2005－CH－66（6）
2005／5／27

異本解析を目的としたオリジナル文書抽出モデルの考察

○三宅真紀，赤間啓之，馬越庸恭*，中川正宣
mmiyake@dp.hum.titech.ac.jp
東京工業大学社会理工学研究科
*東京工業大学学術国際情報センター

本研究では、異本文書から共通したオリジナル文書を抽出するモデルを考案し、新約聖書学の「共観福音書問題」に適用を試みた。そして、人工的に作成した文書（オリジナル文書）を用いて、異本文書からオリジナルの文書の抽出を確認し、提案したモデルの妥当性について検証した。

Model to Extract Original Sources from the Variant Documents

○Maki Miyake，Hiroyuki Akama，
Nobuyasu Makoshi*，Masanori Nakagawa
mmiyake@dp.hum.titech.ac.jp
Department of Human System Science, Tokyo Institute of Technology
* Global Scientific Information Center, Tokyo Institute of Technology

In this paper, we propose a quantitative model to extract original sources from the variant documents. The model is evaluated by simulating a process of intertextuality between the supposed original texts and their derived and modified versions. The first step is to create some artificial texts, which we assume to be dummy "original texts"(in abbreviation, Os), and some "variant documents"(Vs) by randomly distributing all the words contained in each of the Os. Secondly we try to see whether our proposed model is able to extract the traces of the Os in the form of factors, by applying factor analysis to the lexical frequency data gathered from all the subsets of the joint Vs. Based on the result of the simulation, we apply this model to the synoptic problem, which is about the genealogical interdependence between the Synoptic Gospels as one of the controversial subjects in the New Testament (NT) studies.

1　はじめに

本研究では、新約聖書学の「共観福音書問題」に対して、計量モデルを考案し、共観福音書の特徴を質的ばかりでなく量的に表現することを試みている。今回は、異本文書から共通したオリジナル文書を抽出するモデルを考案する。そして、人工的に作成したオリジナル文書を用いて、異本テキストを作成し、モデルの妥当性について検証する。そして、共観福音書の単語出現頻度データを用いて、提案したモデルを適用し、その分析結果から共観福音書の文書成立過程について、いくつかの歴史的仮説を検証していく。このようにして、聖書学の分野において、コーパス言語学的な統計解析を用いた方法論を確立することを目的としている。

2　背景

2.1　共観福音書

新約聖書の文学類型の一つに福音書がある。この文学類型は、キリスト教会において新しく作り出されたもので、宣教的意味を持つ。福音書には、マルコ、マタイ、ルカ、ヨハネ福音書の四文書がある。これらの福音書は、それぞれ別の著者によって書かれたものである。

これら四福音書のうち、マルコ、マタイ、ルカ福音書の三福音書については、互いに密接な類縁関係があり、三つの並行するフレームからなる対観表の形にあらわすことができるため「共観福音書」と呼ばれている[1]。

2.2　共観福音書問題

共観福音書を様々な共通単元のフレームで並べ換え、相互に同時比較できるようにしたものが「共観表」である。これは、共観福音書が相互にどのような文献的な依存関係があるかという、いわゆる「共観福音書問題」を議論する上で、重要な役割を担ってきた[2]。とくに、マルコ福音書を最古の資料とする福音書研究の定説に従い、三福音書の並行箇所をもとに、マタイ・ルカがマルコを引継・変更した形跡について古くから議論がなされてきた。また、マタイ・ルカのみに表れる並行箇所に注目し、マタイ・ルカの間で相互引用の関係がなかったという前提で、「幻の資料集」である「Q資料」が失われた文書として想定された。

新約聖書学においては、マタイ・ルカ福音書が、共通の資料としてマルコ福音書と「Q 資料」をそれぞれ用いたと考える「二資料説」が（図１）、共観表というフレームを用いて説明できると考えられてきた。この仮説は、長い間論議されつづけてきた「共観福音書問題」への最も説得的な解決法としてみなされ、現在の聖書学においては、ほぼ定説化している[3]。

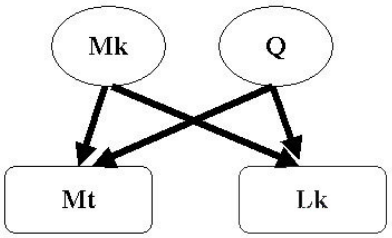

図 1：二資料説

2.3　仮説検証モデル

われわれは、共観福音書から得られた頻度データを７つのカテゴリーに分け、「二資料説」を計量的に説明するようなモデルを考案した[4]。まず。共観福音書の重複部分および独自部分は図１のよう示すことがで

