

## 1 はじめに

### N グラムモデルを用いた古典テキスト研究の現状

最近、近藤泰弘・近藤みゆき両氏による研究 ([10][11][12][13]) に端を発し、東洋学のテキスト研究において N グラム言語モデルを用いた比較の手法が注目を集めている ([6][7][15][22][23][24])。N グラムモデルとは、シャノン[3]によって提唱された情報理論に基づき、「N 個の文字列または単語列」を数えあげその発生する確率を計算することで、ある言語、テキスト、文章の特徴を記述しようとするものである。言い換えれば、単語や文字(「アイテム」と総称)の生起が直前の N-1 個のアイテムのみに依存するという一方向的・線の性質のものとして言語活動をモデル化し、確率・統計的な処理の俎上にのせられるようにしたものである。例えば、日本語の文章で「ばらき」という音(モーラ)の列は「茨城」「原木中山」などの地名、「薔薇・木」「バラキ(マフィア映画のタイトル)」など、単語レベルでもそれなりに高い確率で発生するが、順番を並び替えただけの「らきば」は、「牛は、みはしのひらきばしらにつなぎ」(『大鏡』)や「これから木場へ行く」など複合的な例はあるものの、「ばらき」と比べて発生する確率はかなり低い、というような具合である。

N グラムモデルについては、出力結果が膨大であること、多くのノイズを含むこと、あるいはゼロ頻度問題 ([9], p. 62-71) など、種々の問題点が指摘されているにもかかわらず、音声認識や OCR などの分野ですぐれた成果を挙げている。加えて古典研究の分野では、従来一般的であった研究者によるキーワードの指定や形態素分析による比較分析に対して、

1. 現代人には通常認知できないデータの構造性や規則性を探り出すことができる。
2. キーワードの採取や形態素分析における規準など、曖昧な要素を排除することができる。
3. 形態素分析などで分割されてしまうような比較的長い文字列を比較することができる。

などの利点が指摘されている ([13]など)。特に 1. や 2. については、テキスト読解において助けになると同時に障害にもなる研究者の先入観をある程度克服しうるのでないかと期待できる点で重要である。

石井公成氏はさらに、2 つ以上の N グラム分析による出力結果をマージし、出現頻度付きの表形式で表現することで、異本の比較や翻訳者の判定などにおいて効率よくテキストの比較分析することが可能であることを示した ([6][7])。氏が NGSM (N-Gram based System for Multiple document comparison and analysis) と呼ぶこの方法については、主に漢字文献情報処理研究会内で議論をもとに筆者や近藤

---

<sup>1</sup> moro@ya.sakura.ne.jp、早稲田大学メディアネットワークセンター非常勤講師。

泰弘氏によってツールが開発され、また XML と組み合わせた応用も模索されている ([22])

### N グラム分析結果のクラスター分析への活用

本稿では、この NGSM による古典テキストの比較研究を一步進めて、その出力結果を用いたクラスター分析を試みる。その目的は次の 2 点である。

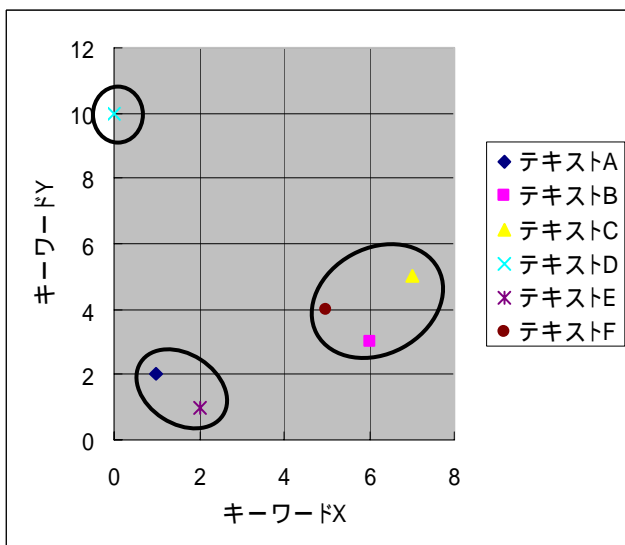
1 NGSM の出力結果を “ 読む ” 際の研究者の先入観や恣意性を、より排除できる可能性が期待できる。

