社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－CH－63 (4)
2004／7／30

単語の共起データに基づく共観福音書の特有性の分析

三宅真紀，赤間啓之，中川正宣，馬越庸恭*

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、比較可能な文書間の類同性と特有性を分析するため、単語の共起データを利用する計量モデルを提案し、そのモデルを共観福音書に適用してマルコ、マタイ、ルカの間の隠れた関係が推定できるようにした。さらに、本研究においては、Mathematica・ WebMathematica を利用して、Tele-COEX という Web 対応の自然言語処理システムを開発したが、これはキーワードの近傍中の連関における意味効果を評価するため、悉皆的に単語の共起データを収集することができる。

Analysis of the Peculiarity of the Synoptic Gospels Using the Co-occurrence Data

Maki Miyake, Hiroyuki Akama,
Masanori Nakagawa, Nobuyasu Makoshi*

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 using a lexical co-occurrence data to analyze the similarity and dissimilarity among a set of parallel text collections and apply this model to the Synoptic Gospels so that we identify the hidden relationship among Matthew, Mark and Luke. The subgoal of our present study is to develop by Mathematica and WebMathematica a Web-based NLP (Natural Language Processing) system named "Tele-COEX", which allows us to gather lexical co-occurrence data to evaluate the semantic effects in the keyword neighborhood associations.

1　はじめに

本研究では、新約聖書中の共観福音書における主要単語の近傍で、いかなる単語が共起するか、悉皆的な計量データを取得し、そうした共起情報から各福音書の類同性・特有性を浮かび上がらせる計量モデルを考案、適用した。共観福音書全体を特徴付ける単語であっても、その周囲の単語の共起パターン次第では、各福音書で異なる意味用法を担い、まったく相違した文脈を形成する場合が予想されるからである。

われわれは、Mathematica のカーネルを利用した WebMathematica を開発言語として、単語の共起情報取得用のアプリケーション、Tele-COEX を作成、それを電子聖書のテキストに適用した。さらに、そこで取得された単語の共起情報をもとに因子分析をおこない、共観福音書の共通部分および各福音書の独自特徴について、新たな計量文献学的解釈を提起した。

2　背景

ここでは、これまで行われてきた、単語の共起データ分析について簡単に述べる。高山らは、共起行列の概念空間を特異値分解により縮退させた Word Space というシステムを開発し[1]、文書を特徴づけない不要語を取り除いた高頻度語である内容表現語の一つの周りで、何語か以内に共起する単語を収めた「ウィンドウ」を設定している。

赤間は、このウィンドイングの手法に想を得、情報検索学の基本概念であるベクトル空間モデルを計量文体論に導入し、哲学思想テキストから単語の共起行列を出力させ、それを因子分析にかけて複雑な文脈の個別抽出を行った[2]。

また清水（由）らは朝日新聞の「インターネット」「ネット」を含む見出しを対象に、新聞見出しがこのメディアをどう扱っているのかを、年毎に変化する高頻度共起語の因子分析結果から考察し、それを通して社会とインターネットというメディアとの関係の変化を辿った[3]。

さらに清水（正）らは、フランス語の非人称主語代名詞 “on”、およびそれと意味的に等価な定冠詞付きの “l'on” の二つについて、それらを使い分ける様々な基準を明らかにするため、文学作品・新聞コーパスの中から、“on” あるいは “l'on” の用法をその共起語と共に悉皆的に収集し、人工知能エンジン C5.0 を用い、共起語の条件や書き手の同一性がその選択にどのように関わってくるのか因果分析をおこなった[4]。

3　共起語ソフトウェア Tele-COEX

3.1　開発目的

前述した共起語情報によるコーパス言語学的研究は、Perl 言語で書かれた共起ウィンドウ制御スクリプト(coexcount)を用いて、共起語データを取得している。このスクリプトは、共起語が中心語に対し前置するか後置するかで別々にデータを取るなど、様々なオプションの追加が重なり、複雑なものになっていた。

そこで、われわれは Mathematica を利用し、体系的に様々なオプションを組み合わせたトータルな共起情報取得システム、Tele-COEX の開発に着手した。本システムは、オンライン操作が可能になり、現行ブラウザの文字コード処理能力を最大限に利用できるように、Web サイトにインタラクティブな計算機能を登載可能な WebMathematica を用いて開発中である。WebMathematica は、Mathematica のカーネルと、Java Servlet 技術に基づいて開発されたツールである。

