# 『論語』の基礎統計

齊藤正高

## はじめに

20世紀末から21世紀にかけて、世界中に存在する文字を一つの体系で扱うユニコードが普及し、漢字文献のデジタル・データベースもこの20年あまりで目をみはるような発展があったことは、多くのひとが感じていることではないだろうか。「漢籍電子文献」「CTEXT」「漢籍リポジトリ」「SAT 大正新脩大藏經テクストデータベース」「CBETA 中華電子仏典協会」など、膨大なデータを公開しているアーカイブをみると、どれ一つとっても、一生をかけても読みきれない驚異的な量であり、そもそも総体として「読む」対象としてよいのかという疑いを禁じえない。しかし、これらのデータベースが提供している「キーワード検索」を使えば、従来、深夜までかかることもまれではなかった「出典さがし」を、比較的短時間で行えるのである。これは確かに利便性が高く、今後も不可欠のツールとしてありつづけるのだろう。

このように「大きなデータベース」から恩恵をうける一方で、いったい一冊の書物がどんな姿をしているのかという「小さな関心」もある。この「小さな関心」を情報学の立場から明らかにしている分野もあり、さまざまな名でよばれている。明白な区別があるわけではないが、文学作品などを分析する場合には「計量文献学」、語彙・語用などをコーパスから論ずる場合には「計量言語学」、社会調査アンケートの自由記述を分析する場合には「計量テキスト分析」、おもにビジネスで有益な情報を発見しようとする場合には「テキスト・マイニング」などと呼ばれる分野である。これらはいずれも、工学でいう「自然言語処理」（NLP）を用いる点で共通しており、人工知能研究の進展によって、その重要性が指摘されている「文理融合」の分野なのであろう。すでに『計量国

語学事典』（朝倉書店、2009年）、『言語処理学事典』（共立出版、2009年）など、大部の事典も出版されている。

小論は現在一般に読むことのできる『論語』を例に、デジタル・テクストの基礎統計で何がわかるのかという問題を、「小さな関心」をもつ一利用者の立場から再考したい。近年展開している高度なテクスト分析の方法[1]は準備がないので扱えないが、文字の使用に即して、「分け方」と「数え方」によってできることも確認しておきたい。

なお、定州漢墓竹簡『論語』（55年没の中山懐王劉脩の墓から1973年に発掘、1993年公開）など、出土文献による研究によって、「呼称や個々の文字の相違に過敏に反応して、篇ごとの成立時代を論断することはきわめて危険である」と指摘されている（湯浅邦弘、2012年、p. 80）。デジタル・テクストによる分析こそ、こうした危険がつねにともなうことを自覚しておかねばならないであろう。したがって、小論では通時的問題を論ずることはできない。現在に伝わっている『論語』について、一つの条件下で現れる姿を述べるにすぎない。

## 1．計量文献学について

村上征勝氏・土山玄氏の指摘[2]によれば、計量文献学の歴史は19世紀中頃にさかのぼる（表1）。もちろん、コンピュータの普及以前は手作業の集計にたよっていた。初期の試みは、それまでの本文批判によって疑問がもたれていた、『聖書』やプラトンの著作の一部について、単語の長さの平均値を用いた分析にはじまる。これは単語を分かち書きする西欧語の特徴をうつしたものであろう。その後、比較的大きな単位である文や句読法（punctuation）も分析対象になった。いずれも基本的には著者の推定や文体の把握といった目的でなされ、パウロやプラトンやシェイクスピアといった作家が分析された。

日本語の文献は一般的に単語の分かち書きをしないので、品詞構成比や特定の品詞・色彩語などの出現確率を用いた研究があらわれた。これによって『源氏物語』の補作部分の特徴などが指摘されている。日本語と同様に、単語を分かち書きしない漢字文献を対象とする研究では、1970年代末に助字（也・矣・焉など）の使用頻度を使って仏典の訳者推定が行われ、80年代にも助字を使って『紅楼夢』の補作部分の分離や著者推定などが行われた[3]。

画期であったのが、2000年、近藤みゆき氏が『古今和歌集』にみえるジェンダー性の抽出にNグラム統計をもちいた研究[4]を発表し、これをうけて石井公成氏がNGSM（N-Gram based System for Multiple document comparison and

