テキストマイニングで作る人物相関図：河上肇『貧乏物語』を例に

林　直樹

はじめに

1917 年 3 月に弘文堂書房から刊行された河上肇『貧乏物語』は，前年 9 月から 12 月にかけて大阪朝日新聞に連載された記事をほぼそのまま採録した「貧乏物語」本編のほか，附録として「ロイド・ジョージ」論を収めている。本稿で扱う『貧乏物語』のテキストは青空文庫所収の岩波文庫版（1965 年改版）であり，その構成は上記原初版に忠実である。したがって，本稿における『貧乏物語』には「ロイド・ジョージ」論を含むという点を，まず明言しておきたい。

『貧乏物語』には管見のかぎり 94 名の人物が登場する（本稿末尾の一覧表を参照されたい）。彼らの相関図をテキストマイニングの手法を用いて描き出すためには，どのような手順を踏めばよいか。この点を具体的に明らかにすることが本稿の第一の目的である。第二の目的は，相関図生成の理論的な根拠を探ることを通じて，その手法上の限界を明らかにすることである。なお，テキストマイニング分析に際して援用したアプリケーションは，樋口耕一氏が開発した KH Coder である。

分析を始める前に，『貧乏物語』の本文中表記に基づく人名一覧のテキストファイルを作成し，それを「タグ」として上記アプリに読み込ませ，マイニング対象の本文テキストから人名だけを強制抽出せねばならない（KH Coder の操作…語の取捨選択＞品詞による語の選択：タグ＞強制抽出する語の指定：ファイルから読み込み＞前処理の実行）。上記 94 名から著者河上肇と「父」および「舎弟」の 3 名を省き，二人のスミスを区別できないため両者を「スミス」で括り，同じくマルクスと「マルクスの妻」および「マルクスの父」を「マルクス」で括ると，タグとして登録される人名から 6 名分がマイナスとなる。また，本文中表記が揺れる峨山昌禎は「峨山」の名で抽出した。そうすることで，相関図の構成要素として 88 の人名が残る。以上が，本稿で必要とした準備作業である[+]。

共起ネットワークによる分析結果

KH Coder（3.Beta.06a を用い，その後リリースされた 3.Beta.07b で再チェックした）の「共起ネットワーク」を用いてテキストの分析を行う。集計単位となる文書，いわゆる「言葉の袋」を段落（全 460）に指定すると，共起関係の強いグループが 15 個生成され，関連付けられた人名は 55/88 件であった。集計単位を文（全 1609）に指定すると，グループ生成数は 14，関連付け人名は 41/88 件となった。共起関係の測定にはいくつかの係数を利用可能だが，Jaccard と Simpson と Cosine のいずれについても，生成されるグループの組成は同じである（Dice は Jaccard の亜種と見て省略する）。しかし共起関係の強さの値には小さくない差が見られる。次ページ以降の図 1 から図 3 に，集計単位を段落および文に指定した場合の Jaccard 係数，Simpson 係数，Cosine 係数の値を順に示しておいた。

[+] 本稿の内容は，尾道市立大学経済情報学科の演習科目「特別演習 IV」（木村文則准教授と筆者との共同担当）ですでに取り扱った題材に基づいている。木村氏および演習受講生諸君のご協力に感謝する。また本稿のドラフトは，2022 年 12 月 10 日に福岡大学で開催された経済学史学会西南部会第 133 回例会で発表済みである。本論中で後述するように，その際，金子創氏（大分大学）と田中秀臣氏（上武大学）から鋭い質問を頂戴した。両氏に厚く御礼を申し上げるとともに，同日の参加者各位に改めて感謝申し上げたい。

