

N グラムによる比較結果からの用例自動抽出

禅宗系の偽経を題材に

師 茂樹¹

1 はじめに

N グラムによる古典テキストの分析は、近藤みゆき・近藤泰弘両氏による国文学・国語学のテキストの研究（[8][9][10][11][12][13][14]）に端を発し、それに触発された谷本玲大氏による漢字文献への応用（[15]）、石井公成氏、報告者（師）による漢字仏教文献の研究（[3][4][19][20][21]）、山田崇仁氏による中国古典の研究（[24]）などで活用されるようになった。この方法において重要なのは、複数のテキストに対する N グラム分析を比較対照することで、テキスト間の共通性を網羅的に調査することができる点であり、その点に注目したものとして石井公成氏による NGSM (N-Gram based System for Multiple document comparison and analysis) の提唱や、報告者によるクラスタ分析への応用などがある。しかし、特に後者においては、テキスト間の「距離」を大雑把に把握し、視覚化することが可能であるものの、実際にどの文字列が近い関係にあるのかという細かな分析に入っていく段階では、膨大な出力結果を人力で探していくという手段しかなかった。

本報告では、膨大な NGSM の出力結果から有意な用例を自動的に抽出することを試み、併せて NGSM によるテキスト分析、クラスタ分析における方法論的な問題点について指摘したい。

2 対象とするテキストとそのクラスタ分析結果

沖本克己「MENSURA ZOILI 禅文献の計量語彙論的研究の試み」

本研究で扱うテキストは、沖本克己氏による禅文献に対する文字単位での統計的分析（[5]）で分析対象とされていたものである（次頁表）。沖本氏の研究は、禅文献に対する計量的分析の先駆的試みであり、文字単位の頻度・占有率・テキスト間の相違度などを分析したものであるが、これまで顧みられることはほとんどなかったと言ってよい。しかし、禅研究の専門家による統計分析として、また統計的なテキスト分析研究には必須である妥当性を検証するための比較材料として、評価されるべきではなからうか。

沖本氏の結論は以下のとおりである。

1. 『起信論』は他のテキストとの共通性が高く、その普遍的性格を示すごとくであるが、『法王経』と『法句経』に対しては語彙の相違が大きい。この傾向は『楞伽経』『仏頂経』などにも見られ、『法王経』と『法句経』が特異なポジションにあることを予想させる。
2. （略）

¹ s-moro@hanazono.ac.jp 花園大学専任講師

A. 禅宗系偽経	B. A に影響を与えたと思われる文献	C. 比較のための中庸的な文献
T273 『金剛三昧経』(650～665) T842 『円覚経』(7世紀末～8世紀初) T945 『楞嚴経』(8世紀) T1484 『梵網経』 T2883 『法王経』(8世紀) T2901 『法句経』(650年ごろ) S5532 『禅門経』(7世紀末～8世紀初)	T670 四巻『楞伽』 T1666 『大乘起信論』	T7 法顕訳『涅槃経』 T1558 玄奘訳『俱舍論』

3. 『楞伽経』はほぼ完璧に『金剛三昧経』、『法王経』、『法句経』を包含しているといつてよい。
4. 『金剛三昧経』は『起信論』に似た数値を示すが、『法王経』に近縁性を示す点が異なる。
5. 『法王経』は『金剛三昧経』に似た数値を示す。また両者の異なり度は低いから、形態的には相似性が高い。

