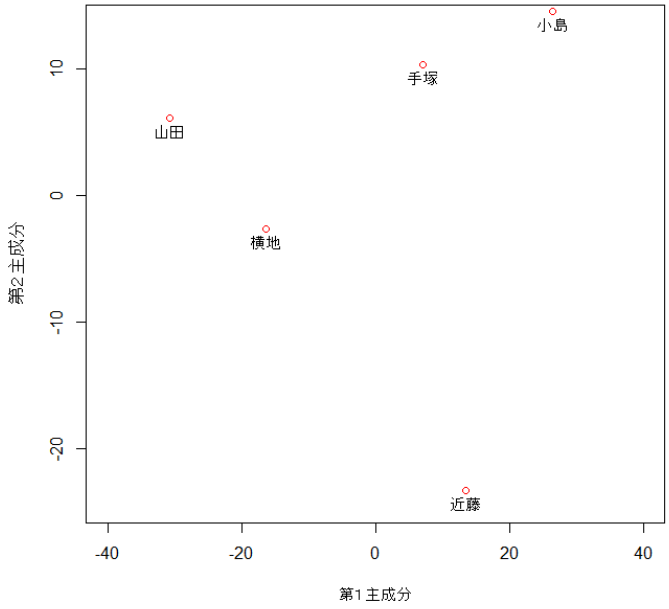


図５　第１主成分に対する因子負荷量

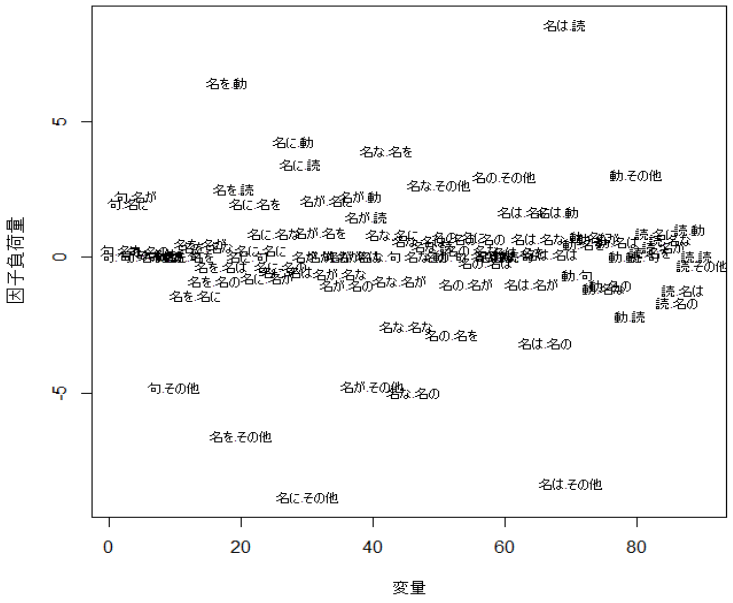


図６　第２主成分に対する因子負荷量

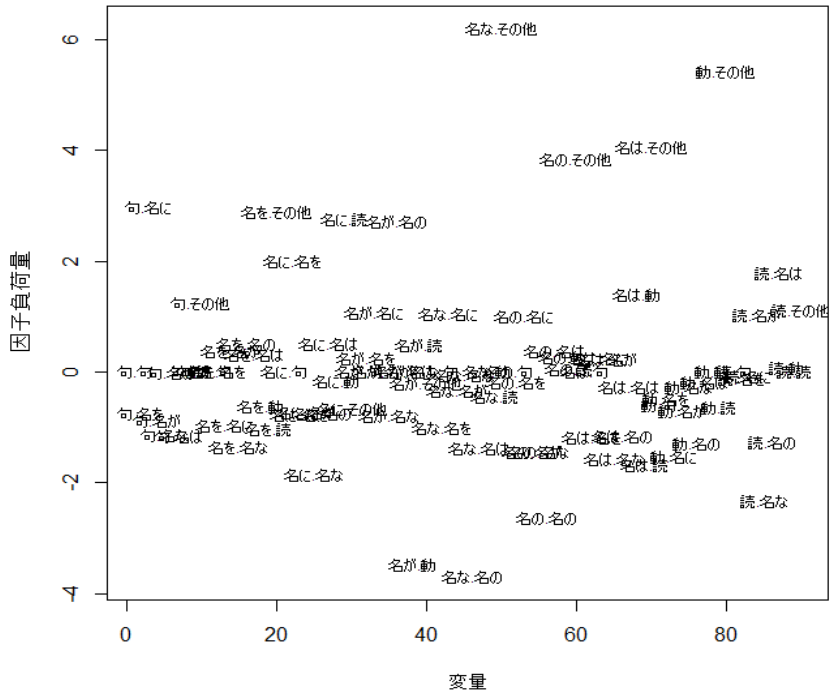
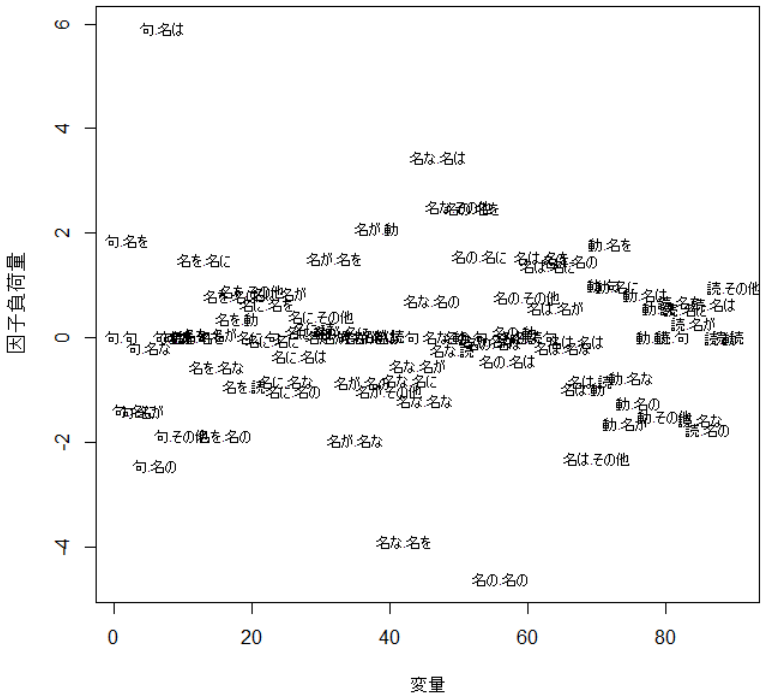


図７　第３主成分に対する因子負荷量

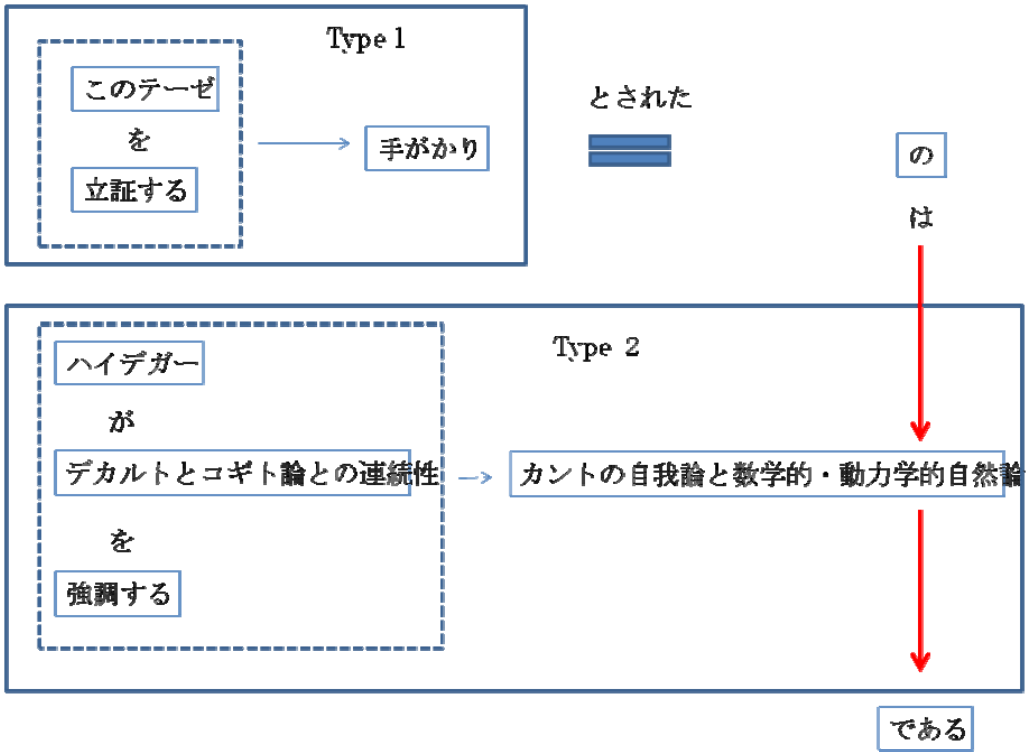


第１主成分に大きく寄与する推移確率の因子として「名詞は→読点」,「名詞を→動詞」,「名詞に→その他の品詞」,「名詞は→その他の品詞」のようなものがあり，これは文の基本的な構造を規定する因子と考えられる。第２主成分に大きく寄与する因子としては「名詞な→その他」,「動詞→その他」,「名詞が→動詞」,「名詞な→名詞の」などの修飾関係に関する因子が見られる。近藤論文はその他の４つの論文と比較すると，第二主成分に大きな違いがあるため，修飾関係にその違いがあると推察される。このように文章を状態空間とみなし，それを推移確率で表現することにより, 単なる品詞の頻度の集計からでは読み取れない動的な文章の特徴を捉える事ができる。