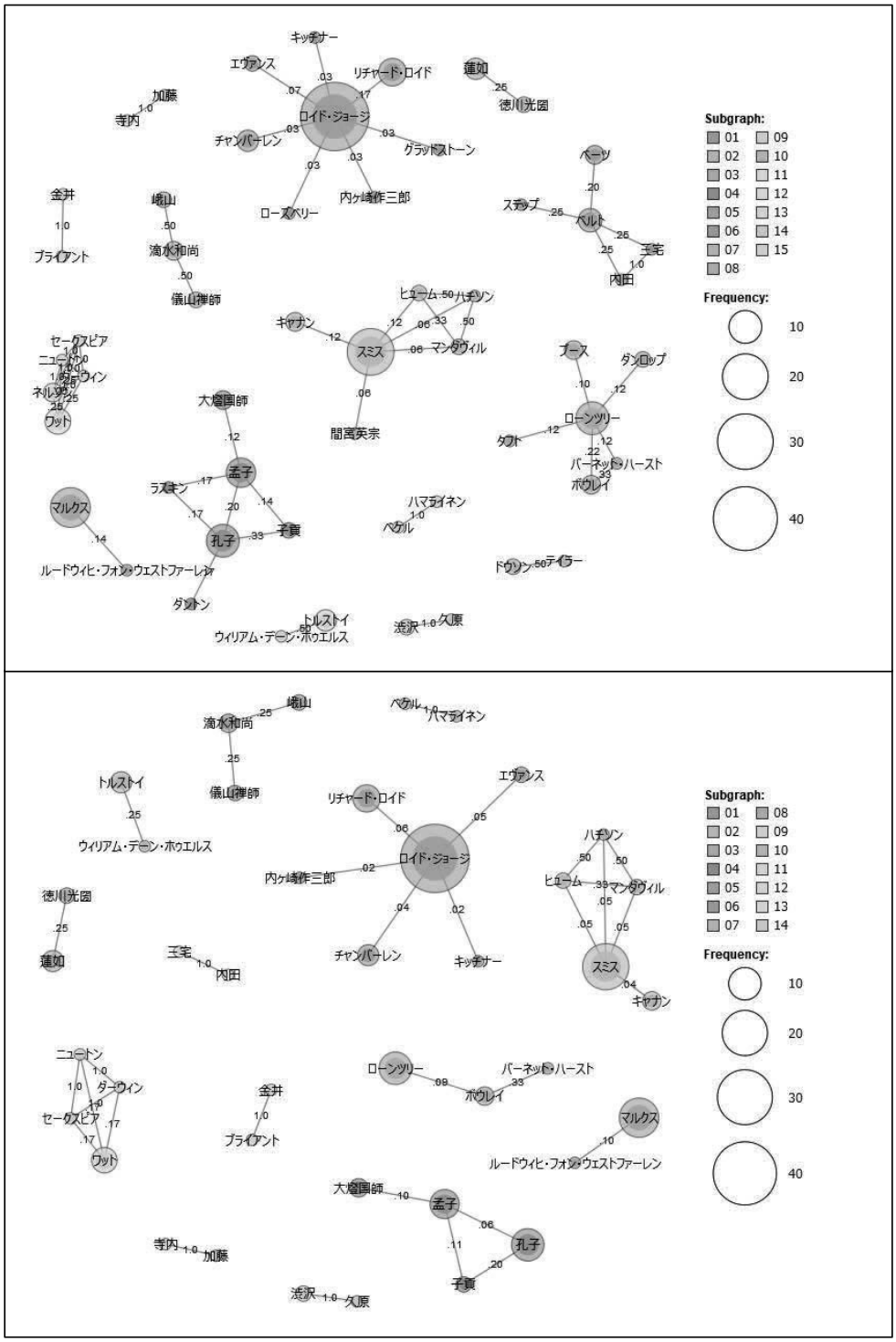

図 1　上段：Jaccard【段落】　下段：Jaccard【文】

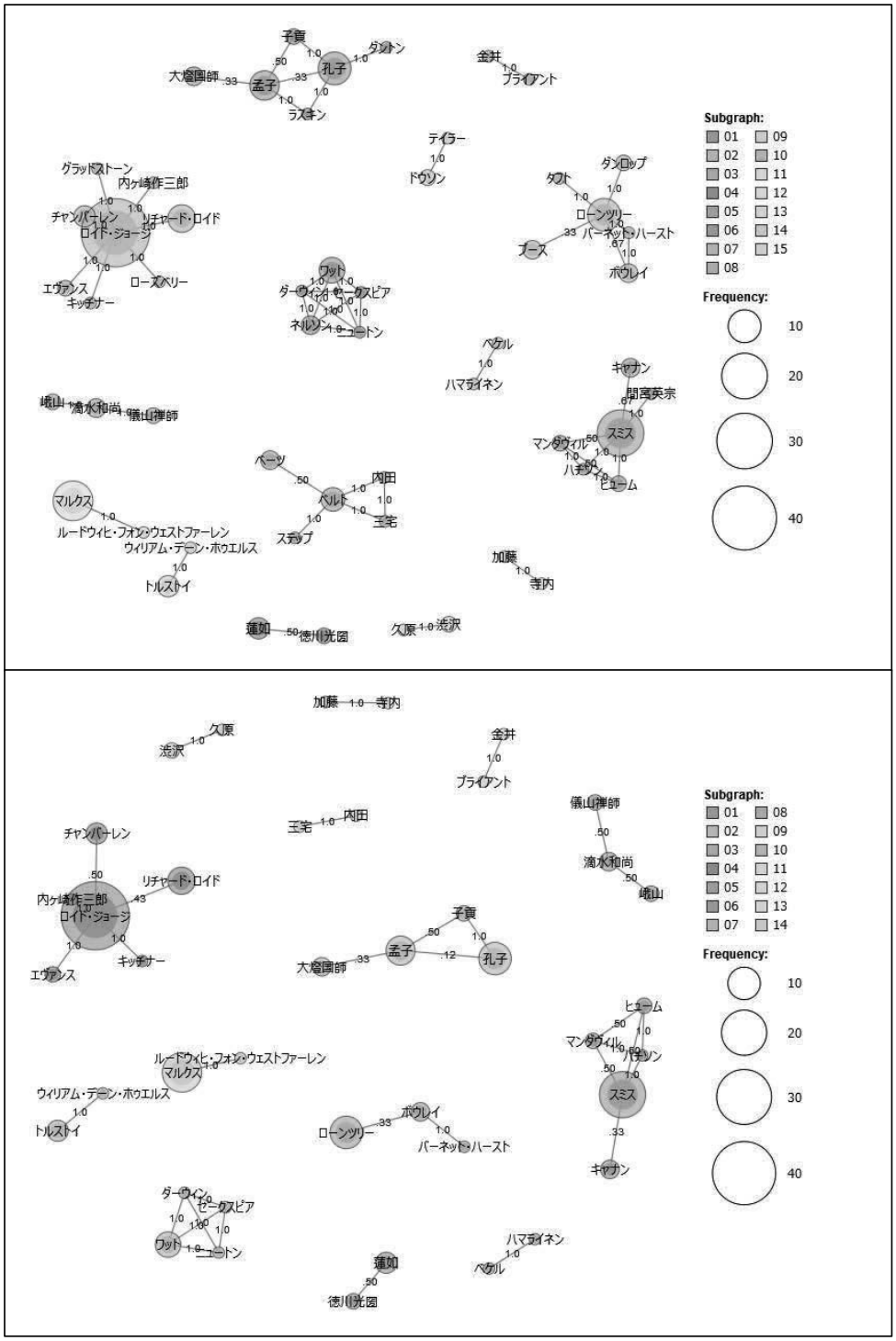

図２　上段：Simpson【段落】　下段：Simpson【文】

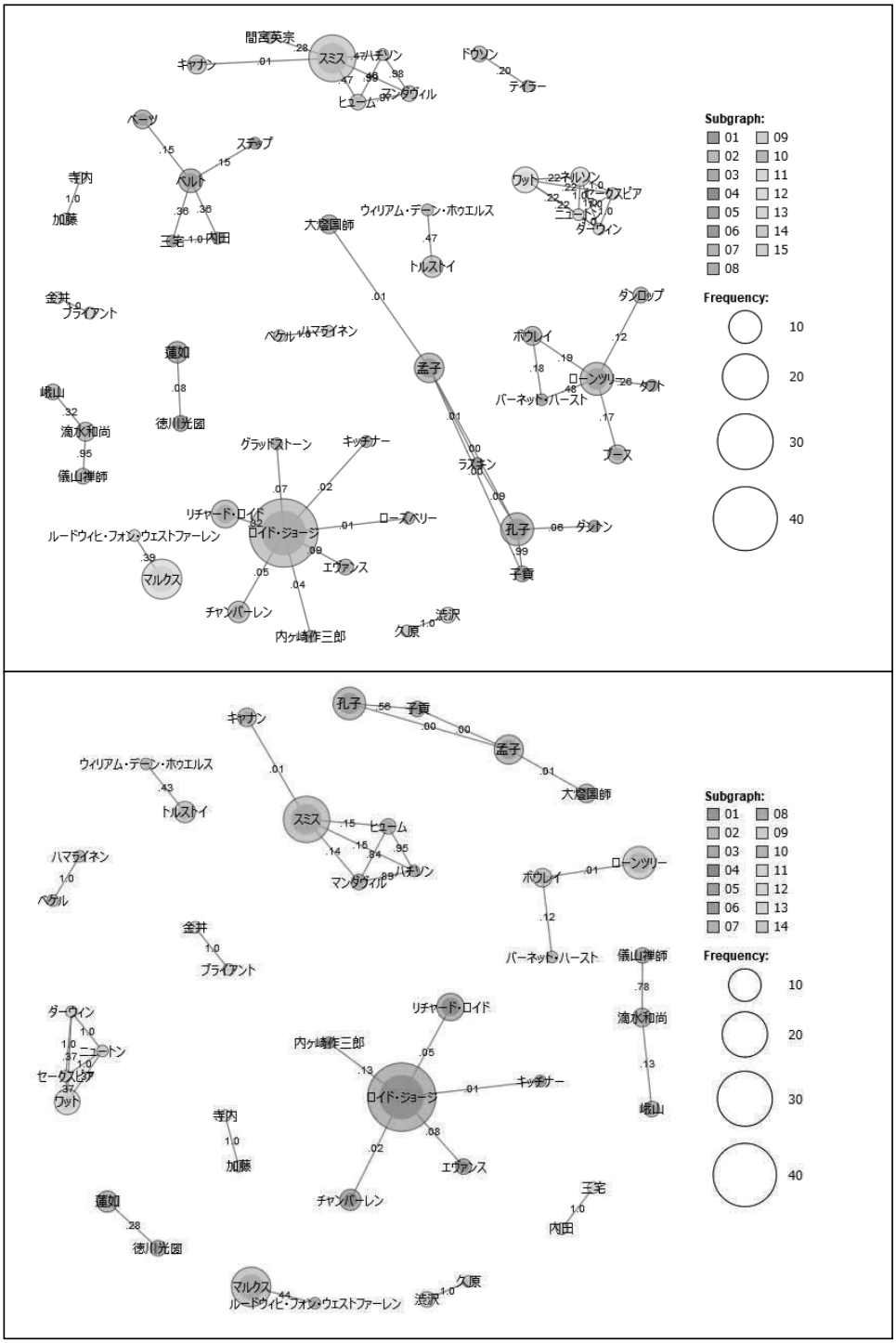

図 3　上段：Cosine【段落】　下段：Cosine【文】

Jaccard と Simpson の値の違いがなぜ生じるかを，具体例を挙げて検証してみよう。ここではトルストイとウィリアム・デーン・ホゥエルスのペアに着目する。『貧乏物語』では，両者は「三の二」節に登場している。以下は登場箇所からの引用である（なお読点「、」は青空文庫のテキストに従う）。

「①私は近ごろウィリアム・デーン・ホゥエルスに会うてトルストイを訪問したことを話したら、氏は次のごとく述べられた。『②トルストイのした事は実に驚くべきものである。それ以上をなせというは無理である。最も高貴なる祖先を有する一貴族としては、遊んでいて食わしてもらうことを拒絶し、自分の手で働いて行くことに努力し、つい近ごろまでは奴隷の階級に属していた百姓らとできうる限りその艱難辛苦を分かって行こうとした事が、彼のなしあとうべき最大の事業である。しかし彼が百姓らとともにその貧乏を分かつという事は、これは彼にとって到底不可能である。③何ゆえというに、貧乏とはただ物の不足をのみ意味するのではない、欠乏の恐怖と憂懼、それがすなわち貧乏であるが、かかる恐怖はトルストイの到底知るを得ざるところだからである。』……」

④げに露国の一貴族としてその名を世界にはせしトルストイにとっては、自発的貧乏のほか味わうべき貧乏はあり得なかったのである。

まずは集計単位として文を選んだ場合から検討しておく。人名が登場する各文に番号①～④を振ってみた。Jaccard であれば，トルストイ「または」ホゥエルスが登場する文の数が分母になるため，4 である。他方 Simpson であれば，最少頻度の人名が登場する文の数が分母になるため，ここではホゥエルスが現れる文の数 1 である。また分子は，どちらの係数の場合でもトルストイ「かつ」ホゥエルスが登場する文の数だから 1 となる。以上をふまえると，トルストイとホゥエルスの共起関係の強さ（相関係数）を Jaccard で求めた値は 1/4＝0.25，Simpson で求めた値は 1/1＝1.0 と分かる。

また，集計単位を段落にした場合，トルストイは連続する二つの段落の両方に，ホゥエルスは最初の段落のみに現れる。トルストイ「∪」ホゥエルスの段落数は 2，トルストイ「∩」ホゥエルスの段落数は 1 である。よって Jaccard の値は 1/2＝0.5 となる。Simpson については，最少頻度の人名は 1 段落にしか登場しないホゥエルスだから，分母の段落数は 1（分子も上記の通り 1）となり，係数値は 1/1＝1.0 と求まる。