表1　計量文献学略年表

| 年 | 人　物 | 事　蹟 | 手法など |
|---|---|---|---|
| 1851 | ド・モルガン | 「ヘブル人への手紙」(『新約聖書』) の分析法を提案 | 単語の長さの平均値 |
| 1867 | キャンベル | プラトン『ソピステス』『政治家』『ピレボス』などを後期対話編であると指摘 | 語の生起状況 |
| 1887 | メンデンホール | サッカレー、ディケンズ、J・S・ミルなどの文を分析し「ワード・スペクトル」を提唱 | 単語の長さの度数分布 |
| 1897 | ルトスワフスキ | プラトンの文体の発展 | ユールに影響 |
| 1900 | ピアソン | カイ二乗検定 | |
| 1901 | メンデンホール | シェイクスピア非実在説を否定 | 10万単位のデータから単語長の分布を測定 |
| 1935 | 波多野完治 | 『文章心理学』刊行 | |
| 1939 | ユール | 『キリストにならいて』の著者推定 | 文の長さの平均値、中央値、四分位範囲など |
| 1944 | ユール | 語彙の集中度をあらわす「K特性値」を提唱 | |
| 1951 | シャノン | Ngram | |
| 1956 | 大野晋 | 大野の語彙法則 | 『万葉集』『源氏物語』などの品詞の構成比 |
| 1957 | ウェイク | プラトン「第七書簡」の偽作説を否定 | 文の長さの比較 |
| | | 計量国語学会設立 | |
| | 安本美典 | 『源氏物語』宇治十帖 | 直喩・声喩・色彩の使用度など12項目を分析 |
| 1958 | ルーン | 自動抄録、KWIC (Key Word In Context) | |
| 1962 | エレゴール | 英国議会を批判した17世紀のユリウス・レターを分析 | Distinctiveness ratio を提案 |
| 1964 | モステラ、ワルラス | 『ザ・フェデラリスト』を分析 | 機能語の使用頻度 |
| 1965 | モートン | 「パウロ書簡」(『新約聖書』) に6人の著者がいると指摘 | 共通語彙の頻度 |
| | 水谷静夫 | 「大野の語彙法則」の数式化 | |
| 1966 | オドネル | 一般女性の書いた手紙 | 句読点の頻度を分析 |
| 1972 | スパーク | TFIDF（キーワードの重みづけ） | キーワードの重みづけ |
| 1973 | 韮沢正 | 『由良物語』の著者推定 | 語の使用率を分析して推定 |
| 1974 | ウィリアムス | メンデンホール(1901)に疑義を示す | 散文と韻文ではワードスペクトルは異なる |
| 1978 | 後藤義乗 | 『無量寿経』の著者推定 | 機能語の使用頻度 |
| 1984 | ヘスツア | 『静かなるドン』の盗作説を否定 | 単語長の分布・のべ語数と異なり語数の比 |

| 年 | 人　物 | 事　蹟 | 手法など |
|---|---|---|---|
| 1987 | 李賢平 | 『紅楼夢』81〜120章の著者を推定 | 41種の機能語を主成分分析・クラスター分析 |
| 1992 | 伊藤瑞叡、村上征勝 | 日蓮遺文の真贋鑑定 | クラスター分析など |
| 1992 | | TREC（情報検索評価ワークショップ）の開始 | |
| 1994 | ロバートソンら | BM25（TFIDFを補正したキーワードの重みづけ） | |
| 1998 | | 漢字文献情報処理研究会設立、『電脳中国学』刊行 | |
| 1999 | フォルシスら | キケロ『慰め』（1583発見）が偽作であることを指摘 | 語の長さの判別分析 |
| 2000 | 土橋喜 | 情報視覚化と仮説形成支援 | |
| | 近藤みゆき | 『古今和歌集』のジェンダー | nグラム統計 |
| 2001 | 石井公成 | 仏教文献の異本比較・訳者推定・和臭の発見 | NGSMの提唱 |
| 2002 | 村上征勝 | 『源氏物語』54篇の成立順序の分析 | 助動詞などの使用率を数量化Ⅲ類で分析 |
| | 師茂樹 | 『般若心経』の異訳の比較 | NGSM |
| 2003 | ビノンゴ | 『オズ』の最終巻がトンプソンであると指摘 | 機能語の主成分分析 |
| | 秋山陽一郎 | 『老子』傳奕本来源考—「項羽妾本」介在の検証 | NGSM |
| 2004 | 師茂樹 | 玄奘訳全文献のクラスター分析による仮説形成 | NGSM |
| | 山田崇仁 | 中国戦国期の語彙量 | NGSMとK特性値 |
| | 樋口耕一 | KH Coderの発表 | |
| 2009 | | 『計量国語学事典』『言語処理学事典』刊行 | |

analysis）を提起したことである[5)]。

いわゆるNグラムとは、クロード・シャノン（1916–2001）が提唱した概念で、N文字の連続する文字の固まりをいう。文書の先頭から末尾へ、1字ずつずらしてN文字の固まりを切りだしていくと、Nグラムの集合を抽出できる。とくに1グラムを「ユニグラム」、2グラムを「バイグラム」とよぶこともある。このNグラム分割の結果には意味をなす固まりもあれば、意味をなさない固まりもあるが、意味をなさないようにみえる固まりであっても、それが著者のクセをあらわしていることがあり、Nグラムを大量にあつめて統計をとって比較すると、歴史研究によって意見の分かれる異本系統の研究などに使うことができ