2 大量のテキストを処理することが可能。

2. について付言すれば、「大規模なデータから思いがけない( unsuspected )パターンを発見する」 ([19], p. 5) 手法として近年注目を集めているデータマイニング ( Data Mining ) に通ずる点として注目されるだろう。

ところで、クラスター分析とは、個体 ( テキストの比較であれば各テキスト ) がどのようにグループ分けできるかを調べるための統計的手法である。N グラムモデルによる分析結果をクラスタリングするにあたっては、NGSM としてマージされた各個体の N グラム頻度を座標とみなして空間の点として表現し、点と点との間の距離を計算して近いものからグループ分けしていくという手順をとる。このようなアイデアは、キーワードの出現頻度をベクトルとして表現し、各テキスト間の距離をベクトルの余弦距離によって判定する Salton らのモデル ([2] など) にまで遡ることができ、また主に日本において発展してきた情報幾何 ([4][5][8] など) の理論へと通ずる考え方である。ごく簡単な例で言えば、次表のような頻度のデータがあったとする。

	テキスト A	テキスト B	テキスト C	テキスト D	テキスト E	テキスト F
キーワード X	1	6	7	9	2	5
キーワード Y	2	3	5	10	1	4



このような個体 ( テキスト ) の集合があった場合、キーワード X の出現頻度を x 軸、キーワード Y の出現頻度を y 軸にとってグラフを書きみると左図のごとくなる ( このような空間を特徴空間 feature space あるいは文書空間 document space とする )。この中で、距離的に近いところにある点を楕円で囲い 3 つに分類したが、これがクラスターである。実際の NGSM 分析では変数の数が何万、何千となるのが普通であり、左のような単純な 2

次元のグラフに書くことはできないが、任意の  $n$  次元空間における座標間の距離を計算するのは可能であり、コンピュータの得意とする分野である。

#### 『般若心経』を用いる理由

ところで、後に述べるように、個体間の距離やクラスター間の距離を算出する方法にはいくつかあり、どれを選択するかによって結果が異なってくる。村上征勝氏はクラスター分析によるテキストの真贋判定について「文献の真贋判定においてどの分析結果が最も信頼できるかは（中略）真作の文献が一つのクラスターを構成するか否かで判断すればよい」（[21], p. 48）と述べられ、従来の文献学的方法による先行研究の活用によって、計量的・統計的なテキスト分析の信頼性・妥当性が高まることを示唆している。本稿で比較分析を行う『般若心経』については古来多くの研究がなされており、テキストの系統に関しては評価が固まったと言っても過言ではないが、逆に方法論的な妥当性を検証する材料としては適当なのではないかと考えられる<sup>2</sup>。

クラスター分析は、個体をグループ分けする際にまったく先行研究がなく、仮説や予想をたてることが困難なものを扱うことに適しているのであるが、本稿ではその妥当性を検証し問題点を指摘することで、クラスター分析による未開拓のテキスト分類をより信頼性の高いものとしたいと考えている。文献学者および確率・統計研究者の両方からのご叱正を乞う次第である。

## 2 『般若心経』異訳のクラスター分析

#### 『般若心経』の異訳

本稿において比較する『般若心経』の異訳は以下の通りである<sup>3</sup>。この經典のサンスクリット本には大本と小本の2種類の系統があり、漢訳にも両系統が存する。内容的な差異はないが、大本には小本に相当する部分の前後に序分（導入部）と流通分（經典の普及を勧める結末部）が追加されている。

