

計量的テキスト分析の目的

文体 (style) の分析 他研究方法との連携が必須

- 「ながらく古い文献に泥んでいると、言葉とは、文字に記された記号だと思って疑うことがない。しかし、考えるまでもなく、言葉はいかにも文字面ではあるけれど、それ以上に音声であり、形態でありそしてそれが発せられる様々な状況を反映するものであり、つまり自らをめぐる総体としての環境を引っ括めた現象の一つに他ならない」(沖本克己 1993)。

隠れたパターンの発見

- 「文体に指紋があるとすれば、それはどのようなものだろうか？ それはおそらく、ある著者の文体的な特徴——例えば ‘such as’ の生起度数といった、まったく取るに足りないと言ってもよいような特徴を組み合わせたもの——であって、指紋と同様にその人に特有のものであろう。文体上些細で取るに足りぬ特徴だからといって、文体分析に利用しない理由にはならない。指先にある渦巻や輪が我々の容姿においては大切でも目につくわけでもないが、指紋が一生変わらないように、そういったものこそが著者の叙述において変化することのない特徴となるはずであり、他の書き手には見られないその人だけのものとなるはずであろう」(Kenny, Anthony (1982), 邦訳 p. 24 [一部改訳])。
- 「一人三人作家といわれた長谷川海太郎の作品が、読点の付け方の情報をみる限り一つにまとまった」(村上征勝 1994)。
- 「徹底的に網羅的な研究 (すべての単語・すべての文字の単位にまで網羅性を及ぼすことが可能になる)」「それによって現代人には通常認知できないデータの構造的な規則性を探り出す。それは、現代人の古典語に対する「内省」(introspection) (語感) の欠如を補うことができ、文学研究に貢献する。なぜなら、古典文学の正しい読みにとって、「内省」(文法的直観と言語外知識など) の欠如は大きな障害のひとつだからである」(近藤泰弘 2001)。

大規模な文献の分析

- 「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」(福田他 2001, p. 5)。

N グラムを用いたクラスタ分析

N グラム・モデル

- 「n 個の文字列または単語列」の発生する確率を、ある文章、あるテキストの特徴と見なす言語モデル。
  - Shanon et al. 1949.
  - 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・線的性質のものとして言語活動をモデル化。
  - 例

◇ 文字単位・3 グラム

摩	訶	般							
	訶	般	若						
		般	若	波					
			若	波	羅				
				波	羅	蜜			
					羅	蜜	多		
						蜜	多	心	
							多	心	經

◇ 単語単位・3 グラム

The	quick	brown							
	quick	brown	fox						
		brown	fox	jumped					
			fox	jumped	over				
				jumped	over	the			
					over	the	lazy		
						the	lazy	dog.	

- 形態素解析と比べて単純で網羅性が高い。
- ノイズが多い (「多心經」はノイズか? ——何がノイズなのか)。
- 国文学・国語学の近藤みゆき・近藤泰弘両氏による研究→仏教学、中国古典などへの広がり (石井 2001-2002、Ishii 2001、師 2001-2002、山田 2001)

NGSM (Ishii 2001)

- N-Gram based System for Multiple document comparison and analysis
- N グラム分析の結果を比較しやすいように表にする手法。

	羅什訳	玄奘訳	法月重訳	般若等訳	智慧輪訳	法成訳	施護訳
一切世	0	0	1	0	1	1	0
一切大	0	0	0	0	0	0	1
一切如	0	0	0	1	1	1	1
一切法	0	0	0	0	0	1	1
一切苦	2	2	1	1	1	0	1
一切諸	0	0	0	0	0	2	0
一切顛	1	0	0	0	0	0	1

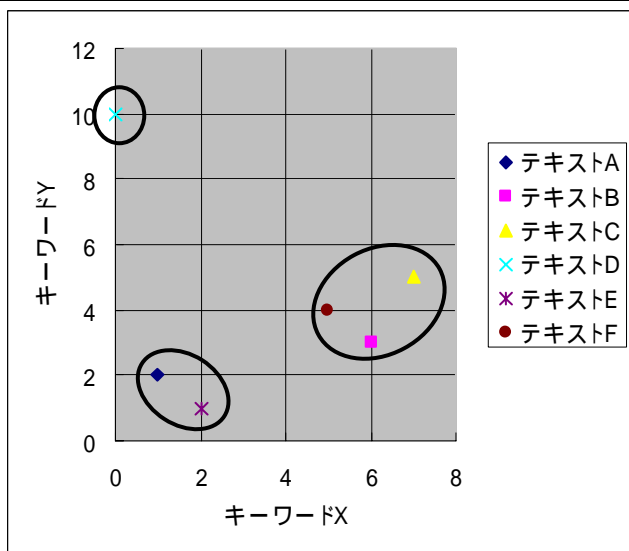
一句庵	0	0	0	0	0	0	1
一時世	0	0	0	0	0	0	1
一時佛	0	0	1	1	0	0	0
一時薄	0	0	0	0	1	1	0
一菩薩	0	0	0	0	1	0	0
七千人	0	0	1	0	0	0	0
七萬七	0	0	1	0	0	0	0
三世一	0	0	0	0	0	1	0
三世諸	1	1	1	1	1	0	1