3.2 仕様

つぎに、Tele-COEX の仕様を簡単に説明する。ユーザーが、任意の文書ファイルと中心語（キーワード）ファイルをシステムにアップロードすると、システムは指定された数幅の共起ウィンドウ内に出現する共起語の数をカウントする。図 1 に Tele-COEX の GUI を示す。本システムでは、共起情報をマトリックスの形で表示し、必要に応じて CSV ファイルに出力することができる。

現在の運用バージョンでは、実装しているオプションは以下の 4 つであるが、実験バージョンではさらに 4 個のオプションを追加している。

1). 文字コードの選択
2). ウィンドウ幅の指定
3). キーワードの頻度数で割る
4). 頻度数の Log 補正

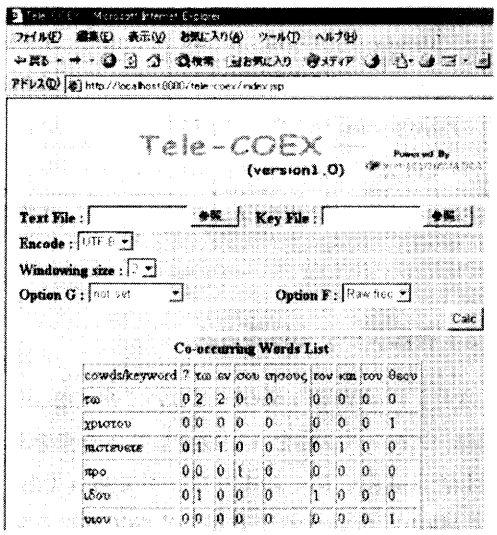

図 1：Tele-COEX の GUI

4 共観福音書

ここで、新約聖書の文学類型の一つ、福音書について述べる。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。それらは、様々な口伝伝承、文献資料を用いて叙述されており、イエスの登場・活動を描き、受難と復活で終わる。

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（以降、場合によりそれぞれ Mk,Mt,Lk と略す）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[4]。そして、この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（Synopsis）」である。これは、J.J.Griesbach が 1974 年に出版した『共観福音書対観表』においてはじめて用いた言葉であり[5]、現在の新約学では、Kurt Aland が作成したギリシャ語共観表が最も信頼性のある共観表として認められている[6]。

5 共観福音書の特徴分析

5.1 類同性・特有性抽出モデル

上述したように共観福音書は、構成・内容の点から一致部分が多く確認される。しかしながら、3 共通部分の他にも、2 文書間の共通箇所や各福音書の独自の編集部分も多く存在している。実際、共観表によって、マタイ・ルカの共通部分が認められ、そこから Q という資料の存在も想定された[7]。

そこで、共観福音書において、「福音書」の

共通特徴及び各福音書の独自特徴を検証するために、共起情報を用いて複数文書の類同性・特有性を抽出する計量モデルを立てた。

まずキーワードとなる中心語の近傍何語か以内に共起する単語をスキャンするため、一定枠の「ウィンドウ」を設定し、それぞれの文書（Mt,Mk,Lk の 3 文書）上をスライドさせ中心語のインスタンスごとにストップして、ウィンドウ内の共起情報を収集する。すべてのキーワードは 3 文書に同時出現し、文書タグを付加した別変数として扱うので、（出現単語，中心語＊3）型の共起行列が取得される。

得られたデータをもとに、因子分析を行うと、中心語の異なり数に近い数の因子が抽出されるはずである。そして、中心語が文書間の類似性を示す場合は、同一の単語、Keyword$i$（iはキーワード番号とする）、

Mt_Keyword i
Mk_Keyword i
Lk_Keyword i

の組が、最大の因子負荷量を持つ変数群としてペアを為す因子が抽出されるであろう。

それに対して、文書間の特有性を示す場合は、上記の中心語の 3 つの組の形が崩れて現れる因子が抽出されるはずである。

5.2 分析方法

5.2.1 中心語の選択方法

まず、中心語となる単語を選択するために、共観福音書の各々の単語の出現頻度数をオンライン聖書アプリケーション（Tele-synopsis）によって取得した [8]。ここで、頻度数は、福音書間の文章の長さの違いによって生ずる偏りを防ぐために、相対頻度数（頻度数／総頻度数）で計算した。得られたデータは、7276 語である。この (7276,3) 型行列を、因子分析し、相関行列の固有値を求め、因子数を 1 に推定した。そして、因子得点の降順に、名詞を３０語抽出し、中心語として選定した（表 1）。また、語彙の一致を厳密にして分析するため、単語は、形態が違う場合は、それぞれ別の単語として扱った。