KH Coder が弾き出した値もこれらと同一であることを，先の図で確認してほしい。

コサイン類似度をめぐる問題提起

Cosine の値も検証してみよう。集計単位は文とし，再びトルストイ＆ホゥエルスに焦点を当て，各文における出現頻度（回数）を要素とする次の単語ベクトルを作る。

	文①	文②	文③	文④
トルストイ（T）	1	1	1	1
ホゥエルス（H）	1	0	0	0

T（1, 1, 1, 1）および H（1, 0, 0, 0）のコサイン類似度を計算すると，$|T| = \sqrt{1^2+1^2+1^2+1^2} = 2$ $|H| = \sqrt{1^2+0^2+0^2+0^2} = 1$ $T \cdot H = 1\times1+1\times0+1\times0+1\times0 = 1$ だから $\mathrm{Sim}(T,H) = \frac{1}{2\times1} = 0.5$ となる。これは KH Coder の算出した値 0.43（図 3 参照）と一致しない。

ここで TF-IDF 概念を導入する（小峯 2021, 35）。TF は集計単位（いまは文）当たりの総単語出現頻度に占める特定単語の出現頻度を示す。したがって，上の表は

	文①	文②	文③	文④
T	1/2＝0.5	1/1＝1	1/1＝1	1/1＝1
H	1/2＝0.5	0	0	0

と書き換えられる。続いて IDF は，集計単位となる文書の総数を分子，特定単語が含まれる文書数を分母とした分数の対数に 1 を加えた値である。したがって，トルストイの場合はすべての文に出現していることから log 4/4＋1＝1，ホゥエルスの場合は 4 文中 1 文にしか出現していないため log 4/1＋1＝2.386 となる。TF-IDF は TF 値に IDF 値を掛け合わせたものだから，再び表を用いて表すと，トルストイとホゥエルスの TF-IDF 値は

	文①	文②	文③	文④
T	0.5	1	1	1
H	1.193	0	0	0

以上よりコサイン類似度を再計算すると，$|\mathrm{T}| = \sqrt{0.5^2 + 1^2 + 1^2 + 1^2} = 1.8$ $|\mathrm{H}| = 1.193$ $\mathrm{T}\cdot\mathrm{H} = 0.5 \times 1.193 = 0.5965$ だから $\mathrm{Sim}(\mathrm{T},\mathrm{H}) = \frac{0.5965}{1.8\times1.193} = 0.27735$ となる。0.43 からさらに乖離してしまった。

相互情報量（岡崎 2016, 53）をふまえた条件付き確率で計算してみよう。単語ベクトルの各要素値を ${}^{Pij}/_{Pi\times Pj}$（Pij は行 i と列 j の同時出現確率，Pi は行 i の出現確率，Pj は列 j の出現確率。これらの確率の分母は単語の総出現頻度，分子は各行各列当たりの単語の出現頻度）で表す。これは，各行各列に対する相互の条件付き確率を重ねたものに相当する。i にトルストイとホゥエルスを，j に文①～④を入れて計算すると，

	文①	文②	文③	文④
T	0.2/(0.8×0.4)	0.2/(0.8×0.2)	0.2/(0.8×0.2)	0.2/(0.8×0.2)
H	0.2/(0.2×0.4)	0	0	0

T（0.625, 1.25, 1.25, 1.25）と H（2.5, 0, 0, 0）のコサイン類似度を計算すると，$|\mathrm{T}| = \sqrt{0.625^2 + 1.25^2 + 1.25^2 + 1.25^2} = 2.253$ $|\mathrm{H}| = 2.5$ $\mathrm{T}\cdot\mathrm{H} = 0.625 \times 2.5 = 1.5625$ だから $\mathrm{Sim}(\mathrm{T},\mathrm{H}) = \frac{1.5625}{2.253\times2.5} = 0.27735$ となる。TF-IDF の場合と同値であり，したがって 0.43 とやはり一致しない。

試みに行列の特異値分解を行ってみても（*ibid.*, 54-57），トルストイ・ベクトルとホゥエルス・ベクトルのコサイン類似度は最初に求めた結果と変わらない。近似値を求めるための分解なのだから当然である。

$$\begin{pmatrix}1 & 1 & 1 & 1\\ 1 & 0 & 0 & 0\end{pmatrix} = \begin{pmatrix}-0.2898 & -0.9571\\ 0.9571 & -0.2898\end{pmatrix} \times \begin{pmatrix}0.8350 & 0\\ 0 & 2.0743\end{pmatrix} \times \begin{pmatrix}0.7992 & -0.3470 & -0.3470 & -0.3470\\ -0.6011 & -0.4614 & -0.4614 & -0.4614\\ 0 & 0.8165 & -0.4082 & -0.4082\\ 0 & 0 & -0.7071 & 0.7071\end{pmatrix}$$

$$= \begin{pmatrix}-0.2898 & -0.9571\\ 0.9571 & -0.2898\end{pmatrix} \times \begin{pmatrix}0.8350 & 0\\ 0 & 2.0743\end{pmatrix} \times \begin{pmatrix}0.7992 & -0.3470 & -0.3470 & -0.3470\\ -0.6011 & -0.4614 & -0.4614 & -0.4614\end{pmatrix}$$

$$= \begin{pmatrix}0.999957 & 0.999991 & 0.999991 & 0.999991\\ 1.000042 & 0.000025 & 0.000025 & 0.000025\end{pmatrix}$$

$|\mathrm{T}| = \sqrt{0.999957^2 + 0.999991^2 + 0.999991^2 + 0.999991^2} = 1.999965$

$|\mathrm{H}| = \sqrt{1.000042^2 + 0.000025^2 + 0.000025^2 + 0.000025^2} = 1.000042$

$\mathrm{T}\cdot\mathrm{H} = 0.999957 \times 1.000042 + 0.999991 \times 0.000025 + 0.999991 \times 0.000025 + 0.999991 \times 0.000025 = 1.000074$

よって $\mathrm{Sim}(\mathrm{T},\mathrm{H}) = \frac{1.000074}{1.999965\times1.000042} = 0.5$ となる。

徳川光圀と蓮如ではどうか。集計単位を文とすると，

①蓮如上人御一代聞書にいう「御膳を御覧じても人の食わぬ飯を食うよとおぼしめされ候と仰せられ候」と。

②徳川光圀卿が常に紙を惜しみたまい、外より来る書束の裏紙長短のかまいなくつがせられ、詩歌の稿には反古の裏を用いたまいたる事はよく人の知るところである。

③また蓮如上人御一代聞書を見ると、「蓮如上人御廊下を御通り候うて、紙切れの落ちて候いつるを御覧ぜられ、仏法領の物をあだにするかやと仰せられ、両の御手にて御いただき候としかじか、総じて紙の切れなんどのようなる物をも、仏物とおぼしめし御用い候えばあだに御沙汰なく候うの由、前々住上人御物語候いき」という記事がある。

④徳川光圀卿の惜しまれた紙、蓮如上人の廊下に落ちあるを見て両手に取っていただかれたという紙、その紙が必要品たるに論はないけれども、いかなる必要品でも使いようによっては限りなくむだにされうるものである。

以上 4 文が対象となる。なお，先のトルストイとホゥエルスの場合とは異なり，これら 4 文は連続していない。複数の段落内に散らばって存在している点を注記しておく。さて，同じく出現頻度を要素とする単語ベクトルを構成すれば

	文①	文②	文③	文④
徳川光圀（TM）	0	1	0	1
蓮如（R）	1	0	2	1

TM（0, 1, 0, 1）および R（1, 0, 2, 1）のコサイン類似度を計算すると，$|\mathrm{TM}| = \sqrt{0^2+1^2+0^2+1^2} = 1.41$ $|\mathrm{R}| = \sqrt{1^2+0^2+2^2+1^2} = 2.45$ $\mathrm{TM}\cdot\mathrm{R} = 0\times1+1\times0+0\times2+1\times1 = 1$ だから $\mathrm{Sim}(\mathrm{TM},\mathrm{R}) = \frac{1}{1.41\times2.45} = 0.289$ となる。こちらは KH Coder の算出値 0.28 にきわめて近い。

次に TF-IDF で計算すると，徳川光圀の IDF は log 4/2＋1＝1.69，蓮如の IDF は log 4/3＋1＝1.29 だから，

	文①	文②	文③	文④
TM	0	1/1×1.69	0	1/2×1.69
R	1/1×1.29	0	2/2×1.29	1/2×1.29

TM（0, 1.69, 0, 0.845）と R（1.29, 0, 1.29, 0.645）のコサイン類似度は $\mathrm{Sim}(\mathrm{TM},\mathrm{R}) = \frac{0.545}{1.889\times1.935} = 0.14907$ である。また条件付き確率の重ね合わせで計算すると，

	文①	文②	文③	文④
TM	0	0.167/(0.333×0.167)	0	0.167/(0.333×0.333)
R	0.167/(0.667×0.167)	0	0.333/(0.667×0.333)	0.167/(0.667×0.333)