1. 小本（漢字数 300 字前後）
  - (ア) 鳩摩羅什訳（400 年ごろ）『摩訶般若波羅蜜大明呪経』
  - (イ) 玄奘訳（649 年）『般若波羅蜜多心経』
2. 大本（漢字数約 600～700 字）
  - (ア) 法月訳（741 年）『普遍智蔵般若波羅蜜多心経』
  - (イ) 般若・利言共訳（790 年）『般若波羅蜜多心経』
  - (ウ) 智慧輪訳（861 年）『般若波羅蜜多心経』
  - (エ) 法成訳（856 年）『般若波羅蜜多心経』

---

<sup>2</sup> これに加えて『般若心経』は、最も有名な玄奘訳が 260 字（流布本では 262 字）という具合に小部のテキストであり、N グラム分析や NGSM による出力結果も小さく抑えられるため、実験に向いているという面もある。

<sup>3</sup> ここに挙げたもののほか、義浄訳（7～8 世紀）と伝えられる『仏説般若波羅蜜多心経』が存するが、玄奘訳とほとんど変わるところがなく、偽撰と見られているため（[18], p. 69-76）、今回の比較材料には含めなかった。

(オ) 施護訳 (1000 年ごろ) 『聖仏母般若波羅蜜多心經』

テキストは『大正新脩大藏經』に収録されたものを使用する。その理由は、SAT<sup>4</sup>やCBETA<sup>5</sup>などでデータベース化されていること、テキストの品質が統一されていること、という2点である。

#### NGSM による般若心經異訳の比較分析

まず、『般若心經』の異訳それぞれを 3 グラムで分析する。分析に際しては、自作のプログラム「morogram」<sup>6</sup>を利用した。このプログラムは、[17]のアルゴリズムに基づいた文字単位での N グラム処理が可能であるが、加えて漢字文献の NGSM 処理を念頭において開発したため他の N グラムツール<sup>7</sup>にはない以下のような特徴がある。

1. 0~16 面の Unicode に対応 (入出力は UTF-8 のみ)
2. 実体参照形式&Mnnnnnn; (1 nnnnnn 131,072) を一文字として扱うことが可能。
3. NGSM 分析に際して重要な頻度 1 の採取が可能。

下にあげたのは左が玄奘訳『般若心經』、右が法月訳『普遍智藏般若波羅蜜多心經』の 3 グラムによる分析結果の一部である。左から頻度、採取された文字列、グラム数になる。この後に行う NGSM 比較のために、頻度 1 から採取している。

2	一切苦	3	1	一切世	3
1	三世諸	3	1	一切苦	3
1	三菩提	3	1	一時佛	3
1	三藏法	3	1	七千人	3
1	三藐三	3	1	七萬七	3
1	上咒是	3	1	三世諸	3
1	不垢不	3	1	三昧力	3
1	不增不	3	1	三昧安	3
1	不淨不	3	1	三昧正	3
1	不減是	3	1	三昧總	3
1	不減不	3	1	三菩提	3
1	不生不	3	1	三藏沙	3
1	不異空	3	1	三藐三	3

<sup>4</sup> <http://www.l.u-tokyo.ac.jp/~sat/>

<sup>5</sup> <http://www.cbeta.org/>

<sup>6</sup> <http://www.ya.sakura.ne.jp/~moro/resources/ngram/morogram.html>

<sup>7</sup> 藤原滋氏作「ngram」(<http://www.jaist.ac.jp/~shigeru/ngram-ja.html>) など。なお、morogram の開発にあたっては藤原氏の ngram より多くの教示を得た。記して感謝申し上げる。

次にこれらの結果を比較するため、NGSM 方式でマージする。ツールとしては、高速な近藤泰弘氏の Perl スクリプト「ngmerge」<sup>8</sup>を利用した。出力結果の一部を下にあげる。この出力結果の読み方は、例えば 1 行目の「一切世」であれば、鳩摩羅什訳、玄奘訳、般若・利言訳、施護訳では 0 回の頻度で、法月訳、智慧輪訳、法成訳では 1 回の頻度でこの文字列が登場するということを表している。