表 1:キーワード

順位	単語	訳	順位	単語	英訳
1	ιησους	イエスは	16	πατηρ	父は
2	θεου	神の	17	υιον	息子を
3	υιος	息子	18	πετρος	ペトロは
4	ανθρωπου	人々の	19	ουρανων	天(複)の
5	μαθηται	弟子達は	20	γραμματεις	聖書を
6	βασιλεια	王国	21	θεος	神は
7	κυριος	主	22	βασιλειαν	王国を
8	κυριε	主よ	23	δαυιδ	ダビデの
9	ιησου	イエスの	24	χειρας	手を
10	κυριου	主の	25	ουρανου	天の
11	ιησουν	イエスを	26	πνευμα	聖霊を
12	γης	大地の	27	μαθηταις	弟子達に
13	ανθρωπων	人々の	28	οικον	家を
14	ημερα	日に	29	πατερα	父を
15	φαρισαιοι	ファリサイ人	30	ονοματι	名前に

5.2.2 共起データの取得方法

次に、共起情報の求め方について述べる。今回は、ウィンドウ幅を 5 に設定してデータを取得した。また、同一のキーワードであっても、Mt,Mk,Lk のいずれに登場するかで異なる限定タグを付け、それぞれ別の変数として扱った。すなわち、前述のキーワード数３０＊３＝９０を変数とした。

共観福音書（Mt,Mk,Lk）の全文を分析の対象としたが、例外として、定冠詞、前置詞、接続詞については、機能語とみなし除外した。また、文書の長さを基準化するために、中心語の頻度（ウィンドウのストップ回数）でそれぞれ共起語エントリーを割る操作を行った。

このようにして得られた共起語の異なり数は、2664 であった。

5.3 分析結果

この（2664,90）型行列のデータを、因子分析し、相関行列の固有値を求めた。スクリープロットを図 2 に示す。図 2 から、固有値 1 以上を基準にして、因子数を 29 に推定し、バリマックス回転を施して、因子分析を行った。

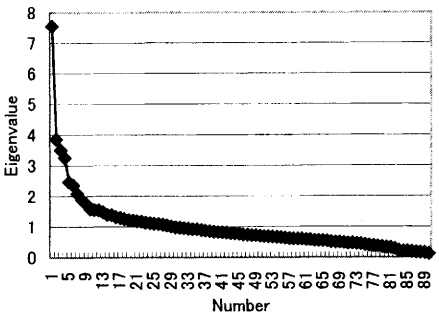

図 2:固有値のスクリープロット

紙面の都合上、バリマックス回転後の因子負荷量の一部を因子負荷量の大きい変数に限定して掲載する。それぞれの結果の解釈は、5.4 考察にまとめて掲載する。

1) 共通性(類同性)

まず、中心語が共通性(類同性)を表すような因子を示す。ここで、中心語 K*i* の番号 *i* は、表 1 中の順位に対応している。

（例：K1＝イエスは（ιησους））

■ 中心語の三つ揃

中心語が共通性(類同性)を表している因子について、表 2 に示す。

表 2:共通性(類同性)因子の因子負荷量

	因子1	因子7	因子9	因子10
Mt_K04	**0.84**	0.07	0.05	0.09
Mk_K04	**0.81**	−0.02	−0.05	0.07
Lk_K04	**0.83**	0.02	0.11	0.05
Mt_K14	0.01	0.11	**0.71**	0.01
Mk_K14	0.06	−0.07	**0.55**	−0.01
Lk_K14	0.17	0.17	**0.63**	−0.05
Mt_K23	0.45	0.05	0.02	**0.53**
Mk_K23	0.22	0.02	−0.07	**0.73**
Lk_K23	−0.02	−0.01	0.03	**0.72**
Mt_K29	−0.02	**0.79**	0.22	0.01
Mk_K29	−0.02	**0.80**	0.17	0.00
Lk_K29	0.18	**0.71**	−0.17	0.00

表 2 から、因子 1 については K4、因子 7 については K29、因子 9 については K14、因子 10 については K23 が、大きな正の負荷量を持っていることがわかる。