き、テキストは７つのカテゴリーに分類することができる。３書共通部分（A)、マタイ・マルコ共通部分（B)、マルコ・ルカ共通部分（C)、マタイ･ルカ共通部分（D)、と､それらの共通部分を除いたマタイ(E)、マルコ（F)、ルカ（G）部分である。

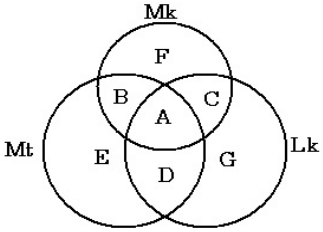

図 2：７つのカテゴリー

ここで､「二資料説」に準じてカテゴリーの特徴を説明すると、マタイ文書がマルコ文書を資料として扱った箇所が A+B 部分、またルカ文書がマルコ文書を資料として扱った箇所が A+C 部分に相当する。さらに、もう一つの資料であるマタイ・ルカ文書が共通して資料としていた Q 資料部分は、D 部分に相当し、マタイ・ルカ文書は、マルコからの資料部分（A+B+C）と Q（D）の２つが大きな割合を占めているとされている。他の諸仮説についても、カテゴリーの関係で表すことができる。このように、文書の成立上で用いたと考えられている資料が、分類した一つのカテゴリー、あるいは複数のカテゴリーの和によって説明されることから、カテゴリー間の類似度を計量的に分析することによって、仮説の検証が可能になる。

2.4　分析方法

分析データは、K.Aland の古典ギリシャ語の「共観表」[5]［6］を用いて、ルカに準じた並行箇所を採用した。ここで、並行箇所に含まれない部分については、それらをまとめて一つの並行箇所とした。そして、並行箇所単位で、出現する単語の頻度数をカウントし、７つのカテゴリーへ分配した。最後に、文書の正規化のために、相対頻度数を求めた。ここで、３つの分配モデル（分配・共観表・共通部分型）に従って、３タイプの頻度データを求めた。全ての並行箇所において出現した単語は 7276 語であり、以上の条件を満たした単語は 7099 語であった[7]。

ここでは、文書間に出現する単語の頻度数表から、上述した７つのカテゴリーへ単語と頻度を分配する３つの方法のうち、分配型モデルについて簡単に説明しておく[8]。具体的に、一つの並行箇所において、出現する単語の頻度数が表１のような場合、データセットの振り分け方法について説明する。

	Mt	Mk	Lk
単語 I	2	1	3
単語 II	3	0	2

表 1：頻度表

分配型セットは、各頻度数を共通部分カテゴリーの順に抽出する方法である。表１のような頻度数が得られたとき、表２のように７つのカテゴリーに振り分けられる。

	A	B	C	D	E	F	G
単語 I	1	0	0	1	0	0	1
単語 II	0	0	0	2	1	0	0

表 2：分配型モデル

さらに、分析データの相関行列の固有値

を基にして、因子数を４つに推定し、バリマックス回転を施し因子分析を行った。7つのカテゴリーについての各因子負荷量をそれぞれ表 3 に示す。ここで、各因子の因子量の絶対値が大きい値については、太字で表した。

	F1	F2	F3	F4
A	**0.41**	**0.72**	-0.11	0.30
B	0.15	**0.79**	0.45	0.06
C	0.16	0.17	0.18	**0.90**
D	**0.80**	0.24	-0.01	0.15
E	**0.70**	0.37	0.23	-0.12
F	0.15	0.14	**0.90**	0.19
G	**0.75**	0.03	0.21	0.37

表 3：因子負荷量

分析結果からは、A+B+C 部分と D 部分がそれぞれ独立した２つの因子を確認することが出来なかった。従って、想定した二資料説のモデルに当てはまらず、因子構造から他の成立過程の可能性を示唆するような結果が得られた。

3　オリジナル文書抽出モデル

これまで、われわれは「二資料説が真ならば、それに見合った因子が抽出さるだろう。ところがそれが見出されなかったので二資料説は偽である」という議論をしてきた。しかし、この背理法的アプローチにおける前提的仮言に関しては、今まで検証なしで天下り的に自然な議論として利用してきたという経緯がある。だが、その妥当性は、実際の数値例をもとに解析的に因子構造を明らかにしない限り、積極的に主張できるものではない。すなわち、もっと根本的に、実際にオリジナル文書を現存させたうえで、実験的に文書成立モデルの妥当性を示さない限り、今までの推論の根拠が薄弱であるという批判が予想される。