TM（0, 3.003, 0, 1.506）と R（1.499, 0, 1.499, 0.752）のコサイン類似度は $\mathrm{Sim}(\mathrm{TM},\mathrm{R}) = \frac{1.133}{3.359\times2.249} = 0.14907$ である。TF-IDF の場合と再び同値だが*，0.28 からは大きく乖離した。

さらに，ウェストファーレンとマルクス（マルクスの細君を含む）で検証してみよう。全部で 10 文あり，先の光圀と蓮如の場合同様，登場順に並べるが，これらの文は必ずしも連続して現れては来ない点に注意されたい。

①否むしろ私は人並み一倍、経済の人心に及ぼす影響の甚大なるものなることを認めつつある者の一人で、その点においては私は十九世紀の最大思想家の一人たるカール・マルクスに負うところが少なくない。

②今私はここにマルクスの伝記をくわしくお話しする余裕ももたなければ、またその必要も感じない。

* 同値なのは偶然ではない。ここでは直観的に明らかなことのみを記しておこう。行列の成分を便宜上セルと，単語の出現頻度を簡潔に単語数と表現するならば，列当たりの単語数（A と置く）に対する当該セルの単語数（B）の比率は，列当たりの単語数（A）を総単語数（N）で割った商に対する，当該セルの単語数（B）を総単語数（N）で割った商の比率に等しい（B/A＝B/N÷A/N）。B/A が TF に対応し，B/N が Pij に，A/N が Pj に対応すると分かる。総単語数（N）に対する行当たりの単語数（C）の比率（C/N）すなわち Pi の場合，その逆数（N/C）をとったとしても概念が異なるため IDF と直接には対応せず，よって値も異なるが，IDF が同行の各セルにおける TF に等しく掛け合わされるのと同様，N/C も同行の各セルにおける B/A に等しく掛け合わされるため，各行を単語ベクトルとして見た際に，その大きさに違いは生じても方向性に変化は起こらない。コサイン類似度はベクトルの方向性だけを評価するものだから，TF＝B/A の要素で決まるベクトルの方向性さえ一致していれば同値になるのである。林（2022, 108-9）も参照されたい。

③すなわちもし諸君が許さるるならば、私はマルクス伝の一鱗を示すがために、ここにマルクスの細君の手紙の一節を抄訳しようと思う。

④これはマルクスの細君が一八四九年にある人に与えた手紙の一節であるが、ここにマルクスの細君というは、マルクスの父の親友なるルードウィヒ・フォン・ウェストファーレンという人の娘である。

⑤当時その人がプロシャの官吏としてザルツウェーデルという所からマルクスの郷里のトリエルに転じて来たのは、今からちょうど百年前の一八一六年のことであるが、その時に連れていた二歳になる女の子は、後にマルクスの細君となった人で、すなわち先に掲げた手紙の主である。

⑥この手紙の主は幼にして容色人にすぐれ､かつ富裕なる名家に人となりしがために、名門の子弟の婚を求むる者も少なくなかったのであるが、たまたまマルクスのせつなる望みにより、四歳年下のこの貧乏人の子にとつぎ、かくてこの女は、かの恐るべき社会主義者として早くより自分の祖国を追い出され、またフランスからもベルギーからも追放されて、ついには英京ロンドンに客死するに至りしところの、世界の浪人にしてかつ世界の学者たるカール･マルクスにその一生をささげ､つぶさに辛酸をなめ尽くしつつ、終始最も善良なる妻として、その遠き祖先の骨を埋めつつある英国に流れ渡り、ついに自身もロンドンの客舎に病死するに至りし人である。

⑦さて私がここにマルクスを持ち出したのは、彼が有名なる唯物史観または経済的社会観という一学説の創設者であるからである。

⑧右はマルクスのごう牙な文章を――しかもわずかにその一節を――直訳したのであるから、これを一読しただけでは充分に彼の意見を了解することは困難であるが、今これを詳しく解説しているいとまはない。

⑨これがマルクスの意見のだいたいである。

⑩今私はマルクスの議論をたどってそれを一々批評して行くというようなめんどうな仕事をばここでしようとは思わぬ。

W をウェストファーレンの，M をマルクスの単語ベクトルとすると，

	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩
W	0	0	0	1	0	0	0	0	0	0
M	1	1	2	3	2	2	1	1	1	1

W（0, 0, 0, 1, 0, 0, 0, 0, 0, 0）および M（1, 1, 2, 3, 2, 2, 1, 1, 1, 1）より $|W| = \sqrt{1^2} = 1$　$|M| = \sqrt{1^2 + 1^2 + 2^2 + 3^2 + 2^2 + 2^2 + 1^2 + 1^2 + 1^2 + 1^2} = 5.196$　$W \cdot M = 1 \times 3 = 3$　したがって $\mathrm{Sim}(W, M) = \frac{3}{1 \times 5.196} = 0.577$となる。KH Coder による値 0.44 との差は小さくない。ウェストファーレンの IDF は log 10/1＋1＝3.3，マルクスの IDF は 1 より，TF-IDF は

	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩
W	0	0	0	0.25×3.3	0	0	0	0	0	0
M	1/1	1/1	2/2	3/4	2/2	2/2	1/1	1/1	1/1	1/1

だから，W（0, 0, 0, 0.825, 0, 0, 0, 0, 0, 0）および M（1, 1, 1, 0.75, 1, 1, 1, 1, 1, 1）のコサイン類似度は $\mathrm{Sim}(W, M) = \frac{0.61875}{0.825 \times 3.09} = 0.24254$である。条件付き確率での計算過程は省略するが，W（0, 0, 0, 4, 0, 0, 0, 0, 0, 0）および T（1.067, 1.067, 1.067, 0.8, 1.067, 1.067, 1.067, 1.067, 1.067, 1.067）より両単語ベクトルのコサイン類似度は $\mathrm{Sim}(W, M) = \frac{3.2}{4 \times 3.299} = 0.24254$，よって TF-IDF と再三同値かつ，0.44 からは大幅に乖離する。

問題解決までの紆余曲折

KH Coder による Cosine 値に手計算値とのズレ，誤差が生じる理由は何だろうか。すで

に確認したように，TF-IDF および条件付き確率を用いないかぎり徳川光圀と蓮如のペアだけが乖離を生まなかった。裏を返せば，光圀と蓮如を含むいずれのペアについても，TF-IDF 等は KH Coder における Cosine 算出には関わりがないと考えられる。光圀と蓮如のペアが他の二つのペアと異なるのは，両者の単語ベクトルがいずれも 0 より大きな要素値を二つ以上有している点である。果たしてこの点が問題解決の糸口になるだろうか。

滴水和尚・儀山禅師・峨山（禅師 or 和尚）の三者内で組まれた，二つのペアを取り上げて検討してみよう。「十二の五」節のひと続きの箇所に，三者は揃って登場する。当該人名を含まない文には下線を引いた。

①峨山禅師言行録にいう「侍者師の室前なる水盤の水を替えけるに、師はそのそばにありて打ち見やりたまいしが、おもむろに口を開き、なんじも侍者となりて半年もたつから、もう気がつくだろうと思っていたが、言っておかぬと生涯知らずに過ごす。物はなァ、大は大、小は小と、それぞれ生かして使わねばならぬ。水を替える時は元の水をそこらの庭木にかけてやるのさ。それで木も喜ぶ、水も生きたというものだ。因地の修行をするものは、ここらが用心すべきところだ。また洗面の水なども、ざっと捨てずに使うたあまりは竹縁に流して洗うのだ……。うむ水一滴もそれで死にはせぬ、皆生きて働いたというものだ。陰徳陰徳と古人がたがやかましく言うのもほかではないぞ」。②水一滴もむだにしてはならぬという這般の消息になると、もはや経済論の外に出た話で、本来はこの物語の中に採録すべき記事ではないのであるが、私は事のついでに峨山和尚のお師匠に当たる滴水和尚の逸話をもここに簡単にしるしておこうと思う。

③滴水和尚かつて曹源寺の儀山禅師に師事されいたるころのことである。④ある日禅師風呂にはいられると、熱すぎるので、滴水和尚を呼んで水を運ぶことを命ぜられた。⑤そこで和尚は何心なくそこにあった手おけを取って、その底にわずかに残っていた一すくいの水を投げ捨てて立ち去ろうとせらるると、浴槽に浸りおられたる儀山禅師、その刹那に大喝一声、ばかッとどなられた。