６．より詳細な文章の特徴づけ

上記までの議論では品詞同士の結合（「名詞＋名詞」）と品詞と助詞との結合（「名詞＋は」）に注目し，文構造の探索を行った。しかしながら文の特徴をより詳しく捉えようとするならば，さらに上位の文構造まで考慮する必要がある。一例として横地論文の次の一文を考察する。「このテーゼを立証する手がかりとされたのは、ハイデガーがデカルトのコギト論との連続性を強調するカントの自我論と数学的・動力学的自然論である」。一見複雑な構造をもつ文章であるが，文の構造に注目すると図８のようになる。

図８　より詳細な文構造



この文章の骨格は矢印で示された「AはBである」という至って単純な構造である。この文章を複雑に見せているものは名詞の修飾関係である。これも大きくわけて Type1 と Type2 があることが分かる。Type1 では「手がかり」という名詞を「目的語（このテーゼ）＋を＋動詞（立証する)」という語が修飾している。Type2 も同様な構造であるが「ハイデガー」という主語が新たに加わった修飾語になっている。このような文の「パターン」に注目し文章または作品全体の特徴づけを行うことにより，より詳細な文構造の分類が可能になる。

７．おわりに

品詞や句による状態空間表現について状態の縮約のプロセスを示した。文は語がつながった一連の系列から成っておりいわば時系列的な表現が必要である。この観点からすると状態空間表現は系列事象としての文あるいは文章を表現するのに適した表現であると考えられる。

これまでの文章の構文解析は，単語の出現割合に注目したものが多かった。これはいわゆる静的な文章解析である。静的な文章解析では、語の連なりは無視される。語や品詞の出現頻度のみが用いられており、文章のなかの語の出現の順序という情報、即ち時系列的な情報は用いられていない。時系列的な表現という立場からすると静的な解析は時間軸に沿った情報をまとめて状態

空間を占める割合だけに注目しているといえる。時間と状態という二つの情報のうち時間についてまとめて状態空間上だけの分布をみているというのが静的な解析である。

一方ここで提案した状態空間による表現は語の連なりといういわば文章の動的な要素を考慮したものとなっており、図3または推移確率に示したような文章の動的な表現を分析の対象としたものである。この状態空間表現による分析は、状態をどう定義するかによって様々な対象の分析に応用できる。名文と呼ばれる文章には読者に訴えるリズムがあり，これが文章の内容理解を深める。名文と駄文の違いは状態空間でどう表現されるのか。理路整然とした文章と理解しがたい文章は状態空間表現でどう違うのかなど、様々な解析に応用できる。この動的要素が色濃く出るものが詩や音楽の世界の歌詞である。このような対象にも状態空間表現が応用できる可能性がある。

今回の試みはまだ試行錯誤の段階であり、状態空間の構成も定まったものではない。今後大量のテキストデータの分析を積み上げることによって、文章の背後にある時系列的な状態の推移の様々なパターンの把握と類型化を試みたい。また大量データの処理にあたり，機械学習に適した構造モデルを構築することも考えている。

参考文献

小島優子（2007）「ヘーゲルにおける「罪責」と「犯罪」―『精神現象学』を中心に」哲学, Vol.58, pp.177-190.

近藤和敬（2009）「ヒルベルトの数学における公理的方法からカヴァイエスの概念の哲学へ」 哲学, Vol.60, pp.169-184.

田中章夫（1974）「句のエントロピーに基づく構文合成」言葉の研究第5集, pp.125-146.

手塚博（2009）「ミシェル・フーコーの権力分析における真理の概念-- 権力の行使としての反省」哲学, Vol.60, pp.217-232.

中野洋（1974）「自動項分解析の構想」言葉の研究第5集, pp.147-157.

町田健（2011）「言語構造基礎論：文の意味と構造」勁草書房.

山田圭一（2008）「最晩年ウィトゲンシュタインの連続性テーゼが意味するもの」哲学, Vol.59, pp.309-325.

横地徳広（2005）「認識論的転回の地平を求めて－ハイデガーとカント『純粋理性批判』」哲学, Vol.56, pp.270-282.

データの出処

国立国語研究所共同研究プロジェクト「文章における語彙の分布と文章構造」（チームリーダー山崎誠）