■ 形態の異なる同一中心語のペア

表 3 に、因子 2 の因子負荷量を示す。Mt,Mk,Lk の K3 と K17 が大きな正の負荷量を持っていることが分かる。

続いて、因子 18 と因子 3 の因子負荷量を表 4 に示す。因子 18 は M t の K6 と K22 が、因子 3 は Mk,Lk の K6 と K22 が大きな正の負荷量を持っていることが分かる。これらは、格が異なるが同一の形態素がペアを為すものである。

表 3:因子2の因子負荷量

中心語	因子2
Mt_K03	**0.72**
Mk_K03	**0.72**
Lk_K03	**0.66**
Mt_K17	**0.61**
Mk_K17	**0.75**
Lk_K17	**0.69**

表 4:因子 3, 18 の因子負荷量

中心語	因子3	因子18
Mk_K22	**0.87**	0.10
Lk_K22	**0.85**	0.08
Lk_K06	**0.79**	0.15
Mk_K06	**0.62**	0.13
Mt_K22	0.30	**0.81**
Mt_K06	0.15	**0.66**

因子 22 と 25 の因子負荷量を表 5 に示す。Mk,Lk のK9 と Lk のK11 が大きな正の負荷量を持っていることが分かる。また、Mt,Mk のK11 については、因子 25 に表れている。

表 5:因子 22, 因子 25

	因子22	因子25
Mt_K09	0.13	0.01
Mk_K09	0.37	−0.04
Lk_K09	**0.68**	−0.05
Mt_K11	−0.11	**0.75**
Mk_K11	0.23	**0.62**
Lk_K11	**0.53**	0.25

2) 特有性

次に、Mt,Mk,Lk の中心語が共通して現れなかった因子、すなわち、特有性を表す因子について示す。

■ **類似概念に対する単語の置換**

因子 4 の因子負荷量を表 6 に示す。Mk,Lk の K2 と Mt の K25 が大きな正の負荷量、また Mt,Mk,Lk の K6 が少し小さいがある程度の大きさの正の負荷量を持っていることが分かる。

表 6:因子4

中心語	因子4
Lk_K02	**0.81**
Mk_K02	**0.75**
Mt_K02	0.41
Mt_K19	**0.79**
Mt_K06	0.29
Mk_K06	0.26

■ 因子解釈上の相補的・対照的な関係

特有性を表している因子として最も興味深いのが因子 5、因子 6、因子 13 であった。その因子負荷量を表 7 に示す。

表 7:因子 5, 6, 13

中心語	因子5	因子6	因子13
Mt_K01	**0.45**	**0.43**	0.14
Mk_K01	0.35	**0.53**	−0.05
Lk_K01	**0.58**	−0.05	0.01
Mt_K05	0.00	0.06	0.39
Mk_K05	−0.01	0.13	0.12
Lk_K05	0.01	−0.01	**0.54**

Mt_K08	0.37	-0.09	**0.42**
Mk_K08	0.06	**0.54**	-0.17
Lk_K08	0.21	-0.02	**0.53**
Mt_K09	0.03	0.08	0.46
Mk_K09	0.29	0.29	0.16
Lk_K09	0.06	-0.02	0.18
Mt_K18	**0.80**	0.05	0.09
Mk_K18	**0.56**	**0.40**	-0.04
Lk_K18	**0.47**	-0.04	0.05
Mt_K27	-0.08	**0.53**	**0.45**
Mk_K27	-0.03	**0.67**	0.09
Lk_K27	0.22	0.02	0.03

5.4 考察

分析結果から、共観福音書の共通性（類同性）と、特有性について考察する。

1）共通性（類同性）

■ 中心語の三つ揃

表2の共通性を表す因子負荷量から、因子1は「人々」、因子7は「父」、因子14は「日」、因子10は「ダビデ」を意味している因子と考えられる。

■ 形態の異なる同一中心語のペア

表3から、K3とK17は形態こそ違うが、ともに「息子」という意味を持つ単語である。このことから、因子2は、「息子」を意味する因子として考えられる。

同様に表4から、K6とK22は形態こそ異なるが、ともに「王国」という意味を持つ単語である。このことから、因子3と18は、「王国」を意味する因子として考えられ、さらに、Mtが Mk,Lk とは「王国」という語の使い方が異なることが確認される。