一切世	( 羅什訳:0 玄奘訳:0 法月重訳:1 般若共利言等訳:0 智慧輪訳:1 法成訳:1 施護訳:0 )
一切大	( 羅什訳:0 玄奘訳:0 法月重訳:0 般若共利言等訳:0 智慧輪訳:0 法成訳:0 施護訳:1 )
一切如	( 羅什訳:0 玄奘訳:0 法月重訳:0 般若共利言等訳:1 智慧輪訳:1 法成訳:1 施護訳:1 )
一切法	( 羅什訳:0 玄奘訳:0 法月重訳:0 般若共利言等訳:0 智慧輪訳:0 法成訳:1 施護訳:1 )
一切苦	( 羅什訳:2 玄奘訳:2 法月重訳:1 般若共利言等訳:1 智慧輪訳:1 法成訳:0 施護訳:1 )
一切諸	( 羅什訳:0 玄奘訳:0 法月重訳:0 般若共利言等訳:0 智慧輪訳:0 法成訳:2 施護訳:0 )
一切顛	( 羅什訳:1 玄奘訳:0 法月重訳:0 般若共利言等訳:0 智慧輪訳:0 法成訳:0 施護訳:1 )
一句唵	( 羅什訳:0 玄奘訳:0 法月重訳:0 般若共利言等訳:0 智慧輪訳:0 法成訳:0 施護訳:1 )
一時世	( 羅什訳:0 玄奘訳:0 法月重訳:0 般若共利言等訳:0 智慧輪訳:0 法成訳:0 施護訳:1 )
一時佛	( 羅什訳:0 玄奘訳:0 法月重訳:1 般若共利言等訳:1 智慧輪訳:0 法成訳:0 施護訳:0 )
一時薄	( 羅什訳:0 玄奘訳:0 法月重訳:0 般若共利言等訳:0 智慧輪訳:1 法成訳:1 施護訳:0 )
一菩薩	( 羅什訳:0 玄奘訳:0 法月重訳:0 般若共利言等訳:0 智慧輪訳:1 法成訳:0 施護訳:0 )
七千人	( 羅什訳:0 玄奘訳:0 法月重訳:1 般若共利言等訳:0 智慧輪訳:0 法成訳:0 施護訳:0 )
七萬七	( 羅什訳:0 玄奘訳:0 法月重訳:1 般若共利言等訳:0 智慧輪訳:0 法成訳:0 施護訳:0 )
三世一	( 羅什訳:0 玄奘訳:0 法月重訳:0 般若共利言等訳:0 智慧輪訳:0 法成訳:1 施護訳:0 )
三世諸	( 羅什訳:1 玄奘訳:1 法月重訳:1 般若共利言等訳:1 智慧輪訳:1 法成訳:0 施護訳:1 )