人文科学では、たしかに永久に失われてしまった文書に関しては間接的な推定作業しか可能ではない。しかし計量モデル論の立場からすると、そのような前提の曖昧な消極的背理法だけでは説得力が薄い。起源をあえて仮想の実体として人工的に作り上げ、シミュレーション実験にかけることで、二資料説批判に因子分析を導入する方法自体の有効性そのものを問い直す必要がある。

確かに、カテゴリーに所属する単語に着目したとき、それらの出現頻度をもとに計算されたカテゴリー間の相関係数は、そのカテゴリー間の計量的相関関係を表しているはずである。しかし、今回提案するオリジナル文書抽出モデルでは、単に実際の文書データから出発しボトムアップ-遡行的にその起源を探索し発見しようとするのではない。その方法論自体の妥当性を検証するため、なんらかの起源テキストからトップダウン-時系列的に引用関係のシミュレーションを行い、それを踏まえた後で、ふたたび、歴史の流れをボトムアップ-遡行するという往復的シミュレーションを行う。すなわち前もって文書の成立上で用いたと考えられている複数の低相関なオリジナル文書を用意しておき、カテゴリーへの単語配分をそれに則って行い、カテゴリー間の相関に基づくクラスタリングを行う。そのようして、オリジナル文書の存在を復元的に示すことにより、仮説の前提となる背理法的推論自体の妥当性を検証することが可能となる。

4　モデルの検証ー二資料説を基にして

われわれが考案した、オリジナル文書抽出モデルの適用性について、人工的に作成したテキストを用いて検証する。まず、二資料説を基にして、２つのオリジナル文書から、ランダムに分配されたテキストを分析データとし、オリジナル文書が抽出できるかどうかについて確認する。

4.1　オリジナルテキストの作成

まず、抽出する独立した２つのテキストを作成する。オリジナルのテキストの選定については、マタイ福音書の頻度データから出現頻度の上位 10000 単語を抽出して使用した。その頻度データを文書１（Original1）、文書１と正反対の頻度分布データを文書２（Original2）として使用し、２つのオリジナル文書を作成した。各テキストの出現頻度分布を図２に示す。ここで、テキスト間の相関係数は、0.02(p<5%)であり、それぞれ独立したテキストとしてみなすことができる。

4.2　テキストの分配

次に、4.1 で生成したオリジナル文書を７つのカテゴリーに分配する方法について説明する。

表６に、共観表の並行箇所ごとに、７つのカテゴリー分配モデルに従って振り分けた結果の、頻度数の割合について示す。この割合に準じて、オリジナル文書を７つのカテゴリーにランダムに分配する。このようにして得られた結果を、分析データをして用いた。

	分配型	共観表	共通部分
A	8	60	9
B	7	4	7
C	4	1	3
D	7	14	7
E	28	7	28
F	12	1	10
G	35	13	36

表 4：カテゴリー別の頻度数の割合

4.3　オリジナル文書の抽出

分析データから相関行列の固有値を求めると図２のようになる。図２のスクリープロットを基にして、因子数を２つに推定することができる。そして、バリマックス回転を施し因子分析を行った。ここで、第２因子までの累積寄与率は 92.6%であった。

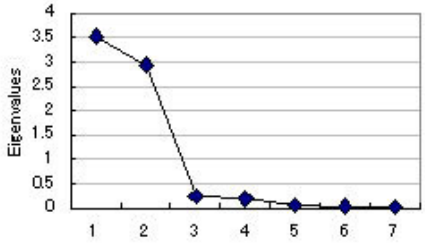

図 3：スクリープロット

	Factor1	Fctor2
A	**0.97**	−0.01
B	**0.95**	−0.01
C	**0.92**	−0.01
D	−0.01	**0.99**
E	−0.01	**0.99**
F	**0.91**	−0.01
G	−0.01	**0.99**

表 5：因子負荷量

バリマックス回転後の 7 つのカテゴリーについての各因子負荷量をそれぞれ表８に示す。ここで、各因子の因子量の絶対値が大きい値については、太字で表した。表８において、第 1 因子は、A,B,C,F 部分に大きな正の負荷量を持つ因子が抽出された。第 2 因子においては、D,E,G 部分が大きな正の負荷量を持っている。このように、オリジナル文書１（A+B+C+F）と文書２（D+E+F）がそれぞれ独立した因子として抽出され、２つのテキストが基となってテキストが形成されていることが分る。図７に抽出した因子の得点をプロットしたものを示す。これと、図６のオリジナル文書の頻度分布を比較すると、同様な分布であることが確認でき、抽出した因子が、オリジナル文書を再現していることがわかる。他のテキスト分配モデルの分析データに対しても、同様の結果が得られた。

