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

文章は文の連なりからなる。文は一連の語や記号の系列で成り立っている。そこで、語や記号の連なりを“状態”の系列とみなし状態間の推移で文を表現することにより文章の構造を表現し様々な分析に使えないかという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文の構造の分析には適している。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、場面に応じて定義される。

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。文章の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。

２．品詞による表現

品詞状態による表現の例を示す。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表２のように数値を対応させた。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。

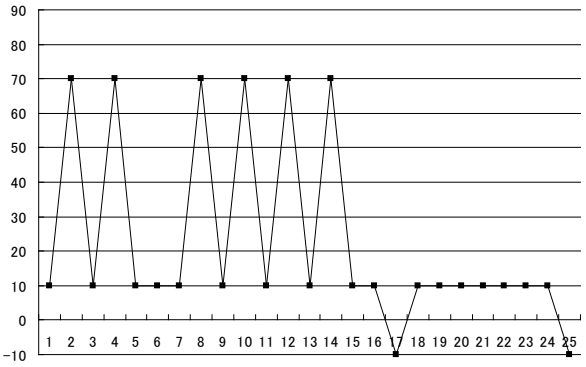

図１　形態素解析（品詞）による
タイトル部分の時系列表現

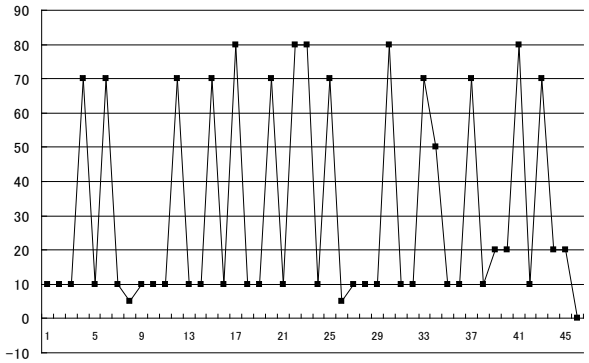

図２　形態素解析（品詞）による文の
時系列表現

表２　品詞等の状態表現

品詞等	状　態
句点	0
読点	5
名詞	10
動詞	20
形容詞	30
副詞	40
連体詞	50
接続詞	60
助詞	70
助動詞	80
接頭詞	90
その他	-10

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

表３　句等による状態表現

句等	状態
句読点	0
名詞句	10
動詞句	20
形容詞	30
副詞	40
連体詞	50
接続詞	60
その他	−10

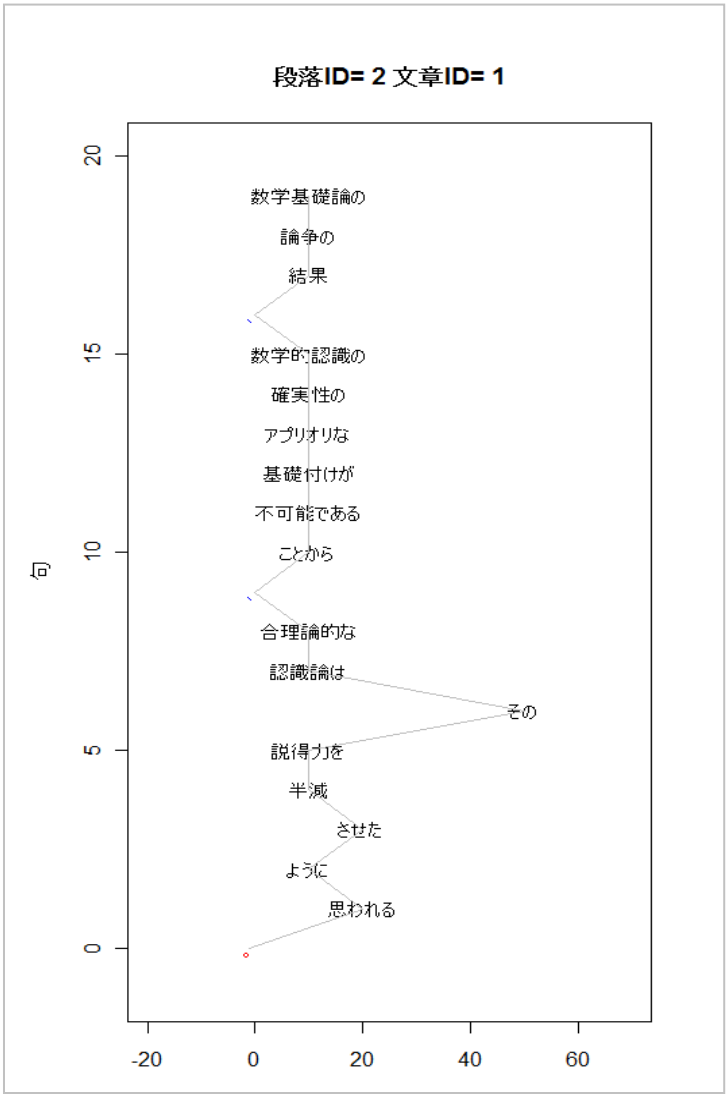

図３　助詞の違う要素表現（句）による文の時系列表現

３．句による表現

品詞による時系列を観察すると、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。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行列を作ったものが表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。なお句等はその他を除いて、出現頻度の多い順に並べてある。