単語ベクトルを組むと，三者ともに二つ以上の要素値において 0 を上回ることが分かる。

	文①	文②	文③	文④	文⑤
滴水和尚（TS）	0	1	1	1	0
儀山禅師（GI）	0	0	1	0	1
峨山（GA）	1	1	0	0	0

TS（0, 1, 1, 1, 0）および GI（0, 0, 1, 0, 1）のコサイン類似度は，$|\mathrm{TS}| = \sqrt{0^2+1^2+1^2+1^2+0^2} = 1.73$ $|\mathrm{GI}| = \sqrt{0^2+0^2+1^2+0^2+1^2} = 1.41$ $\mathrm{TS}\cdot\mathrm{GI} = 1\times1 = 1$ だから $\mathrm{Sim}(\mathrm{TS},\mathrm{GI}) = \frac{1}{1.73\times1.41} = 0.408$ である。また TS（0, 1, 1, 1, 0）と GA（1, 1, 0, 0, 0）のコサイン類似度は，$|\mathrm{TS}| = 1.73$ $|\mathrm{GA}| = \sqrt{1^2+1^2+0^2+0^2+0^2} = 1.41$ $\mathrm{TS}\cdot\mathrm{GA} = 1\times1 = 1$ だから $\mathrm{Sim}(\mathrm{TS},\mathrm{GA}) = \frac{1}{1.73\times1.41} = 0.408$ である。つまり手計算によるコサイン類似度は同一であるが，これは，儀山禅師ベクトルと峨山ベクトルの要素値構成が同型であり，かつ，滴水和尚ベクトルとの内積に正の値を与える要素値も，上記の両ベクトルにおいて一つずつしか存在しないことによる。ところが KH Coder による Cosine 値は，TS と GI の場合が 0.78，TS と GA の場合は 0.13 である。大きな食い違いが生じてしまった。

他の同型ベクトルに目を向けてみよう。例えば「六の一」節にそろって出現しワットと共起するセークスピア，ダーウィン，ニュートンの各単語ベクトルがこれに当てはまる。先の例と同様，当該人名を含まない文には下線を引いた。

①されば一昨昨年（一九一三年）の末始めてロンドンに着き、取りあえず有名なウェストミンスター寺院を訪問して、はからずもゼームス・ワットの大理石像を仰ぎ見たる

時なども、私は実に言うべからざる感慨にふけった者である。②仰ぎ見れば、彼ワットはガウンを着て椅子に腰を掛け、大きな靴をはいて、左の足を後ろに引き、右の足を前に出し、紙をひざにのべ、左手にその端をおさえ、右手にはコンパスを握っている。そうして台石の表面には、次のような文字が彫り付けてある。「③この国の国王、諸大臣、ならびに貴族平民の多くの者どもが、この記念像をゼームス・ワットのために建てた。そは彼の名を永遠に伝えんとてにあらず、彼の名は平和の事業にして栄ゆる限り、かかる記念像をまたずして必ずや永遠に伝わるべきものである。むしろこの像は人間が……彼らの最上の感謝に値するところの人々を尊敬することをわきまえているという証拠を示すためにのみ、ただ建てられたものである。」

④彼ワットとは言うまでもなく蒸気機関の発明者である。しかしてこの蒸気機関の発明者こそ機械時代の先駆者の一人であってみれば、彼の名は実に人間にして滅びざる限り永遠に伝わるべきものである。

⑤ウェストミンスター寺院には、ダーウィンがいる、ニュートンがいる、セークスピアがいる、そうしてまたこのワットがいるのである。寺院のすぐ前は、ロンドンで最もにぎやかな場所の一つたるトラファルガル・スケアであって、そこには空にそびゆる高い高い柱の頂上に、ネルソン将軍が突き立っている。昔トラファルガルの海戦でスペイン、フランスの連合艦隊を一挙にしてほとんど全滅させ、自分もその場で戦いに倒れた英国海軍の軍神ネルソン卿の銅像が、灰色の空に突き立って下界を見おろしているのである。そのネルソン卿の見おろしている下の広場は、自動車や人間の往来に目もくらむばかりであって、道一ツ横切るにも私たちのようないなか者はいつもひやひやしたものである。カフェーにはいると、地下室になっている。そこへ腰を掛けて茶を飲んでいると、天井の明かり取りのガラス板の上をおおぜいの人が靴を踏み鳴らしながら通る。その騒々しさにはわれわれの神経もすり減らされるような気持ちであるが、さて戸を一つあけて寺院の内にはいると、たとえば浅草の公園でパノラマ館にはいったよう、空気はたちまち一変して、外の騒々しさはすべて拭いたように消されてしまって、寺院の内は靴音さえ慎まれるほどの静けさである。⑥私はそういう空気の中で彼ワットの像を仰ぎ見ながら、低徊去るあたわず、静かにさまざまの感想にふけったものであるが、今またこの物語を草して機械のことに及ぶに当たり、ゆくりなくも当時を追懐して、ここに無用の閑話に貴重なる一日の紙面をふさぐに至りし次第である。

ワット以外の三つの単語ベクトルは同等だが，例えばセークスピアを採って，

	文①	文②	文③	文④	文⑤	文⑥
ワット（W）	1	1	1	1	1	1
セークスピア（S）	0	0	0	0	1	0
ダーウィン（D）	0	0	0	0	1	0
ニュートン（N）	0	0	0	0	1	0

W（1, 1, 1, 1, 1, 1）および S（0, 0, 0, 0, 1, 0）のコサイン類似度を計算すれば，$|W| = \sqrt{1^2+1^2+1^2+1^2+1^2+1^2} = 2.45$　$|S| = \sqrt{0+0^2+0^2+0^2+1^2+0^2} = 1$　$W \cdot S = 1$　だから $\mathrm{Sim}(W,S) = \frac{1}{2.45\times1} = 0.408$となる。残り二つのベクトルをペアの相手にしても同値となることは明らかである。ここで重要なのは，KH Coder が算出した Cosine 値もまた，セークスピア，ダーウィン，ニュートンの三単語ベクトルのいずれかとワットベクトルとを比較したとき，つねに同じ 0.37 となっている事実である。

先に見た TS と GI，TS と GA では Cosine 値に大きな差が出たにもかかわらず，W と S，W と D，W と N の場合には Cosine 値が一致するのはなぜか。滴水和尚ベクトルやワットベクトルとペアを組んだベクトル同士は同型であったにもかかわらず，である。ここで，

KH Coder による Cosine 値には集計単位の位置関係で見た場合の単語の出現位置に応じた重み付けがなされているのではないか，と推測してみた。

ワットとセークスピア，ダーウィン，ニュートンを例にとる。共起（同一文内に登場）する場合の距離を 0 と置き，ペアの一方の単語がそこからどれほど離れた位置に存在するかを，共起文からペアの一方の単語を含む文に至るまでの最短文数で表現した場合，文⑤（共起文）の直前に位置する文④にワットが出現していることから，セークスピア，ダーウィン，ニュートンのいずれから見てもワットの距離は 1 である。滴水和尚と儀山禅師の例でも文③（共起文）から文④に至るまでの距離は 1 だが，滴水和尚と峨山の例では文②（共起文）から文①に至るまでの距離は 8 となっている。すでに確認した通り，前者の Cosine 値は 0.78，後者のそれは 0.13 で，距離の近い前者のほうが係数は大きい。

ところが徳川光圀と蓮如について改めて見てみると，文①は「十二の四」節の最終文，文②および文③は「十二の五」節の冒頭付近に位置するのに対し，共起文に当たる文④は「十三の二」節に存在し，文③と文④のあいだの距離は莫大なのにもかかわらず Cosine 値は 0.28 となっており，上の 0.13 よりも大きい。下記を参照されたい（同じく，当該人名が登場しない文には下線を引いてある）。

①蓮如上人御一代聞書にいう「御膳を御覧じても人の食わぬ飯を食うよとおぼしめされ候と仰せられ候」と。思うにこの一句、これを各戸の食堂の壁に題することを得ば、恐らく天下無用の費えを節する少なからざるべし。

私の倹約論は主として金持ちに聞いてもらいたいのだと言ったが、しかし私のいう意味のぜいたくは、多少の差こそあれ、金持ちも貧乏人も皆それ相応にしていることである。

②徳川光圀卿が常に紙を惜しみたまい、外より来る書束の裏紙長短のかまいなくつがせられ、詩歌の稿には反古の裏を用いたまいたる事はよく人の知るところである。現に水戸の彰考館に蔵する大日本史の草稿はやはり反古を用いある由、かつて実見せし友人の親しく余に物語りしことである。

③また蓮如上人御一代聞書を見ると、「蓮如上人御廊下を御通り候うて、紙切れの落ちて候いつるを御覧ぜられ、仏法領の物をあだにするかやと仰せられ、両の御手にて御いただき候としかじか、総じて紙の切れなんどのようなる物をも、仏物とおぼしめし御用い候えばあだに御沙汰なく候うの由、前々住上人御物語候いき」という記事がある。

【この間 100 文以上にわたり当該人名は登場しない】

④徳川光圀卿の惜しまれた紙、蓮如上人の廊下に落ちあるを見て両手に取っていただかれたという紙、その紙が必要品たるに論はないけれども、いかなる必要品でも使いようによっては限りなくむだにされうるものである。

よって，距離による重み付け仮説は誤りだと判明した。加えて，徳川光圀ベクトルと蓮如ベクトル間の Cosine 値が手計算値にきわめて近接したのは偶然に他ならなかったこともはっきりした。滴水和尚とのペアについて確認した通り，二つ以上の要素値において 0 を上回るという条件を共有する他の単語ベクトルが，ただちに反例となるからである。