N グラム + クラスタ分析による分析

次に、同じテキスト群²に対して N グラムによる分析と、そのクラスタ分析を試みたので併せて報告したい ([21])。

各テキストを 3 グラムで分析³し、それらをマージした NGSM の結果は、変数数 47 万強 (CSV ファイルで 20MB 強) であった (文末表)。これをクラスタ分析に用いる際、文献の文字数にばらつきがあるため、各文献を長さ 100 のベクトルとして正規化した (文末表)。個体 V の変数 $V_1, V_2, V_3 \dots V_n$ について、個体 V の長さを

$$L = \sqrt{\sum_{i=1}^n V_i^2}$$

とすれば、長さ 100 に正規化された個体 V の変数は

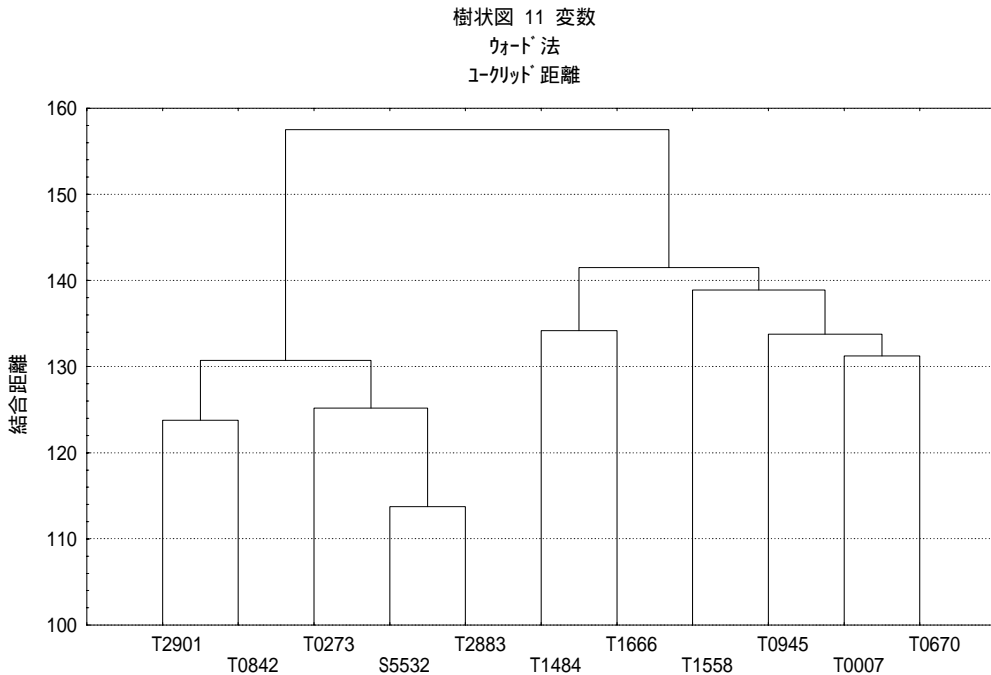
² 厳密に言えば、まったく同じ電子テキストを用いたわけではない。沖本氏が自身で入力したものを使っているのに対して、報告者は CBETA (<http://www.cbeta.org>) や SAT (<http://www.l.u-tokyo.ac.jp/~sat/>) で公開されているものを使っているため、校正レベルで言えば恐らく沖本氏の使った電子テキストよりもより正確なものを用いていると思われる。もっとも、ここで言う「より正確な」とは、底本である大正新脩大蔵経等により近いという意味であって、厳密なテキスト批判 (というものがそもそもあり得るのか、という問題はあがあるが) を経たものではない。

³ N グラム分析においては、自作のプログラム (morogram) を用いた。morogram は現在のところ <http://www.ya.sakura.ne.jp/~moro/resources/ngram/morogram.html> で公開中であるが、オープンソースにすべく Sourceforge.jp に移行中である (<http://sourceforge.jp/projects/morogram/>)。