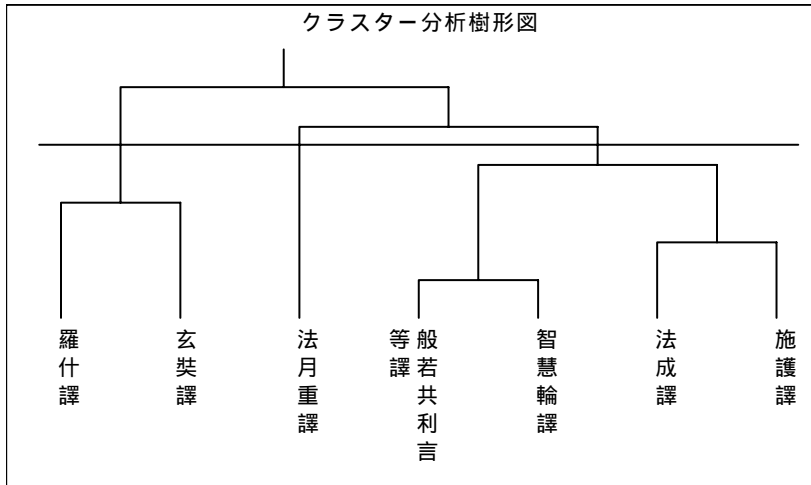
この結果にざっと目を通すだけでも、同じ小本系である鳩摩羅什訳と玄奘訳に共通する場合が多いことに気づく等、NGSM の有効性を知ることができる。しかしながら、『般若心経』というごく小部のテキストでさえ 7 つのテキストの 3 グラムにおける NGSM 結果が 1,738 行にものぼることを考えれば、通常の分量をもつテキスト群に対して任意のグラム数で分析すれば NGSM 結果は爆発的に増加し「ざっと目を通す」ことさえ困難になることが予想される。大量の NGSM 結果を網羅的に活用するためには、クラスター分析のような手法が必要となってくるのである。

#### 『般若心経』異訳のクラスター分析

続いて、上の NGSM 分析の結果についてクラスター分析をし樹形図に描くと次ページのような力が得られる。クラスタリングの方法は、例えば鳩摩羅什訳であれば特徴空間上に (0, 0, 0, 0, 2, 0, 1, 0...) という座標の点を置き、各点 (= テキスト) 間の距離の遠近にしたがって近いところからクラスターへとまとめていく。個体間の距離の測定には標準化ユークリッド距離を、クラスター間の距離

<sup>8</sup> <http://klab.ri.aoyama.ac.jp/tool/index.html> なお、利用には Unix 系のコマンド sort が必要。

の測定にはワード法を、それぞれ一般的であるという理由で用いた（他の方法による結果とその問



題点については後述する。計算と作図については、「EXCEL多変量解析 Ver. 4.0」(株式会社エスミ<sup>9</sup>)を使用した。『般若心経』には小本系(羅什訳・玄奘訳)と大本系(その他)の二系統があることは前述したとおりであるが、こ

れがきちりと分けられている点は注目される。渡辺章悟氏は「羅什訳とされる『大明呪経』が『大品般若』と玄奘訳の『般若心経』を定本として作られたもの」([25], p. 79)と分析されており、鳩摩羅什訳と玄奘訳の近さがこれによっていよいよ傍証されると言えるだろう。

また大本の分類においても、先行研究([1], p. 68など)が以下のように指摘する翻訳年代やテキストの系統を反映した分類になっているように思われる。

- 法月訳(741年)…東インド系
- 般若・利言共訳(790年)…カシュミール系
- 智慧輪訳(861年)…中央アジア系
- 法成訳(856年)…チベット系(敦煌発見)
- 施護訳(1000年ごろ)…ウディヤーナ(ガンダーラ地方)系