問題は果たして解決したか

かくして行き詰まった果てに，樋口氏自らが KH Coder について解説したオフィシャルブックを点検してみた。そして，その初版および第二版のいずれにも，共起ネットワークではなく「多次元尺度構成法」の解説部分で，Cosine については「文書中における語の出現回数（粗頻度）をそのまま使うのではなく，1,000 語あたりの出現回数に調整したもの（調整頻度）を計算に使用する。文書の長さのばらつきに左右されない形で計算を行うた

めである」旨の但し書きが現れることを見出した（樋口 2014, 155; 樋口 2020, 180）。つまり，ここまでの分析では粗頻度を用いていた点にこそ問題の根があったと考えられる。

調整頻度を用いて滴水和尚・儀山禅師・峨山ベクトルのそれぞれを再構成してみよう。KH Coder が実装する形態素解析器 MeCab による総抽出語数は文①が 74，文②が 78，文③が 20，文④が 28，文⑤が 60 である。一千を総抽出語数で割った商に滴水和尚・儀山禅師・峨山の粗頻度を掛け合わせて，一千語当たりの調整頻度を求めたものが次表である。

	文①	文②	文③	文④	文⑤
TS	0	12.821	50	35.714	0
GI	0	0	50	0	16.667
GA	13.514	12.821	0	0	0

TS（0, 12.821, 50, 35.714, 0）および GI（0, 0, 50, 0, 16.667）のコサイン類似度は，$|\mathrm{TS}| = \sqrt{0^2 + 12.821^2 + 50^2 + 35.714^2 + 0^2} = 62.768$ $|\mathrm{GI}| = \sqrt{0^2 + 0^2 + 50^2 + 0^2 + 16.667^2} = 52.705$ $\mathrm{TS} \cdot \mathrm{GI} = 50 \times 50 = 2500$ だから $\mathrm{Sim}(\mathrm{TS}, \mathrm{GI}) = \frac{2500}{62.768 \times 52.705} = 0.7557$ である。また TS（0, 12.821, 50, 35.714, 0）と GA（13.514, 12.821, 0, 0, 0）のコサイン類似度は，$|\mathrm{TS}| = 62.768$ $|\mathrm{GA}| = \sqrt{13.514^2 + 12.821^2 + 0^2 + 0^2 + 0^2} = 18.628$ $\mathrm{TS} \cdot \mathrm{GA} = 12.821 \times 12.821 = 164.378$ だから $\mathrm{Sim}(\mathrm{TS}, \mathrm{GA}) = \frac{164.378}{62.768 \times 18.628} = 0.1406$ である。KH Coder の Cosine 値 0.78 および 0.13 と概ね一致した。

徳川光圀・蓮如についても一千語当たりの出現頻度を求めておこう。総抽出語数は文①が 34，文②が 49，文③が 102，文④が 60 より，

	文①	文②	文③	文④
TM	0	20.408	0	16.667
R	29.412	0	19.608	16.667

ここで文③において蓮如が 2 回登場することに注意されたい。つまり粗頻度が 2 だから，調整頻度は 2 に 1000/102 を掛け合わせて求めている。TM（0, 20.408, 0, 16.667）および R（29.412, 0, 19.608, 16.667）のコサイン類似度は，$|\mathrm{TM}| = \sqrt{0^2 + 20.408^2 + 0^2 + 16.667^2} = 26.349$ $|\mathrm{R}| = \sqrt{29.412^2 + 0^2 + 19.608^2 + 16.667^2} = 39.081$ $\mathrm{TM} \cdot \mathrm{R} = 16.667 \times 16.667 = 277.789$ だから $\mathrm{Sim}(\mathrm{TM}, \mathrm{R}) = \frac{277.789}{26.349 \times 39.081} = 0.2698$ である。KH Coder の Cosine 値 0.28 とほぼ一致する。

しかしながら，わずかとはいえ手計算値と測定値にズレが生じていることは否めない。金子創氏より，この乖離は形態素解析器の差違に由来するのではないかと指摘があった。そこで，KH Coder が MeCab とともに実装する ChaSen を用い，徳川光圀と蓮如のペアについて改めて計測してみたところ，総抽出語数は文①が 33，文②が 48，文③が 103，文④が 62 となった。一千語当たりの調整頻度は下表の通りである。

	文①	文②	文③	文④
TM	0	20.833	0	16.129
R	30.303	0	19.417	16.129

TM（0, 20.833, 0, 16.129）および R（30.303, 0, 19.417, 16.129）のコサイン類似度は，$|\mathrm{TM}| = \sqrt{0^2 + 20.833^2 + 0^2 + 16.129^2} = 26.347$ $|\mathrm{R}| = \sqrt{30.303^2 + 0^2 + 19.417^2 + 16.129^2} = 39.439$ $\mathrm{TM} \cdot \mathrm{R} = 16.129 \times 16.129 = 260.145$ だから $\mathrm{Sim}(\mathrm{TM}, \mathrm{R}) = \frac{260.145}{26.347 \times 39.439} = 0.2504$ である。MeCab を用いた値として先に見た 0.2698 に比べると，0.28 からより大きく遠ざかった。

ちなみに，KH Coder で算出した「総抽出語数」（延べ語数）と「異なり語数」，および，一部品詞等を省いた各々の「使用」語数は次の通りである。括弧内が使用語数を示す。

		文①	文②	文③	文④
MeCab	総抽出	34(14)	49(25)	102(45)	60(25)
	異なり	29(12)	41(24)	64(37)	44(22)
ChaSen	総抽出	33(16)	48(25)	103(48)	62(25)
	異なり	28(14)	39(23)	62(38)	46(22)

上表に基づき，一千語当たりの調整頻度をそれぞれ算出したものが下表である。

			文①	文②	文③	文④
TM	Me	総	0	20.408(40)	0	16.667(40)
		異	0	24.39(41.667)	0	22.727(45.455)
	Cha	総	0	20.833(40)	0	16.129(40)
		異	0	25.641(43.478)	0	21.739(45.455)
R	Me	総	29.412(71.429)	0	19.608(44.444)	16.667(40)
		異	34.483(83.333)	0	31.25(54.054)	22.727(45.455)
	Cha	総	30.303(62.5)	0	19.417(41.667)	16.129(40)
		異	35.714(71.429)	0	32.258(52.632)	21.739(45.455)

MeCab の総抽出∧使用語数を用いた調整頻度によるコサイン類似度：

TM（0, 40, 0, 40）および R（71.429, 0, 44.444, 40）

$|TM| = \sqrt{40^2 + 40^2} = 56.569$ $|R| = \sqrt{71.429^2 + 44.444^2 + 40^2} = 93.152$ $TM \cdot R = 40^2 = 1600$ $Sim(TM, R) = 0.3036$

MeCab の異なり語数を用いた調整頻度によるコサイン類似度：

TM（0, 24.39, 0, 22.727）および R（34.483, 0, 31.25, 22.727）

$|TM| = \sqrt{24.39^2 + 22.727^2} = 33.337$ $|R| = \sqrt{34.483^2 + 31.25^2 + 22.727^2} = 51.79$ $TM \cdot R = 22.727^2 = 516.517$ $Sim(TM, R) = 0.2992$

MeCab の異なり∧使用語数を用いた調整頻度によるコサイン類似度：

TM（0, 41.667, 0, 45.455）および R（83.333, 0, 54.054, 45.455）

$|TM| = \sqrt{41.667^2 + 45.455^2} = 61.663$ $|R| = \sqrt{83.333^2 + 54.054^2 + 45.455^2} = 109.235$ $TM \cdot R = 45.455^2 = 2066.157$

$Sim(TM, R) = 0.3067$

ChaSen の総抽出∧使用語数を用いた調整頻度によるコサイン類似度：

TM（0, 40, 0, 40）および R（62.5, 0, 41.667, 40）

$|TM| = \sqrt{40^2 + 40^2} = 56.569$ $|R| = \sqrt{62.5^2 + 41.667^2 + 40^2} = 85.102$ $TM \cdot R = 40^2 = 1600$ $Sim(TM, R) = 0.3324$

ChaSen の異なり語数を用いた調整頻度によるコサイン類似度：

TM（0, 25.641, 0, 21.739）および R（35.714, 0, 32.258, 21.739）

$|TM| = \sqrt{25.641^2 + 21.739^2} = 33.616$ $|R| = \sqrt{35.714^2 + 32.258^2 + 21.739^2} = 52.808$ $TM \cdot R = 21.739^2 = 472.584$ $Sim(TM, R) = 0.2662$

ChaSen の異なり∧使用語数を用いた調整頻度によるコサイン類似度：

TM（0, 43.478, 0, 45.455）および R（71.429, 0, 52.632, 45.455）

$|TM| = \sqrt{43.478^2 + 45.455^2} = 62.901$ $|R| = \sqrt{71.429^2 + 52.632^2 + 45.455^2} = 99.691$ $TM \cdot R = 45.455^2 = 2066.157$ $Sim(TM, R) = 0.3295$