### クラスタ分析

- 個体間の距離によってグループ化する手法。

➢ 例

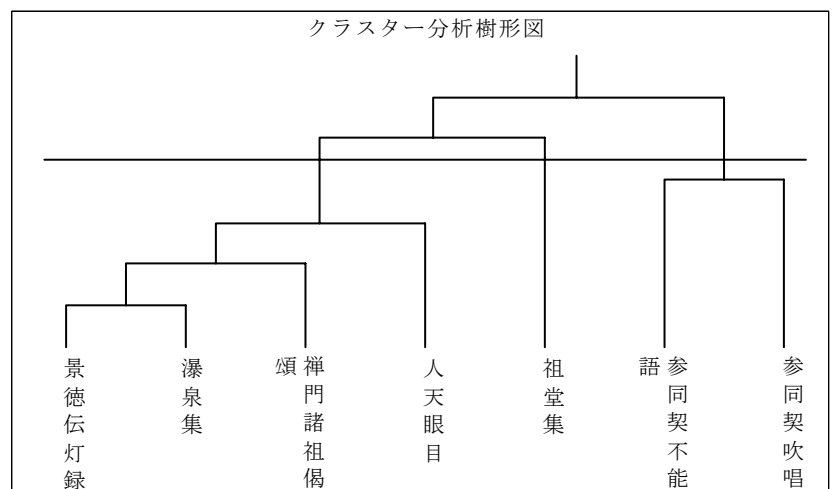
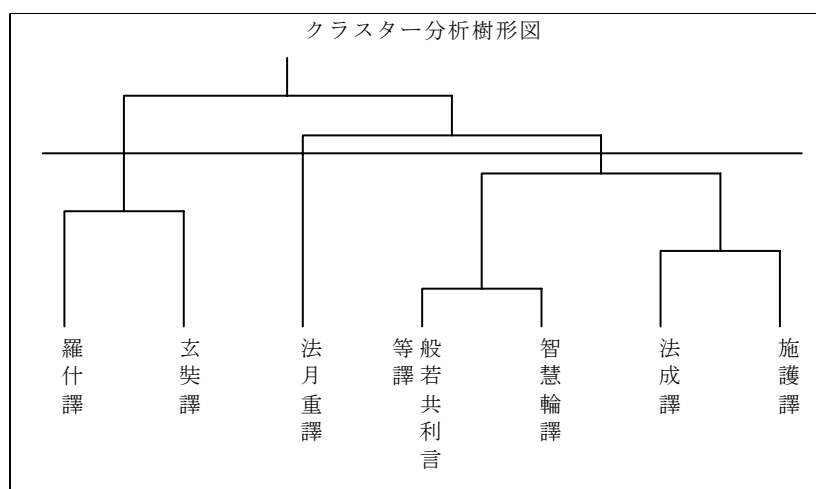
		個体					
		テキスト A	テキスト B	テキスト C	テキスト D	テキスト E	テキスト F
変数	キーワード X	1	6	7	9	2	5
	キーワード Y	2	3	5	10	1	4



- 個体間の距離、グループ間の距離を算出するためにさまざまな計算方法がある。
- 樹形図によって表現。

- NGSM を使ったクラスタ分析

➢ 実例 (師 2002)



- 従来の研究は形態素解析が主であった (村上 1994)。
- ☆ N グラムは大規模文献の分析に有利。

### 沖本 1993 の再検討

- 禅文献に対する計量的分析の先駆的試み。
- 文字単位の頻度・占有率・テキスト間の相違度などを分析。

#### 比較対照

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)	T670 四卷『楞伽』	T7 法顕訳『涅槃経』
T842 『円覚経』(7世紀末~8世紀初)	T1666 『大乘起信論』	T1558 玄奘訳『俱舍論』
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)		

### 結論

- 『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。
- (略)
- 『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。
- 『金剛三昧経』は『起信論』に似た数値を示すが、『法王経』に近縁性を示す点異なる。
- 『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。