特に、東インド系である法月訳が他の中央アジア系と分離している点は注目されよう。

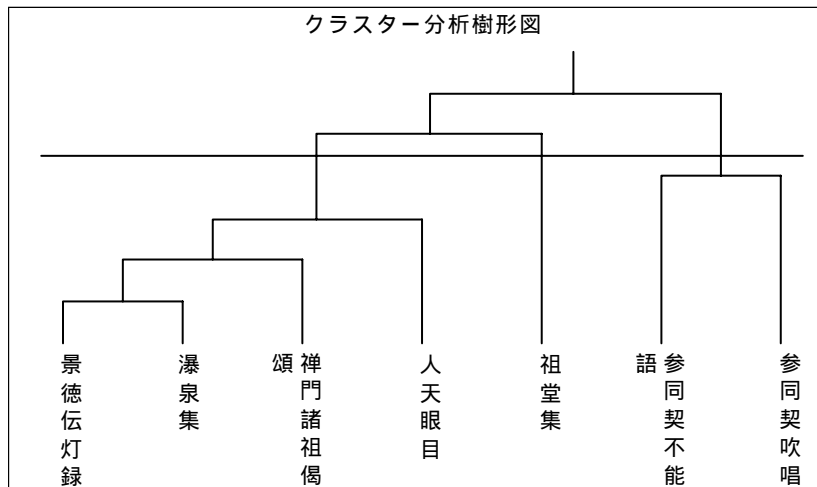
#### 附録・『参同契』異本のクラスター分析

参考資料として、石頭希遷によって著わされたとされる『参同契』の異本を石井公成氏がクラスター分析を試みた結果を紹介する。比較した異本は以下の通りである。

1. 『景德伝灯録』所収テキスト(1004年)
2. 『瀑泉集』所収テキスト(~1052年)
3. 『禅門諸祖偈頌』所収テキスト(南宋末頃)
4. 『人天眼目』所収テキスト(1188年)

<sup>9</sup> <http://www.esumi.co.jp/>

5. 『祖堂集』所収テキスト（高麗版本）
6. 『参同契不能語』所収テキスト（江戸、1736年）
7. 『参同契吹唱』所収テキスト（江戸、1767年）



これら異本の 2 グラム分析による結果を、先と同様の方法(標準化ユークリッド距離・ワード法)でクラスター分析すると、左のような樹形図が得られる。この分析においてもほぼ時代順・国別に分類され

ている上、椎名宏雄氏の、

- (一) 『祖堂集』本は、最も文字の異同が著しく、独特のテキストである。
- (二) 『景德伝灯録』以下、中国撰述の四文献には、それぞれ独自の文字が二～四字ずつみられるものの、たがいに近似の関係にある。
- (三) 本邦江戸期の注解書二種は、たがいに独自の文字が三～四字ずつみられるが、ほぼ類似の関係にあり、全体としては(一)よりも(二)に近い。

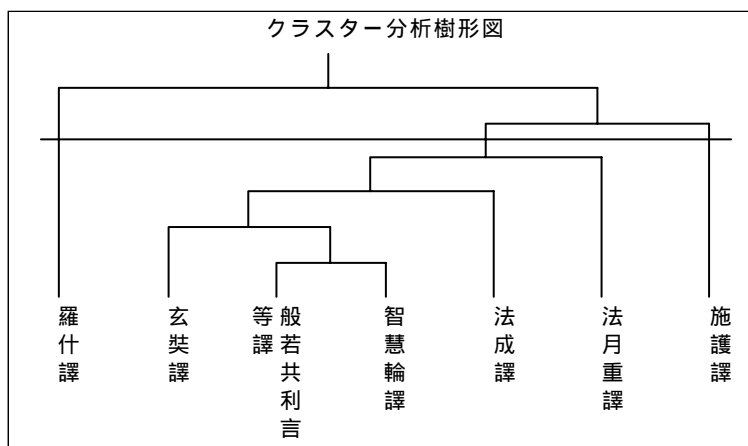
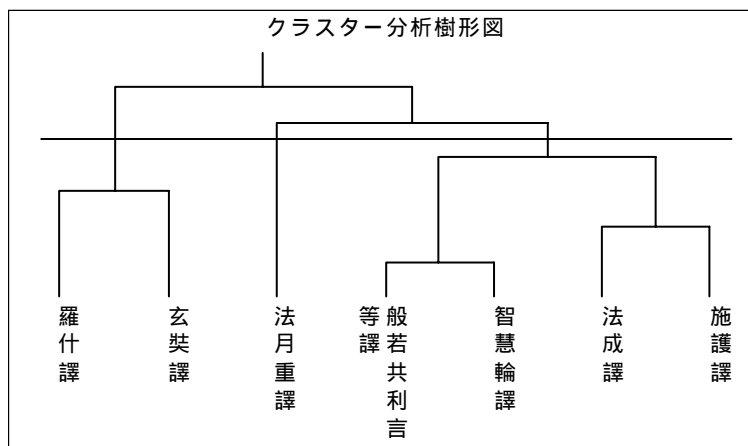
という分析 ([14], p. 193) にほぼ一致する。しかしながら、上の(三)の最後に指摘されている中国系と日本系が近いという点については樹形図と異なる。

