計量分析を利用した仏教説話のパラレル検出の試み
――『賢愚経』を中心として――

石田勝世

1．はじめに

本稿の目的は，効率的にパラレル（並行話）を探し出すための計量的手法を提案し，先行研究の結果と比較して，その有効性を検討することである[1)]．

仏教説話の研究においてパラレル探しはよく行われている．多くの場合は，登場人物の名前など固有名詞を中心とするキーワードに注目して，それを大蔵経で検索する方法がとられる．しかし，類話の中には登場人物の名前がほとんど可変項となって大きく違っている場合もあり，その場合には「全体的に似た話」，「状況的に似た話」，「文面的パラレル」等でもうまく探し出すような分析が必要となるはずである．そのような「全体的に似た話」や「場面的に似た話」等を含めて高速に探し出す方法への試みとして，自然言語処理を応用した計量的手法を提案し，実際に漢訳『賢愚経』に適用した結果を報告する．本計量分析手法は，テキストの文字列をＮ文字に区切ってその出現頻度の類似度を計算して，類似度の高いテキストをパラレル候補として出力する．人手による分析と比較すると精度的に劣る可能性があるものの，高速にパラレル候補が求められるので，広範囲のテキスト群のなかでのパラレル探しへの利用が期待される．

本稿では，まず，Ｎグラムを利用した計量的手法を提案した．次に，この手法の有効性を検討するために『賢愚経』と類似した仏教説話探しに適用してみたところ，先行研究の結果とほぼ同じものが得られた．このＮグラムを利用した計量的手法は，パラレルを探し出すのに有効であることが確認された．

2．手法の説明

Ｎグラムを利用し，２つのテキストの類似度を計量的に求める手法について説明する．パラレルを計量的に探し出すためには，２つのテキスト間で類似している度合いを数値で表現する必要がある．このために２つのテキスト間の類似度を

表現する尺度を導入する．本研究ではNグラムを利用してテキストの類似度を表現する[2)]．

まず，用語を定義する．「Nグラム」とは，テキストをN個の文字列（または単語列）で区切ったものである．Nは正の整数値で2や3などが使われることが多い．たとえば，次のようになる．

例）「私は学校に行きます」
　1グラム　私，は，学，校，へ，行，き，ま，す
　2グラム　私は，は学，学校，校へ，へ行，行き，きま，ます
　3グラム　私は学，は学校，学校へ，校へ行，へ行き，行きま，きます

Nグラムは，「は学」「校へ」など意味のない（とり難い）文字列になることもある．テキストに現れるNグラムの出現頻度（出現する回数）を「Nグラム出現頻度」と本稿では呼ぶことにする．以下に例をあげる．

例）ある漢訳テキストの3グラムとその出現頻度
　佛世尊4　←「佛世尊」という文字列が4回出現している．
　佛報恩5
　佛提婆2
　作何物2　など

Nグラム出現頻度は，ベクトルで表現され，テキストの何らかの特徴を表現していると考えられる．2つのテキストの類似度をNグラム出現頻度の相関係数[3)]で定義する．相関係数が大きければ2つのテキストの類似度が大きく，相関係数の値が小さければ2つのテキストの類似度が小さいと判断される．あるテキストのパラレル検出は，類似度の高い（＝相関係数の高い）テキストを探し出すことになる[4)]．

筆者が調べた限りでは，仏教説話においてNグラムを用いたパラレル検出の先行研究はなかった．Nグラムの他の応用分野での先行研究として漢訳『般若心経』の異訳を比較分類したものがある．異訳テキスト7本（大本5本と小本2本）をNグラムで分析し，その類似度にもとづいたクラスタ分析により「大本と小本が分離した」「大まかな系統推定が得られた」等の興味深い結果が報告されている（師 2002）．

3．手法の適用

Nグラムを利用した計量的手法の有効性を検討するために，『賢愚経』（大正202，全13巻，69品）と類似したテキストを探すパラレル検出に適用してみた．

検索範囲は，『撰集百縁経』（大正200，全10巻，100品）と『大方便仏報恩経』（大正156，全7巻，9品）である．いずれも本縁部に属する漢訳テキストで，『賢愚経』と類似した説話が多く含まれていることが指摘されている．計算方法は，2つの品ごとにNグラム出現頻度の相関係数を求める．相関係数は，2つの品ごとに計算する．すなわち，『賢愚経』と『撰集百縁経』とで69×100＝6,900（個），『賢愚経』と『大方便仏報恩経』とで69×9＝621（個）の相関係数を計算することになる．結果を表1と表2に示す．紙幅の都合上，先行研究との比較を表わす表のみをあげた．表の見方は，たとえば表1の1行目の場合，「『賢愚経』の品1と『撰集百縁経』品34」のペアが干潟1978でパラレルと指摘されている．そのペアの相関係数は－0.01であり，「『賢愚経』の品1」と「『撰集百縁経』品1から品100まで」のペア100個の相関係数のなかでの順位が6位であることを表現している．