$$V'_n = \frac{V_n}{L} \times 100$$

とすることができる。長さを 100 にしたのは、変数が極端に小さくならないようにするためである。

これをユークリッド距離およびワード法によってクラスタ分析した結果が以下のデンドログラムである。



この結果を、先に引いた沖本氏の研究 ([5]) の結論と比較してみると、一致点と相違点とが見られる。

一致点を見てみると、デンドログラムの左側のクラスタに『法句経』、『円覚経』のグループ (1-a) と『金剛三昧経』、『禅門経』、『法王経』のグループ (1-b) ができており、右側のクラスタに『梵網経』、『起信論』(2-a) と玄奘訳『俱舍論』、『楞嚴経』、法顕訳『涅槃経』、四巻『楞伽』のグループ (2-b) がそれぞれできている点は、沖本氏の結論 1、4 の後半、5 と一致する。特に、1-a および 1-b については、石井公成氏が「『金剛三昧経』に『法句経』や『円覚経』などを含めたこれらの一群の如来蔵系の偽経は、禅と華嚴との交流の中で生まれ、その系統に伝えられていったようである」([1], p. 367) と述べるように、文献学的な分析を通じた分類とも一致するのが興味深い。

一方、相違点を見てみると、結論 3 によれば『法句経』(1-a)、『金剛三昧経』、『法王経』(以上 1-b) と同じグループに入るべき『楞伽経』(2-b) が、結論 4 の前半によれば『金剛三昧経』(1-b) と同じグループに入るべき『起信論』(2-a) が、それぞれまったく反対のグループに属している点をあげられよう。加えて、先学によって『円覚経』(1-a) と『起信論』(2-a) との密接な関係が指摘されてき

たが ([6][18][23])、クラスタ分析の結果はそれとも相違する。この点についても、検討課題としてあげられよう。

3 用例の自動抽出

さて、以上のような問題意識を前提に、各テキスト間で強い共通性のある文字列について検討したいと思うが、47 万を超える用例の中からそれを抽出するのは人力では容易なことではない。したがって、それを自動的に抽出する方法が必要になってくるであろう。今回は、用例を抽出したいテキスト群について、

1. 各語における偏差値を求め (文末表)
2. 平均値からの距離 (偏差値 - 50) を求め、それを合計し、
3. 合計値の大きな順にソートする。

というごく単純な方法で、用例が抽出できないか試みた⁴。結果の一部を文末表 に掲載している。

この結果を見てみると、全体としては、1-a および 1-b において「～菩薩」「菩薩～」や「善男子」といった語が共通して用いられていることが推測される一方、2-a および 2-b においては、そのような単語レベルでの共通性ではなく、語のつながり (文体) の共通性によってクラスタが形成されているように思われる。

前者については、沖本氏の文字単位での分析では指摘されなかった 2 文字の用例について抽出することができたことは有意義であったと思われる。特に「菩薩」については、サンスクリット語 bodhisattva の音訳語であり、漢字の表語文字としての性格よりも表音文字としての性格が強くていう例であるといえるだろう。これを比較材料にできたかできなかったかが結果の相違につながったと考えられないだろうか。

ただし、近藤みゆき・近藤泰弘氏が指摘するように ([10][12])、形態素分析等に基づくテキスト分析に対して、N グラムによる分析の有利な点として、

1. 形態素 1 語の単位を認定する基準が一通りではないのに対して、網羅的な分析が可能。
2. 複合語や強い共起性のある単語群 (連語、慣用句など) の分析に有利。