5　モデルの検証－四資料説を基にして

5.1　四資料説

四資料説は、二資料説を拡張した仮説であり、Mk と Q 資料の他に、Lk と Mt がそれぞれ別々の資料（L,M）を参照して、四資料から形成されたという説である（図４）。

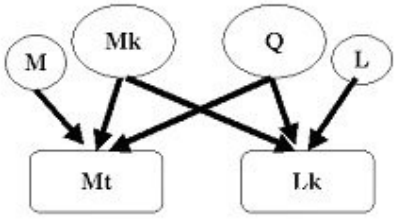

図 4：4 資料説

5.2　オリジナルテキストの作成

オリジナルのテキストの選定については、マタイ福音書の頻度データから出現頻度の上位 10000 単語を抽出して使用した。４つ独立したオリジナル文書を作成した。各テキスト間の相関係数を表 7 に示す。4.2 と同様に、３つの分配モデルに従って、オリジナル文書を７つのカテゴリーに分配した。

	Ori1	Ori2	Ori3	Ori4
Ori1	1	-0.08	-0.06	-0.02
Ori2	-	1	-0.02	0.01
Ori3	-	-	1	-0.01

表 6：テキスト間の相関係数

5.3　オリジナル文書の抽出

分析データから相関行列の固有値を求めると図２のようになる。図５のスクリープロットを基にして、因子数を４つに推定することができる。そして、バリマックス回転を施し因子分析を行った。ここで、第４因子までの累積寄与率は 72.2%であった。

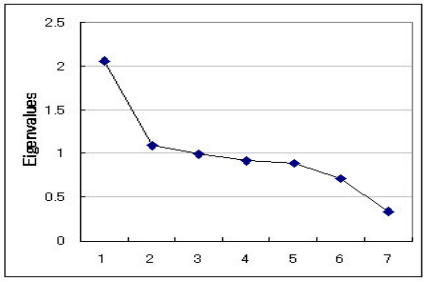

図 5：スクリープロット

バリマックス回転後の 7 つのカテゴリーについての各因子負荷量をそれぞれ表８に示す。ここで、各因子の因子量の絶対値が大きい値については、太字で表した。表８において、第 1 因子は、A,B,C,F 部分に大きな正の負荷量を持つ因子が抽出された。第

2,3,4 因子においては、G,D,E 部分がそれぞれ大きな正の負荷量を持っている。

	F1	F2	F3	F4
A	**0.87**	-0.11	-0.07	-0.06
B	**0.79**	-0.09	-0.07	-0.03
C	**0.66**	0.24	0.19	0.05
D	**0.92**	-0.13	**0.92**	-0.02
E	-0.03	-0.23	-0.01	**0.99**
F	**0.43**	-0.37	-0.32	-0.06
G	0.01	**0.89**	-0.15	-0.04

表 7：因子負荷量

このように、オリジナル文書１（A+B+C+F）、文書２（D）、文書３（E）、文書４（G）がそれぞれ独立した因子として抽出され、４つのテキストが基となってテキストが形成されていることが分る。図 9 に抽出した因子の得点をプロットしたものを示す。この図と、図 8 のオリジナル文書の頻度分布を比較すると、ほぼ同様な分布であることが確認でき、抽出した因子が、オリジナル文書を再現していることがわかる。他のテキスト分配モデルの分析データに対しても、同様の結果が得られた。

6　今後の課題

今回は、聖書学で立てられた有力な２つの仮説に対してモデルを作り、オリジナル文書抽出モデルの検証を行った。今後は、他の諸仮説についても同様なオリジナル抽出モデルを立て、計量的分析を試みる。特に、オリジナル文書に相互関係がある場合のモデルについて検討していく。

これらの計量分析結果をもとにして、聖書学で立てられた仮説とは別の成立方法についても提案していく予定である。

我々のオリジナル文書抽出モデルは、現在のところ共観福音書を対象として、その適用を検討した。このモデル発展させ、聖書に限らず、ソシュールの言語講義ノートのような異本からなる他のコーパスも対象にして、異本分析モデルについて考案してゆきたいと考える。

7　謝辞

本研究は、２１世紀 COE プログラム（研究拠点形成補助金）「大規模知識資源の体系化と活用基盤構築」の言語・文献、知識資源分野に関する研究の一環として行われたものである。

【参考文献】

[1]. Conzelmann, H. & Lindemann, A., Interpreting The New Testament, trans. by Siegfried S. Schatzmann, Hendrickson Publishes,45-53, 1988.