また、表5から、因子22は、「イエス」を意味する因子として考えられ、さらに、Lkが Mt,Mk とは「イエス」という語の使い方が異なることが確認される。

以上から、形態素のグルーピングをした因子が抽出できることが確認できた。

2）特有性

■ 類似概念に対する単語の置換

表6から、因子4は、Mk,Lk の「神の」と Mt の「天の」、及び Mt,Mk,Lk の「王国」から特徴付けられていることが分る。

各福音書を参照すると、K2の「天の」をいう単語は、マタイに多く見られるが、Mk,Lkには一回しか使用されていないという、Mtに特有な単語であることが分かる。さらに、Mk,Lk では「神の国」として使用している部分に対して、Mtでは「天の国」と言い換えをしている事実も認められている[9]。このことから、因子4は「神の国」を表す因子で、語の言い換えに関する情報も同じ因子の中に含まれていることが確認できた。

■ 因子解釈上の相補的・対照的な関係

表7から、因子5は、Mtの K8 と Mt,Mk,Lk の K1、K18 が大きな正の負荷量を持っている。Mtの K8 を外して考えると、共通性（類同性）を表す、「イエスーペトロ」の関係を示すような因子として考えられる。Mtの K8 を考慮すると、「主ーイエスーペトロ」の関係を示すような因子としても考えられる。対して、因子6では、Mk の K8、K18 と Mt,Mk の K1、K18 が大きな正の負荷量を持っている。このことから、Mk の特有性を MtMk 共通部分に含んだ「イエス（主）ー弟子（ペトロ）」の二元関係を示すような因子として考えられる。対して、

因子13では、Lkの K8、K5が大きな正の負荷量を持っている。このことから、Lkの特有性を表す因子であり、「主－弟子」の二元関係を示すような因子とみなしうる。

以上の3つの因子から、共観福音書の「主－イエス－ペトロ」の関係を、マタイは、「イエス－弟子」の二元関係、また、マルコは、「イエス－ペトロ」といった二元関係に集約したと考えられる。さらにマルコに於いては、「イエス－ペトロ」の関係と「主－弟子」の関係は同等であるとも考えられる。対して、ルカでは、「主－弟子」の二元関係に簡略化された可能性も見られる。

6　まとめと今後の課題

本研究では、共観福音書の共起情報を用いて因子分析を行い、「福音書」の共通性、各福音書の独自特徴について考察した。

共通性（類同性）については、形態素のグルーピング、語の置き換えを示すような因子が抽出された。また特有性については、同じキーワードでも福音書によって使い方が違うことが確認された。

今後は、考察の特有性を表す因子を形成している中心語が実際にどのような文章で使用されているのか調べ、各文書の独自の特徴についてさらに言及する予定である。

さらに、共起語情報を取得するために開発したTele-COEXについては、オプションの追加や自然言語処理ツール(n-gramのカウント等)の導入を行い、汎用性のあるアプリケーションへと発展させる予定である。

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。また、Tele-COEXの開発にあたって、Mathematicaのアドバイスをいただいた東工大学術国際情報センターの松田裕幸先生に感謝します。

【参考文献】

[1] 高山, Flournoy, Kaufmann, Peters,単語の連想関係に基づく情報検索システム InfoMap, *情報学基礎* 53-1, 1999.

[2] 赤間啓之, ベクトル空間モデルに則った、近代ストア主義とメスメリスムの類似性に関する計量文体論的分析, *情報処理学会・人文科学とコンピュータ研究会 2001-CH-50*, 2001.

[3] 清水由美子, 大谷紀子, 赤間啓之, 同義語の意味の違いを測る－「インターネット」と「ネット」を例に－, *人工知能学会研究会資料 SIG-SLUD-A203-01*, p.1-6, 2003.

[4] 清水正勝, 赤間啓之, 清水由美子, 新聞コーパスの悉皆調査に基づくフランス語人称代名詞の使い分け基準について（on と l'on を例に）, *情報処理学会人文科学とコンピュータ研究会 2003-CH-60*, No.107, p.9~16 , 2003.

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

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

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

[8] 三宅真紀, 赤間啓之, 佐藤研, 中川正宣, 使用単語の因子得点に基づく福音書ジャンルの特徴考察, *文理シナジー学会平成 16 年度大会発表要旨集*, p.18, 2004.

[9] Hawkins, J.C., *HOAE SYNOPTICAE*, Oxford University Press, 1968.