という点があげられるが、単語レベル (形態素レベル) での結びつきが強くと出してしまうと、このような特長が失われる可能性もあるかもしれない。

4 まとめ・課題

以上、大雑把な方法による分析ではあったが、いくつかの知見と課題が得られたのでまとめておきたい。

⁴ 本報告で使用したスクリプトについては、近く Sourceforge.jp 内の morogram のサイト (前述) において公開する予定である。

- クラスタを特徴付ける用例の自動抽出については、NGSM 結果を“読む”にあたって生じるであろう研究者のバイアスを排除する意味でも意義のあることではないか。
- ただし、その方法については、今後さらなる検討を要する。特に、NGSM によって得られた数値の評価の仕方については、確率・統計的な厳密さが要求されるのではないか。
- また、クラスタの結びつき方が単語に依存するのか、文体に依存するのかについてはこれまで明確に考察されては来なかったので、N グラム分析におけるグラム数の設定の仕方や、キーワードや形態素によるテキスト分析との比較を視野に入れながら検討されなければならないだろう。
- 補足として、柳田聖山氏が『金剛三昧経』や『法句経』の出現を考慮するならば、達摩系の主張が偽経に影響した時代を、此に先立って考える必要があり、更に古い『最妙勝定経』との関係も問題であり、後に『禅門経』や、『法王経』、『円覚経』、『首楞嚴経』、及び『起信論』や『釈摩訶衍論』、『円明論』、『宝蔵論』などの偽経の出現と、禅思想の発展を迹づけることも可能となる」([22], p. 484) と述べるように、比較対照の拡大が今後必要であろう。

参考文献

- [1] 石井公成「金剛三昧経の成立事情」(『印度学仏教学研究』92、1996)
- [2] Ishii, Kosei. “Classifying the Genealogies of Variant Editions in the Chinese Buddhist Corpus”. (『電子佛典』第3輯、東國大専校 EBTL、2001)
- [3] 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)
- [4] 石井公成「仏教学における N-Gram の活用」(東京大学東洋文化研究所附属東洋学研究情報センター編『明日の東洋学』8、2002)
- [5] 沖本克己「MENSURA ZOILI 禅文献の計量語彙論的研究の試み」(『禅文化研究所紀要』19、1993)
- [6] 鎌田茂雄『宗密教学の思想史的研究』(東京大学出版会、1975)
- [7] 鎌田茂雄「円覚十二菩薩の形成 円覚経の造像化」(『印度学仏教学研究』93、1998)
- [8] 近藤みゆき「平安時代和歌資料における特殊語彙抽出についての計量的研究と利用ツールの公開 古今和歌集の歌語と表現のジェンダー性について」(『科学研究費特定領域研究 人文科学とコンピュータ 研究成果報告書 コンピュータ支援による人文科学研究の推進 1999』)
- [9] 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)
- [10] 近藤みゆき「n-gram 統計による語形の抽出と複合語 平安時代語の分析から」(『日本語学』20、2001年8月号)
- [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「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>)
- [13] 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)
- [14] 近藤泰弘・近藤みゆき「N-gram の手法による言語テキストの分析方法 現代語対話表現の自動抽出に及ぶ」(『漢字文献情報処理研究』2、好文出版、2001)
- [15] 谷本 玲大「曖昧検索性を持たせた N-gram サーチの手法 『新撰萬葉集』と菅原道真の詩の比較を例に」(『漢字文献情報処理研究』2、2001年10月、好文出版)
- [16] 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).
- [17] 水野弘元「菩提達摩の二入四行説と金剛三昧經」(『印度学仏教学研究』6、1955)
- [18] 望月信享『仏教経典成立史論』(法蔵館、1946)
- [19] 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』第2号、好文出版、2001)
- [20] 師茂樹「N グラムモデルとクラスター分析を用いた漢文古典テキストの比較研究 『般若心経』の異訳の比較を例に」(京都大学大型計算機センター第69回研究セミナー「東洋学へのコンピュータ利用」、2002)⁵
- [21] 師茂樹「N グラムを用いたクラスタ分析による禅文献分類の試み」(禅学研究会第73回学術大会における口頭発表、2002年11月30日、花園大学)⁶
- [22] 柳田聖山『初期禅宗史書の研究 中国初期禅宗史料の成立に関する一考察』(禅文化研究所『研究報告』第1冊、1966)
- [23] 柳田聖山『中国撰述経典 I 円覚経』(筑摩書房、1987)
- [24] 山田崇仁「『世本』と『国語』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号、2001)

⁵ 予稿集に載せる予定だった原稿: <http://www.ya.sakura.ne.jp/~moro/resources/20020322moro.pdf>

⁶ 当日配布予定だったレジュメ: <http://www.ya.sakura.ne.jp/~moro/resources/20021130NgramZen.pdf>