結果をまとめると次のようになる．

①『撰集百縁経』：全般に相関係数の値は小さい．3つの先行研究のなかでは，出本1995に近い結果が得られた（順位がすべて1位）．

②『大方便仏報恩経』：『撰集百縁経』の場合と比べて相関係数の値はやや大きい．順位からみて干潟1978よりも梁2002に近い結果が得られた．

③全体を通して：先行研究との違いは視点の違いと考えられる（詳細は要検討）．相関係数の値そのものは小さいものの，相関係数でソートして順位づけすると先行研究との類似性が見い出される．順位が高いものには先行研究では示されていない未知のパラレルが発見された可能性もある．相関係数の値が小さいのは，共通した文字列が少なく頻度がゼロのデータが多いためと考えられる．

表1　先行研究との比較（『撰集百縁経』）

	『賢愚経』	『撰集百縁経』	相関係数	順位
干潟1978	1	34	－0.01	6
	1	35	0.16	1
	3	59	－0.31	1
	4	72	－0.47	12
	6	98	0.64	1
	8	79	0.28	1
	9	83	－0.17	1
	18	51	－0.51	84
	26	73	－0.08	2

表2　先行研究との比較（『大方便仏報恩経』）

	『賢愚経』	『大方便仏報恩経』	相関係数	順位
干潟1978	1	2	0.02	5
	7	1	－0.22	8
	42	6	0.40	1
	61	9	0.04	3
	69	5	0.09	2
梁2002	1	3	0.12	1
	7	2	0.18	1
	16	7	0.03	1
	31	7	0.21	1

干潟 1978	36	88	− 0.37	1
	28	87	− 0.48	42
	38	31	− 0.53	6
	60	60	− 0.03	1
出本 1995	1	35	0.16	1
	3	59	− 0.31	1
	6	98	0.64	1
	8	79	0.28	1
	36	88	− 0.37	1
	60	60	− 0.03	1
梁 2002	1	33	− 0.05	14
	1	34	− 0.01	6
	1	35	0.16	1
	3	59	− 0.31	1
	4	72	− 0.47	12
	6	98	0.64	1
	8	79	0.28	1
	9	83	− 0.17	1
	18	51	− 0.51	84
	26	73	− 0.08	2
	36	88	− 0.37	1
	38	31	− 0.53	6
	39	66	− 0.39	50
	46	60	− 0.69	80
	60	60	− 0.03	1
	62	5	− 0.45	34
梁 2002	32	5	− 0.01	2
	42	6	0.40	1
	61	9	0.04	3
	68	5	0.01	1
	69	5	0.09	2

4．まとめ

先行研究とほぼ類似した結果が得られたこと（それ以上の結果が得られた可能性もある）[5]，さらに，「人手による探索と比べて広範囲な探索が短時間で可能なこと」「grep 検索のように検索パタンを事前に指定する必要がないこと」「事前に想定していなかった視点でのパラレル探索が可能なこと」などの利点もあることから，N グラムを利用した計量的手法の実用可能性は高いと評価できる．今後の課題として「探索範囲の拡大（本縁部や律部の全体に）」「探索機能の強化[6]」「漢訳以外への適用（蔵，梵）」等があげられる．

1）テキスト校訂において「写本の系統樹作成」や「パラレルなテキスト探索」はテキスト成立過程を検討するうえで重要である．筆者は，これらの作業をサポートする計量

的手法を開発して実際のテキスト校訂に役立てたいと考えている．　　2）この他にも2つのテキストの類似性を計測する手法は自然言語処理の分野で種々提案されているが，本研究では「簡易に利用できること」「途中結果が見えること」等からNグラムを採用した．　　3）相関係数は，2つのデータの相関（類似）の度合いを表現する統計数値である．−1から+1の実数値をとり，大きい値ほど相関が高い．　　4）パラレルには「全体的に似た話」「場面的に似た話」「登場人物などの固有名詞が一致」「定型的パタン」「丸ごとコピー」などがある．本稿で提案した手法では，事前にパラレルの種類を指定して探索することはできないが，想定していなかった視点でのパラレルが検出される可能性はある．　　5）本稿では，先行研究との比較という点から手法を評価したが，未知のパラレル探索という点からの評価は行っていない．先行研究では指摘されていないパラレルが得られた可能性もある．　　6）同じ意味で異なった字句（同義異表現）が数多くある場合，探索の精度が低くなる可能性がある．「事前にテキストを修正する」のは難しいので「処理の途中結果であるNグラム出現頻度のデータを修正する」ことで対応したいと考えている．

〈参考文献〉

出本充代（1995）「『撰集百縁經』の訳出年代について」，『パーリ学仏教文化学』8，99–108．干潟龍祥（1978）『本生経類の思想史的研究改訂増補版』，山喜房佛書林．師茂樹（2002）「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」予稿集』．内藤竜雄（1955）「大方便仏報恩経について」，『印度学仏教学研究』，3（2），695–697，梁麗玲（2002）『《賢愚經》研究』，法鼓文化事業．

〈キーワード〉　仏教説話，パラレル，統計解析，Nグラム，『賢愚経』，『撰集百縁経』，『大方便仏報恩経』

（九州大学大学院）