[2]. Theissen, G., *Das Neue Testament*, Beck, Mchn, 2002.

[3]. Kloppenborg, John S., et al. *Q Thomas Reader* , Polebridge Press, 1990

[4]. 三宅真紀、赤間啓之、佐藤研、中川正宣：因子分析による共観福音書問題の解析、*統計数理*、48 巻, 2 号, p.327-337, 2002.

[5]. Nestle-Aland, *Novum Testmentum Graece 26th edition*, German Bible Society Stuttgart

[6]. Kurt Aland, *Synopsis of the Four Gospels*, German Bible Society Stuttgart.

[7]. Miyake, M., AKAMA, H., Sato, M., Nakagawa, M., Makoshi, N., " Tele-Synopsis for Biblical Research", *Proceedings of the IEEE ICALT*, 931-935, 2004

[8]. 三宅真紀, 赤間啓之, 中川正宣,聖書ソフトウェアの開発と因子得点に基づく福音書の特徴分析, *文理シナジー学会誌,* 8 巻, 199-207, 2004

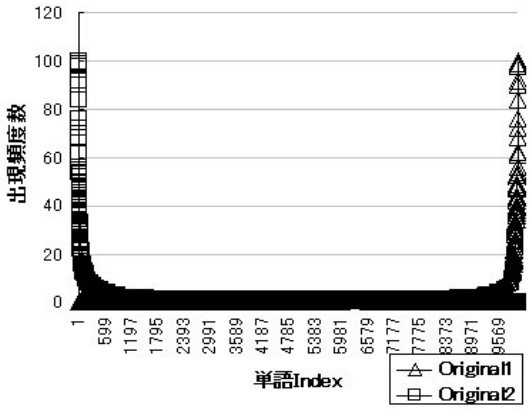

図 6：２オリジナル文書出現頻度

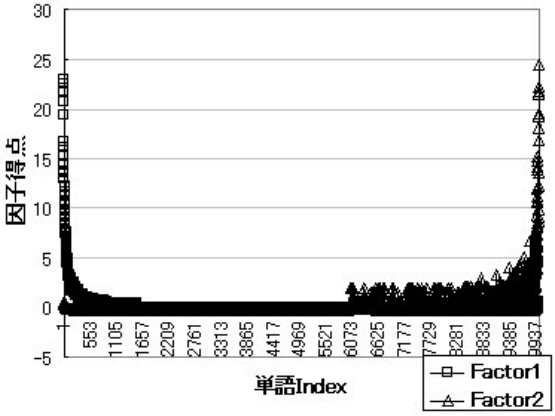

図 7：因子得点

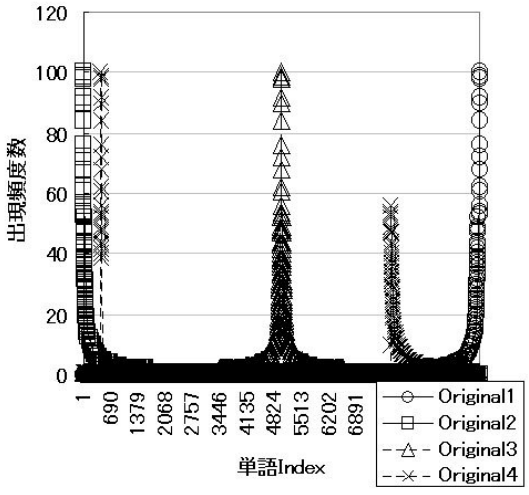

図 8：４オリジナル文書出現頻度

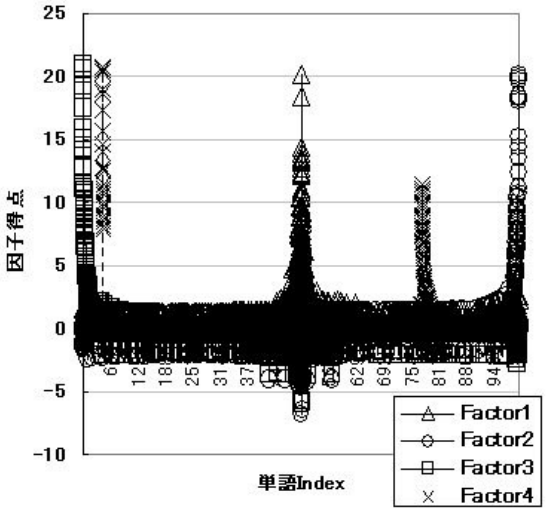

図 9：因子得点