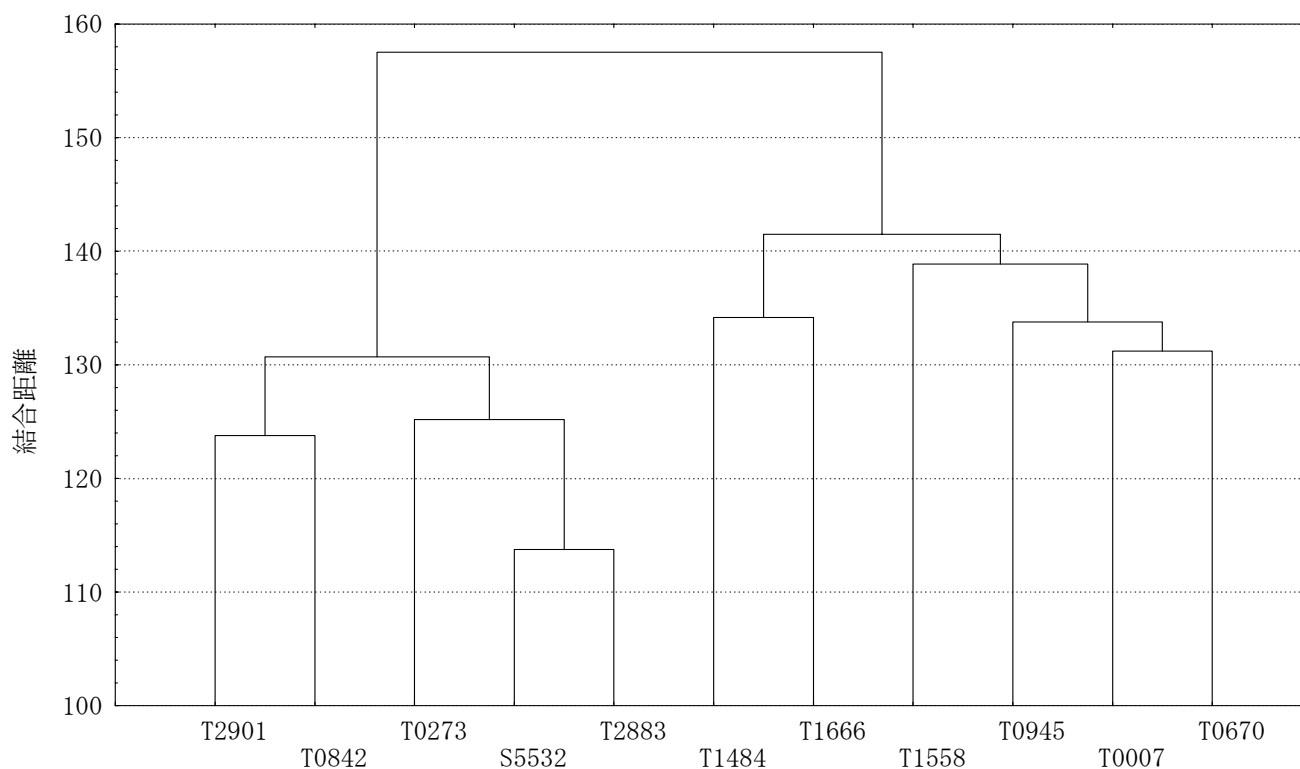
### Nグラム+クラスタ分析による分析

- NGSMの結果、変数数47万強(CSVファイルで20MB強)。
- 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。

➤ 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 $L = \sqrt{\sum_{i=1}^n V_i^2}$ とすれば、長さ100に正規化された個体Vの変数 $V'_n = \frac{V_n}{L} \times 100$

樹状図 11 変数

ワード法  
ユークリッド距離



- 沖本 1993 の結論との一致点 (1, 4 の後半, 5)
  - 第1グループ
    - 法句経、円覚経
    - 金剛三昧経、禅門経、法王経
    - ◇ 『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」(石井 1996, p. 367)
  - 第2グループ
    - 梵網経、起信論
    - (玄奘訳俱舍論,) 楞嚴経、(法顕訳涅槃経,) 四卷楞伽
- 沖本 1993 の結論との相違点 (3, 4 の前半)
  - 語彙(用字)的に『楞伽経』は『金剛三昧経』等と重なるが、文体的には離れるということの意味。
  - 望月 1946、鎌田 1975、柳田 1987 等では『円覚経』と『起信論』との密接な関係が指摘されている。

## まとめ・課題

- 計量的テキスト分析の有効性が確認できたのではないか。
  - 従来の研究との相違点については要検討。
- 比較対照の拡大が必要。
  - 『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、禅思想の発展を述べつけることも可能となる」(柳田 1966, p. 484)。

## 参考文献

- 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③、共立出版、2001)
- 石井公成「金剛三昧経の成立事情」(『印仏研』92)
- Ishii, Kosei. "Classifying the Genealogies of Variant Editions in the Chinese Buddhist Corpus". (『電子佛典』第3輯、東國大學校 EBTI、2001)
- 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』第2号、好文出版、2001)
- 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』第8号、2002)
- 鎌田茂雄『宗密教学の思想史的研究』(東京大学出版会、1975)
- Kenny, Anthony. *Computation of Style: An Introduction to Statistics for Students of Literature and Humanities*. 1982; 吉岡健一訳『文章の計量 文学研究のための計量文体学入門』(南雲堂、1996)
- 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』第29号、2000)
- 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)
- 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>)
- 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)
- 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』第2号、好文出版、2001)
- 水野弘元「菩提達摩の二入四行説と金剛三昧経」(『印仏研』6、1955)
- 望月信享『仏教経典成立史論』(法蔵館、1946)
- 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)
- 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)
- 村上征勝『真贋の科学 計量文献学入門』(朝倉書店、1994)
- 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)
- 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>
- 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』第19号、1993)
- Shanon, 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)
- 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』第2号、好文出版、2001)
- 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』第2号、好文出版、2001)
- 山田崇仁「『世本』と『國語』韋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)
- 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)
- 柳田聖山『中国撰述経典 I 円覚経』(筑摩書房、1987)