自動抽出による上位 40 語

1-a

27.3525916095022	明善男
27.3525916095022	守護是
27.3525916095022	子當知
27.0211392503145	動善男
26.7668642447423	男子當
26.742827487413	菩薩普
26.6252261937444	大衆及
26.5047123535368	師利菩
26.4819554312378	利菩薩
26.0935596679381	自在菩
25.9765083309122	音菩薩
25.9765083309122	賢菩薩
25.9765083309122	普賢菩
25.9765083309122	彌勒菩
25.9765083309122	勒菩薩
25.9516813188295	剛藏菩
25.5984523536829	與諸衆
25.5968639075434	如是善
25.4429815798255	於是普
25.4429815798254	養不惜
25.4429815798254	除諸病
25.4429815798254	護眼目
25.4429815798254	識應當
25.4429815798254	證於如
25.4429815798254	證實相
25.4429815798254	諸佛名
25.4429815798254	行者及
25.4429815798254	衆說於
25.4429815798254	衆徳本
25.4429815798254	薩普賢

25.4429815798254	薩彌勒
25.4429815798254	菩薩彌
25.4429815798254	若遇如
25.4429815798254	經已一
25.4429815798254	經名及
25.4429815798254	純以七
25.4429815798254	礙善男
25.4429815798254	知法不
25.4429815798254	生得聞
25.4429815798254	無礙善

1-b

20.0935053899842	爲虚妄
19.9553074677476	子如是
19.938449618567	解脫佛
19.5802689364051	如是之
19.5752254071381	一菩薩
19.4623211086646	五十年
19.4591715665879	救衆生
19.4547862474572	男子於
19.4362743653532	經五十
19.4218213312229	無住處
19.4018994864102	法若如
19.3886152952529	縛何以
19.3886152952529	法是如
19.3886152952529	即是動
19.3886152952529	不有何
19.3886152952529	一不一
19.3828696895764	離蓋纏
19.3828696895764	間動不
19.3828696895764	諸境不
19.3828696895764	諦聽爲
19.3828696895764	言無説

19.3828696895764	行亦無
19.3828696895764	衆皆大
19.3828696895764	衆生宣
19.3828696895764	處住心
19.3828696895764	薩若有
19.3828696895764	薩若化
19.3828696895764	薩能以
19.3828696895764	薩汝能
19.3828696895764	薩名者
19.3828696895764	菩薩我
19.3828696895764	菩提虚
19.3828696895764	菩提菩
19.3828696895764	菩提汝
19.3828696895764	若心無
19.3828696895764	若在淨
19.3828696895764	聽爲汝
19.3828696895764	而化衆
19.3828696895764	眞性不
19.3828696895764	眞實令

1-a・1-b 共通

14.9078269160204	千萬億
14.7855202567028	生死故
14.7288002948978	爲諸衆
14.6713806123764	爲諸大
14.6393121060731	清淨大
14.5330985587737	於此經
14.4822019874844	無起無
14.4822019874844	滅善男
14.4822019874844	心善男
14.4822019874844	宣説一
14.4797969172627	人俱其
14.4738563619669	提何以

14.4724390880166	菩薩名
14.4632103925132	薩皆悉
14.4395558281604	是經名
14.3913871709997	男子此
14.3648501387909	如是人
14.3542760000304	心若無
14.2685416774028	百千萬
14.2543908039896	菩提何
14.21632545968	心不動
14.1799414872683	得阿耨
14.1770022427951	尊云何
14.1075326195625	子菩薩
14.0838159580745	空無所
14.0733560901108	善男子
14.069124551111	男子菩
14.0653768152764	解脫佛
14.0492998194851	如護眼
14.0491024847644	世尊云
13.973698168138	菩提佛
13.9727081313996	三菩提
13.9496440558413	於虛空
13.9406903164294	子如是
13.9229947595396	施不如
13.9229947595396	布施不
13.9186560833771	男子若
13.9174477517057	如是事
13.9088787631769	爲虛妄
13.8958158609109	薩及諸