### 3 問題点と今後の課題

#### 方法の選択の問題

クラスター分析において、変数間の距離の計算方法には市街地距離・ユークリッド距離・標準化ユークリッド距離・マハラノビスの距離などがあり、クラスター間の距離計算には最短距離法・最長距離法・群平均法・重心法・ワード法などがある。一般的に、変数間の距離には標準化ユークリッド距離が、クラスター間の距離は最長距離法またはワード法がよいとされており、先の『般若心経』および『参同契』の分析においても、標準化ユークリッド距離とワード法に基づく分析によって先行研究からそれほど逸脱しない結果が得られた。しかしそれは、経験的に得られた妥当性であって、現在のところ何らかの理論的な裏づけがあるものではない。

また、先行研究とクラスター分析の結果とを比較する基準についても、解釈者の恣意が強く働いていることは否定できない。先行研究と異なる結果がクラスター分析によって得られた場合 実際、



先の分析においても一部相違する部分が見られた  
それをどのように考えるかについては大きな問題であろう。

左の図は、『般若心経』異本に対して、上が最長距離法、下が最短距離法によってクラスター分析をした樹形図の出力結果である。前者が先に示したワード法による出力結果とほとんど同じである反面、後者はまったく異なった結果を出力していることが興味深い。後者の結果が、テキスト群に内在するこれまで見出されることのないテキストの関係性を反映したものなのか、

あるいはまったく的外れなモデルに基づいたことによる間違った結果なのかについては、今後の検討が必要である。

#### 統計処理における妥当性

また、本稿で紹介した『般若心経』および『参同契』のクラスター分析の例は、いずれも小部のテキストについて行ったものであり、変数の数も少ない(とは言え、Nグラム分析結果は前述の通り大変多くなる)。この少なさが、統計上の誤差をどの程度生んでいるのかについては、今後検証の必要があるのではないだろうか。

さらに言えば、本稿で行った分析では、各変数については重み付けなどの処理をまったく行っていない。異なる大きさのテキストの距離を測る際、各テキストの大きさを標準化(特徴空間上の点を長さ1のベクトルに揃えるなど)したり、有意な語やキーワードに特別な重みをつける(例えば「摩訶般若」の「摩訶」「般若」と「訶般若」とでは異なった重みをつける)など、様々な処理が考えられる。ただし、どの文字列を「有意味」と判断するかについては、分析者の恣意性が働くことは否定できないだろう。

その他にも、前述した N グラム自体が抱える問題点や、クラスター分析における順序依存性の問題 ([20]) など、考慮すべき課題は山積している。

#### 学際的研究の必要性

以上の問題については(これらの結果を問題提起とした)幅広い文献学的な研究が必要となると同時に、確率・統計を専門とする研究者との学際的な共同研究が必要となってくるのではないだろうか。古典テキストに対する N グラム分析や NGSM、クラスター分析については、豊かな可能性が想像しうるだけに、本稿がその議論のたたき台となれば幸甚である。