以上六つのコサイン類似度はいずれも KH Coder の Cosine 値 0.28 から乖離し，最も近傍の 0.2662 でさえ MeCab による総抽出語数を用いた先の計測値 0.2698 に及ばない。したがって，KH Coder における Cosine 値の算出に用いられているのは総抽出語数であって異なり語数ではなく，また使用語数でもないとの判断に，間違いはないだろう。

では，形態素解析器についてはどうか。MeCab と ChaSen のどちらが計測に用いられているのか。光圀と蓮如の例を採れば MeCab と考えられよう。しかし，このことの最終判定は，より多くの事例につき検証して以後にしたほうが無難である。下表は，本稿で扱った人名について，MeCab と ChaSen とに場合分けしたうえで，調整頻度およびそれに基づくコサイン類似度をまとめてみたものである。徳川光圀と蓮如は重複するため割愛した。

	MeCab				ChaSen			
文	①	②	③	④	①	②	③	④
語数	26	13	53	36	32	13	53	36
T	38.46	76.92	18.87	27.78	31.25	76.92	18.87	27.78
H	38.46	0	0	0	31.25	0	0	0
cos	0.4166				0.3489			
kh	0.43							

文	①	②	③	④	⑤	①	②	③	④	⑤
語数	56	25	38	46	78	55	25	38	47	76
W	0	0	0	21.74	0	0	0	0	21.28	0
M	17.86	40	52.63	65.22	25.64	18.18	40	52.63	63.83	26.32
文	⑥	⑦	⑧	⑨	⑩	⑥	⑦	⑧	⑨	⑩
語数	173	35	62	10	33	176	35	62	10	34
W	0	0	0	0	0	0	0	0	0	0
M	11.56	28.57	16.13	100	30.3	11.36	28.57	16.13	100	29.41
cos	0.4425					0.4349				
kh	0.44									

文	①	②	③	④	⑤	①	②	③	④	⑤
語数	74	78	20	28	60	74	79	19	27	60
TS	0	12.82	50	35.71	0	0	12.66	52.63	37.04	0
GI	0	0	50	0	16.67	0	0	52.63	0	16.67
GA	13.51	12.82	0	0	0	13.51	12.66	0	0	0
cos	0.7557(TS,GI) 0.1406(TS,GA)					0.765(TS,GI) 0.1319(TS,GA)				
kh	0.78(TS,GI) 0.13(TS,GA)									

文	①	②	③	④	⑤	⑥	①	②	③	④	⑤	⑥
語数	55	61	29	13	29	78	54	61	31	13	27	78
W	18.18	16.39	34.48	76.92	34.48	12.82	18.52	16.39	32.26	76.92	37.04	12.82
S	0	0	0	0	34.48	0	0	0	0	0	37.04	0
D	0	0	0	0	34.48	0	0	0	0	0	37.04	0
N	0	0	0	0	34.48	0	0	0	0	0	37.04	0
cos	0.3622(W,S/W,D/W,N) 1(S,D,N)						0.3881(W,S/W,D/W,N) 1(S,D,N)					
kh	0.37(W,S/W,D/W,N) 1.0(S,D,N)											

明らかなように，ChaSen を使用して調整頻度を求めた場合のほうが，MeCab を使用した場合よりも，コサイン類似度の手計算値（cos 欄）が KH Coder の算出値（kh 欄）に近づく例も見られる。よってここでの暫定的結論は，Cosine 算出に際しての形態素解析器には MeCab が用いられていると想定したほうが不都合は小さい，という程度のことに過ぎなくなる。またウェストファーレンとマルクスの例を除き，どちらの形態素解析器を用いて手計算を行った場合にも Cosine 値からの乖離が多かれ少なかれ生じている。上述した通り，調整頻度の算出に総抽出語数が用いられていることは否定できないと考えられるが，算出式（粗頻度×一千÷集計単位当たり総抽出語数）に何らかの別項が付け加えられるなど，隠れた重み付け操作がなされている可能性も否定できないだろう。

結びにかえて

本稿では，河上肇『貧乏物語』のテキストデータを素材として，著作に登場する諸人物の相関図を KH Coder の共起ネットワーク分析を用いて描き出す手法について検討した。その際の前提は，「言葉の袋」の内部で共起する単語（本稿の場合は人名）は何らかの共通の特徴のもとに置かれている蓋然性が高いという仮説である（小峯 2021, 34-35）。本稿において「言葉の袋」に相当するのは集計単位となる文書であった。このことをふまえつつ本稿での分析を通じて得られたいくつかの知見につき，テキストマイニングの裏側に張り付いている限界ないし制約を明るみに出すためにも，以下で改めてまとめておきたい。

すでに見たように，Jaccard や Simpson や Cosine といった相関係数で測られる人名同士の共起の強さの値は，同一のペアを対象とした場合でも係数の種別に応じて異なるが，相

関図自体についてはどの係数を用いようとも全く同等のものが作り出される。理由は明確である。測定値が 0 を上回る，すなわち相関が認められるためには，集計単位中において必ず一例は人名同士の共起が見られなければならず，テキストデータをめぐるこの構造的制約はすべての読み取り手にとっての共通の条件として，初めから与えられているからである。同一テキストを対象とするかぎり，係数の種別を問わず，共起を認めるための条件は等しい。そしてこの条件が満たされなければ，どの係数についても定義式の分子の値が 0 になる。特に Cosine については単語ベクトルの内積が必ず 0 になると言える。

つまり人物同士の相関は，究極的には，集計単位中で少なくとも 1 回共起しているか否かだけで判定されている。たとえ文脈上きわめて重要なペアであったとしても，集計単位中で共起することが一度もなければ相関は絶無と見なされる。この点はきわめて重要な，場合によっては致命的とも評しうる，分析手法上の弱点だろう。ただし，これもすでに見たように，集計単位を文から段落に広げると関連付けられる人名数が増加する可能性が高い。これは（原則として）段落が複数の文の集合体であるため，一文よりも一段落を最小集計単位とするほうが，通常は，共起する人名の出現確率が高まることによる。上述した通りテキストデータの構造は当初より不動だが，集計単位を何に設定するかについては読み取り手に裁量の余地があり，見様によってはそこにこそテキストを超出した自由が存在する。ただし集計単位を段落以上に大きく拡張すると人名の関連付けがおびただしい量に上るため，人物相関図の態をなさなくなることには注意を払うべきである。

この点について，田中秀臣氏より，KH Coder を通じて拾い上げることのできない人名，すなわち影の部分にこそむしろ注目するべきであるとの指摘を受けた。集計単位における共起が 1 回以上という量的基準だけで人物相関図を描出することは明らかに乱暴であり，共起していないにもかかわらず重要な相関を有するペアを探り当てることで，いわば質的側面から量的側面を補完，時に掣肘することが，テキストマイニングの成果をより説得性あるものとしていくうえで不可欠との示唆である。例えば河上が引用する間宮英宗の講話はスミス『道徳感情論』の「注脚」と特筆されているにもかかわらず（河上 2008, 44），間宮とスミスは同一文中に登場しないため，集計単位を文とした場合には相関係数値は 0 である。しかし一段落中では共起するため，集計単位を段落とした場合にはスミスを中心とする人脈の一角を占める（図を参照）。またマルクスと孔子の関係も興味深い。集計単位を文にしても段落にしても両者は一度として共起しないため相関係数値は 0 だが，著者河上は「経済組織がまず変わってしかるのちに人の思想精神が変わる」というマルクスの命題の東洋版として，事の順序をめぐる孔子の言葉「食を足し，兵を足し，民をして之を信ぜしむ」を引いているからである（*ibid.*, 153-54）。間宮とスミス，マルクスと孔子のような事例を逐一確かめれば，集計単位選択上の按配についての知見をより深めたり，小手先の技術的裁量では埋め合わせ不可能な，情報の決定的欠落が生じるリスクの程度を見極めたりすることも，可能になるのではないだろうか。

このような制約を意識しながら，本稿では単語＝人名ベクトル同士のコサイン類似度に焦点を絞り，KH Coder 算出の Cosine 値と手計算値とが乖離する問題を取り上げ，それに一定の解決策を与えた。隔靴掻痒の感を残す不本意な結論に落ち着いたと言わざるを得ないが，しかし安易に開発者に答えを求めることなく，それぞれの実践者が自ら立てた問いの解決を模索する中でこそ，テキストマイニングという技法の習熟度が高まるとともに，算出された数値の持つ有効性とその限界とがより明瞭に浮かび上がるはずである。