2-a

27.2420885857251	爲他人
27.2420885857251	求法者
27.2420885857251	果畢竟

27.2420885857251	於後代
27.2420885857251	心者入
27.2420885857251	心念念
27.2420885857251	之心貪
27.1700520216305	請法師
26.8307135032516	當成佛
26.7012135967763	方佛土
26.5030852655619	風動心
26.5030852655619	道轉法
26.5030852655619	自知有
26.5030852655619	盜不姪
26.5030852655619	生正信
26.5030852655619	生根行
26.5030852655619	殺不盜
26.5030852655619	成道轉
26.5030852655619	平等空
26.5030852655619	因名爲
26.5030852655619	可窮盡
26.5030852655619	像菩薩
26.5030852655619	以之爲
26.5030852655619	亦能使
26.5030852655619	二十卷
26.5030852655619	不變以
26.5030852655619	不盜不
26.5030852655619	不可窮
26.5030852655619	三寶之
26.3650015294697	心者有
26.2075316552744	而生一
26.2075316552744	心貪著
26.2075316552744	如是信
26.2075316552744	中若見
26.0596872310656	無量行

25.9834056623382	法體性
25.8612754386773	菩薩今
25.8591379130043	法乃至
25.7610086757777	習畢竟
25.7313958439327	衆生平

2-b

17.3442352549772	我於此
17.2910073760677	非得非
17.2910073760677	種說有
16.9818382105111	然世尊
16.923339775836	說所以
16.8378187250572	當知若
16.8082556515528	非有外
16.8082556515528	道世尊
16.8082556515528	轉是故
16.8082556515528	謂得無
16.8082556515528	癡云何
16.8082556515528	異世尊
16.8082556515528	所謂因
16.8082556515528	不動何
16.6683688736994	名爲上
16.5778181972123	爲三種
16.5778181972123	世間現
16.5347942269223	有無想
16.5260678640515	羅如是
16.4768004971584	以是因
16.4021686446756	生世尊
16.3974272576313	見彼諸
16.3974272576313	五受陰
16.3471729087018	有八無
16.3471729087018	和合性
16.3445494552579	在此中

16.2873948258466	名爲色
16.2765925894207	所以者
16.2765925894207	以者何
16.2675936967111	中見有
16.0497442436787	所說二
16.0328194152106	於爾時
15.9964908385191	此名爲
15.9964039055334	解脫我
15.9423655624386	地如來
15.9088715725906	入無所
15.7749341422359	應於此
15.7588085728074	善根生
15.7322953371645	者則不
15.7284801681106	然無有

2-a · 2-b 共通

13.3513077073798	爾時即
13.3302007299765	地一切
13.1045333037606	我於此
13.06431668414	非得非
13.06431668414	種說有
13.0629524696239	然世尊
12.9662169765272	譯爾時
12.9129816130224	有十種
12.870514585416	所說二
12.8444819256587	生世尊
12.8411640397647	如是過
12.8364302344713	殺生者
12.8128281686081	有無想
12.8109660945293	有如是
12.807950087157	故說一
12.7865233861872	說所以
12.7779129036536	其身如

12.7219074811543	當知若
12.7215920767797	外一切
12.6995709367288	非有外
12.6995709367288	道世尊
12.6995709367288	轉是故
12.6995709367288	謂得無
12.6995709367288	癡云何
12.6995709367288	異世尊
12.6995709367288	所謂因
12.6995709367288	不動何
12.6808731588229	故種種
12.6647754746523	差別而
12.6470953350357	故我於
12.6341069971165	現前不
12.6104604676856	在此中
12.6098888557231	有八無
12.6098888557231	和合性
12.6013229079352	樂如是
12.5938787045729	名爲上
12.5932917630001	所以者
12.5932917630001	以者何
12.5601912338274	方便生
12.5344708483187	中見有