表４　助詞による状態表現の推移確率(%)（近藤論文）

	名詞の	動詞	読点	句点	名詞を	名詞に	名詞が	名詞な	名詞は	名詞	その他
名詞の	18.6	0.5	0	0	13.6	7.1	11.3	7.3	6.8	1.3	34.2
動詞	9.7	0	7.2	26.2	5.5	9.4	7.5	2.8	7.2	5	20.1
読点	26.2	1.5	0	0	7.2	6.3	6.6	9.6	3.9	3	35.4
句点	11.1	0	0	0	3.6	3.6	8	4.4	12.9	9.3	46.4
名詞を	3.5	41.4	7.1	0	0	10.1	0.5	3.5	0.5	3.5	29.6
名詞に	4.3	41.7	7.5	0	1.6	3.2	1.6	6.4	0	2.7	30.5
名詞が	5.6	26	6.8	0	4	9	0	4.5	0	1.7	43.1
名詞な	18.3	0	1.8	0	20.7	3.7	9.8	2.4	7.3	0.6	34.9
名詞は	9.9	3.5	47.2	0	7	4.2	0.7	5.6	0.7	0.7	20.3
名詞	1.1	15.1	21.5	34.4	0	0	0	0	0	2.2	26.1

表５　句等の出現頻度
（近藤論文）

	度数	割合(%)
名詞の	382	11.7
動詞	362	11.1
読点	332	10.2
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
名詞は	142	4.3
名詞	93	2.8
その他	1003	28.6

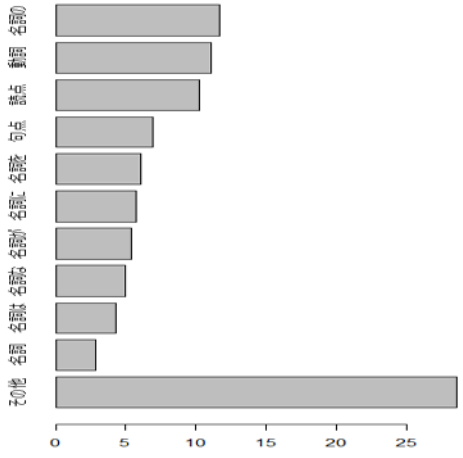

図４　句等の出現頻度（近藤論文）

４．論文の特徴の比較

比較のためにもう一つ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。推移行列を表６に、各句等の状態の出現頻度を表７に示してある。この推移行列の状態も表４と同様に出現頻度の順に並べてある。

表４、表５、表６、表７を見ることにより、２つの論文の表現の比較が可能になる。

表６　助詞による状態表現の推移確率(%)（横地論文）

	読点	動詞	名詞の	名詞を	句点	名詞に	名詞は	名詞	名詞が	動詞て	その他
読点	0	0.4	22.9	8.2	0	6.5	9.5	4.8	8.2	0.4	38.7
動詞	10.2	0	6.6	8	28.8	8.4	8.4	8.8	1.8	0	18.4
名詞の	0	1.1	4.4	22	0	11	6.6	0.5	10.4	0	42.9
名詞を	0.6	32.9	1.9	0	0	13.3	3.2	5.1	0.6	13.9	28.1
句点	0	0	5.3	6.1	0	3	23.5	8.3	0.8	0.8	52.7
名詞に	7.6	35.6	1.7	3.4	0	1.7	0.8	3.4	3.4	15.3	26.4
名詞は	31.9	1.7	15.5	8.6	0	5.2	0	2.6	3.4	0	31.5
名詞	41	12	0	0	15.7	1.2	2.4	1.2	0	4.8	21.6
名詞が	5.1	23.1	9	6.4	0	7.7	0	1.3	0	11.5	36.1
動詞て	12	48	4	2.7	0	1.3	1.3	0	1.3	4	25.1

表７　句等の出現頻度
（横地論文）

	度数	割合(%)
読点	231	11.4
動詞	226	11.2
名詞の	182	9
名詞を	158	7.8
句点	133	6.6
名詞に	118	5.8
名詞は	116	5.7
名詞	83	4.1
名詞が	78	3.9
動詞て	75	3.7
その他	624	27.6

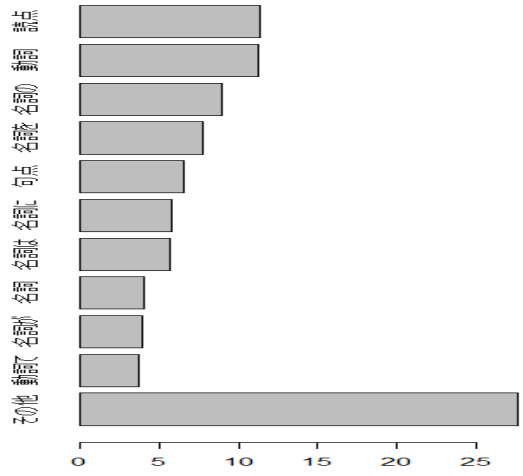

図５　句等の出現頻度（横地論文）

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等の傾向が存在していることが分かる。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行列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行列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．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。これまでの文章の構文解析は，表４，表５，表６，表７に示した単語の出現割合または推移確率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。文章の構造理解には動的な要素も重要であり，図３に示したような時系列的な表現と解析が必要である。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。この動的要素が色濃く出るものが詩また音楽の世界の歌詞である。今回の試みはまだ試行錯誤の段階であるが，今後大量のテキストデータの分析をすることによって、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

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.

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

横地徳広（2005）「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』」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）