『貧乏物語』登場人物一覧

本文中表記	回	備考
01 アダムス	1	Brooks Adams/1848-1927：米の歴史家
02 啄木	1	石川啄木/1886-1912：歌人
03 ウィザース	1	Hartley Withers/1867-1950：英のジャーナリスト
04 ウィルソン	2	William T. Wilson/1855-1921：英の労働運動家
05 ルードウィヒ・フォン・ウェストファーレン	1	Ludwig von Westphalen/1770-1842：マルクスの義父
06 内ヶ崎作三郎	1	1877-1947：文学者・政治家
07 内田	1	内田清之助/1884-1975：鳥類学者
08 エヴァンス	2	Beriah Evans/1848-1927：英のジャーナリスト
09 エモット	2	Alfred Emmott/1858-1926：英の政治家
10 オスボーン	2	Henry F. Osborn/1857-1935：米の古生物学者
11 小野塚	1	小野塚喜平次/1871-1944：政治学者
12 峨山禅師,峨山和尚	2	峨山昌禎/1853-1900：臨済宗僧侶
13 加藤	1	加藤高明/1860-1926：政治家
14 金井	1	金井延/1865-1933：経済学者
15 舎弟	1	河上左京/1889-1971：本書著者の弟。水彩画家
16 父	1	河上忠/1848-1927：本書著者の父。元岩国藩士
17 河上肇	1	1879-1946：本書著者
18 顔淵	1	顔回/前 520 頃-490 頃：孔子十哲
19 儀山禅師	2	儀山善来/1802-78：臨済宗僧侶
20 ギゾー	1	François Guizot/1787-1874：仏の歴史家・政治家
21 キッチナー	1	Herbert Kitchener/1850-1916：英の軍人
22 エドウィン・キャナン,キャナン	3	Edwin Cannan/1861-1935：英の経済学者
23 キング	2	Willford I. King/1880-1962：米の統計学者
24 久原	1	久原房之助/1869-1965：実業家・政治家
25 熊沢蕃山	1	1619-91：陽明学者
26 グラッドストーン	1	William Gladstone/1809-98：英の政治家
27 クローレイ	1	Ralph H. Crowley/1869-1953：クローリー 英の医師
28 孔子	10	前 550 頃-前 479 頃：儒学の祖
29 阪谷	1	阪谷芳郎/1863-1941：官僚・政治家
30 子貢	2	前 520-前 456 頃：孔子十哲
31 渋沢	1	渋沢栄一/1840-1931：実業家
32 釈雲解	1	文政十二（1829）年頃に「生財弁」執筆
33 ステップ	1	Edward Step/1855-1931：英の博物学者
34 アダム・スミス,スミス	19	Adam Smith/1723-90：経済学の父
35 ゼームス・ハルデーン・スミス,ゼー・エチ・スミス	2	James Haldane Smith：1916 年にロンドン，翌年にニューヨークで『経済上の道徳』*Economic Moralism* を刊行し，後者はフランク・ナイトが JPE 誌で書評
36 セークスピア	1	William Shakespeare/1564-1616：シェークスピア 劇作家

本文中表記	回	備考
37　セリグマン	1	Edwin R. A. Seligman/1861-1939：米の経済学者
38　大燈国師	3	宗峰妙超/1283-1338：臨済宗僧侶
39　ダーウィン	1	Charles Darwin/1809-82：英の進化論者
40　タフト	1	William H. Taft/1857-1930：米の政治家
41　ダントン	1	Georges Danton/1759-94：仏の政治家
42　ダンロップ	2	George H. M. Dunlop/1859-1916：英の医師
43　チャンバーレン	4	Joseph Chamberlain/1836-1914：チェンバレン 英の政治家
44　テイラー	1	Pierre Teilhard de Chardin/1881-1955：テイヤール 仏のイエズス会士・古人類学者
45　滴水和尚	3	由理(利)滴水/1822-99：臨済宗僧侶
46　寺内	1	寺内正毅/1852-1919：軍人・政治家
47　デュブア	2	Eugène Dubois/1858-1940：蘭の古人類学者
48　ドウソン	2	Charles Dawson/1864-1916：英のアマチュア考古学者。英 Piltdown 原人を捏造。河上は「ビ」ルトダウンと記す
49　徳川光圀	2	1628-1701：水戸藩主
50　トルストイ	4	Lev N. Tolstoy/1828-1910：露の文豪
51　ニュートン	1	Isaac Newton/1642-1727：英の数学者・物理学者
52　額田	1	額田豊/1878-1972：医師
53　ネルソン	3	Horatio Nelson/1758-1805：英の軍人
54　バーネット・ハースト	1	Alexander R. Burnet-Hurst：印 Muir Central College 教授
55　ハチソン	1	Francis Hutcheson/1694-1746：英の道徳哲学者
56　ハマライネン	1	Hamalainen：ヘルシンキ大教授
57　ハンター	3	Robert Hunter/1874-1942：米の著述家
58　ビレル	1	Augustine Birrell/1850-1933：米の法律家・政治家
59　ヒューム	2	David Hume/1711-76：英の歴史家・哲学者
60　フォルソム	1	Justus W. Folsom/1871-1936：米の昆虫学者
61　福田	1	福田徳三/1874-1930：経済学者
62　チャーレス･ブース,ブース	3	Charles Booth/1840-1916：英の統計学者
63　ブライアント	1	Louise S. Bryant/1885-1956：米の公衆衛生技師
64　プレンゲ	2	Johann Plenge/1874-1963：独の経済学者
65　ベケル	1	Becker：ヘルシンキ大教授
66　ベーツ	3	Henry W. Bates/1825-92：英の博物学者
67　トマス・ベルト,ベルト	5	Thomas Belt/1832-78：英の博物学者・鉱山技師
68　ウィリアム・デーン・ホゥエルス	1	William D. Howells/1837-1920：米の著述家
69　ボウレイ	3	Arthur L. Bowley/1869-1957：ボウリー 英の経済統計学者
70　ホランダー	1	Jacob H. Hollander/1871-1940：米の経済学者。poverty を economic inequality と economic dependence と economic insufficiency に三区分し，河上『貧乏物語』による区別に先駆けた。この点に気づいた福田が河上と論争
71　マグレゴア	2	James H. McGregor/1872-1954：マグレガー 米の動物学者

	本文中表記	回	備考
72	チオザ・マネー	1	Leo C. Money/1870-1944：伊出身。英の政治家
73	間宮英宗	1	1871-1945：臨済宗僧侶
74	カール・マルクス,マルクス	10	Karl H. Marx/1818-83：『資本論』著者
75	マルクスの細君	4	Jenny von Westphalen/1814-81：マルクスの妻
76	マルクスの父	1	Heinrich Marx/1777-1838：独の法曹
77	マルサス	4	Thomas R. Malthus/1766-1834：『人口論』著者
78	マンダヴィル	2	Bernard Mandeville/1670-1733：『蜂の寓話』著者
79	皆川淇園	1	1735-1807：儒学者
80	三宅	1	三宅恒方/1880-1921：昆虫学者
81	ミル	1	John Stuart Mill/1806-73：英の哲学者・経済学者
82	孟子	8	前 372 頃-前 289 頃：儒学者
83	フレデリック・モーリス,モーリス	2	John F. Maurice/1841-1912：英の軍人
84	ラスキン	1	John Ruskin/1819-1900：英の美学者
85	ランゲ	1	Friedrich A. Lange/1828-75：独の哲学者
86	ウィリアム・レーン,レーン	8	William Lane/1861-1917：英のユートピア思想家
87	蓮如	4	1415-99：浄土真宗僧侶
88	リチャード・ロイド	7	Richard Lloyd/1834-1917：英の牧師
89	ロイド・ジョージ	47	David Lloyd-George/1863-1945：英の政治家
90	ローズベリー	1	A. P. Primrose, Earl of Rosebery/1847-1929：英の政治家
91	ロレンズ	3	Max O. Lorenz/1876-1959：ローレンツ 米の統計学者
92	ローンツリー	10	Seebohm Rowntree/1871-1954：英の社会改良家
93	渡辺	1	渡辺銕蔵/1885-1980：経済経営学者
94	ゼームス・ワット,ワット	6	James Watt/1736-1819：英の技師。河上が眺めた像は現在スコットランド国立博物館に安置

注）回は出現粗頻度を表す。表記が二様に揺れる場合には「,」で区切り，併記した。

主要参考文献

青空文庫 https://www.aozora.gr.jp
岡﨑直観 2016.「単語の意味をコンピュータに教える」岩波データサイエンス刊行委員会編『岩波データサイエンス Vol.2』岩波書店.
河上肇［1947,1965］2008.『貧乏物語』岩波文庫.
小峯敦編 2021.『テキストマイニングから読み解く経済学史』ナカニシヤ出版.
杉原四郎 1979.「福田徳三と河上肇」『経済論叢』第 124 巻第 5・6 号.
住谷一彦編 1984.『河上肇』中央公論社.
林直樹 2022.「社会思想史研究とテキストマイニング」『愛知学院論叢「経済学研究」』第 9 巻第 2 号.
樋口耕一 2014.『社会調査のための計量テキスト分析：内容分析の継承と発展を目指して』ナカニシヤ出版.
樋口耕一 2020.『社会調査のための計量テキスト分析：内容分析の継承と発展を目指して第２版』ナカニシヤ出版.