## 4 参考文献

- [1] Conze, Edward. *The Prajnaparamita literature*. Tokyo: Reiyukai, 1978.
- [2] Salton, G., A. Wong and C. S. Yang. "A Vector Space Model for Automatic Indexing". *Communications of the ACM*, 18-11 (1975).
- [3] Shannon, Claude E. and Warren Weaver. *The Mathematical Theory of Communication*. 1949; 長谷川淳・井上光洋訳 『コミュニケーションの数学的理論』(明治図書、1969)
- [4] 甘利俊一・長岡浩司 『情報幾何の方法』(岩波講座応用数学 21 [対象 12]、1993)
- [5] Amari, Shun'ichi and Hiroshi Nagaoka. *Methods of Information Geometry (Translations of Mathematical Monographs, Vol 191)*. AMS & OUP, 2000.
- [6] Ishii, Kosei. "Classifying the Genealogies of Variant Editions in the Chinese Buddhist Corpus". 2001 EBTI International Conference, Seoul, Korea, May 2001 口頭発表 (Proceedings 刊行予定)
- [7] 石井公成 「N-gram 利用の可能性 仏教文献における異本比較と訳者・作者判定」(『漢字文献情報処理研究』2、好文出版、2001)
- [8] 今井浩 「幾何クラスタリングの情報計算幾何構造」(森下真一・宮野悟編 『発見科学とデータマイニング』、共立出版、2001)
- [9] 北研二 『確率的言語モデル』(辻井潤一編・言語と計算 4、東京大学出版会、1999)
- [10] 近藤みゆき 「n グラム統計処理を用いた文字列分析による日本古典文学の研究 『古今和歌集』の「ことば」の型と性差」(千葉大学 『人文研究』29、2000)
- [11] 近藤泰弘 「《文化資源》としてのデジタルテキスト 国語学と国文学の共通の課題として」(『国語と国文学』平成 12 年 11 月特集号)
- [12] 近藤泰弘・近藤みゆき 「平安時代古典語古典文学のための N-gram を用いた解析手法」(『言語処理学会第 7 次年次大会発表論文集』、2000)
- [13] 近藤泰弘 「コンピュータによる文学語学研究にできること 古典語の「内省」を求めて」(全国大学国語国文学会夏季大会シンポジウム「情報技術は文学研究をいかに変えるか」要旨、2001、<http://klab.ri.aoyama.ac.jp/public/paper/20010602.pdf>)
- [14] 椎名宏雄 「「参同契」の性格と本文」(『宗学研究』23、1981)

- [15] 谷本玲大「曖昧検索性をもたせた N-gram サーチの手法 『新撰万葉集』と菅原道真の詩の比較を例に」(『漢字文献情報処理研究』2、好文出版、2001)
- [16] 長尾眞・森信介「大規模日本語テキストの n グラム統計の作り方と語句の自動抽出」(情報処理学会研究報告 1993-NL-96)
- [17] Nagao, Makoto and Shinsuke Mori. "A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese". In Proceedings of the 15th International Conference on Computational Linguistics (1994).  
<http://www-lab25.kuee.kyoto-u.ac.jp/member/mori/postscript/Coling94.ps>
- [18] 福井文雅『般若心経の総合的研究 歴史・社会・資料』(春秋社、2000)
- [19] 福田剛志・森本康彦・徳山豪『データマイニング』(データサイエンスシリーズ、共立出版、2001)
- [20] 宮本定明『クラスター分析入門』(森北出版、1999)
- [21] 村上征勝『真贋の科学 計量文献学入門』(朝倉書店、1994)
- [22] 師茂樹「XML と NGSM によるテキスト内部の比較分析実験 『守護国界章』研究の一環として」(『漢字文献情報処理研究』2、好文出版、2001)
- [23] 山田崇仁「初めての N-gram Cygwin もしくは Perl を用いて」(『漢字文献情報処理研究』2、好文出版、2001)
- [24] 山田崇仁「『世本』と『國語』章昭注引系譜資料について N-gram 統計解析法による分析」(『立命館史学』22、2001)
- [25] 渡辺章悟「般若心経成立論序説 『摩訶般若波羅蜜大明呪経』と『大品般若経』の関係を中心として」(『仏教学』31、1991)

付記 本研究にあたっては、データをご提供いただいた石井公成氏をはじめ、漢字文献情報処理研究会のメンバーに多大なご教示を頂いている。記して感謝申し上げます。