る。このNグラムは谷本玲大氏によって、いちはやく漢字文献の検索に応用された[6]。

師茂樹氏はNGSMを使い、『般若心経』の異訳を比較して樹状図（デンドログラム）に示すなど多くの研究を発表し[7]、土橋喜氏の仮説論を参照して[8]、玄奘（602–664）訳とされる全仏典の分類を行い、「文体上筆受の影響が少なくなく、また大きく二つに分類することができる」と指摘している。また、山田崇仁氏は先秦の文献を対象に数多くの研究を発表し、NGSMとK特性値（後述）をつかい、諸子百家の語彙量の比較を行い、「儒家・道家は多様な言葉を保持し、墨家は貧困、法家（韓非子）は洗練、雑家（『呂氏春秋』）はまさしく諸家を折中した言葉を用いている」と指摘した[9]。秋山陽一郎氏はNGSMを使い、『老子』傅奕本に「項羽妾本」の参照された部分を指摘するという興味深い研究を発表した[10]。

以上に概観したように、計量文献学はまず西欧で著者の推定や文体を把握するためにはじまった。これらは作品の背後に「著者」がいるという前提で成立するものであろう。しかし、中国で個人の著作が成立するのは漢代以降で、先秦の著作は「一時一人の作」ではなく、後の挿入があることが一般的であると指摘されている（余嘉錫『古書通例』平凡社、2008年、第11章）。こうした古代中国の文献がもつ特徴をうつして、Nグラムを用いた漢字文献の分析は著者の推定にとどまらず、言語や文献のさまざまな問題を扱うように進展しているようにみうけられる。

なお、基礎的統計量である字数については、中国においても古くから問題になってきた。司馬遷は老子が「道徳の意、五千余言を言いて去る」（『史記』老荘申韓列伝）と書き、後漢の趙岐（201年没）は『孟子』について、「二百六十一章、三万四千六百八十五字」（「孟子題辞」）と書いた。これについて清の焦循（1763–1820）は明の陳士元（1516–1597）『孟子雑記』を引き、次のようにいう。

> いま、字数を計るに「梁恵王篇」上下共せて五千三百六十九、「公孫丑篇」上下共せて五千一百四十四、「滕文公篇」上下共せて五千零四十五、「離婁篇」上下共せて四千七百八十九、「万章篇」上下共せて五千一百二十五、「告子篇」上下共せて五千二百五十五、「尽心篇」上下共せて四千六百八十三、これを統べれば、実に三万五千四百一十字あり。趙説と較べて七百二十五字多し。（焦循『孟子正義』「孟子題辞」疏）

これについて、焦循は別本で数えなおして趙岐より「実に五百四十一字多

し」（同前）という結果を得ている。このように字数についての疑問は古くからある。なお、この文が興味ぶかいのは「計」と「統」が文字数について使われていることであろう。

また、文献の性質を語った古い言葉に、「詩三百、一言を以て之を蔽（おお）えば、曰く思い邪（よこしま）なし」（『論語』為政）がある。この孔子の言葉は『詩』311篇の概数を言い、その7000をこえる句から「思い邪なし」（魯頌「駉（けい）」）という句によって、『詩』を要約している。つまり、全体を量によって提示したのち、その「代表例」を抽出して全体の性質を要約しているのである。しかも、これは単なる「代表例」の抽出ではなかった。孔子の学団では『詩』の本来の意味に関係なく、自らの言いたいことを『詩』を引いて言う「断章取義」の風があり、孔子がここに取りだした「代表例」も『詩』では異なる意味であると指摘されている（吉川幸次郎訳『論語』上、p. 50）。つまり、孔子のこの言葉は (1) 文献の全体を数によって把握し、(2)「代表例」をぬきだし、(3) 再解釈をくわえて、(4) 一言でまとめるということをしており、たいへん高度な情報処理の結果に似ている。

このように、大きなデータを小さくまとめるという困難は、現代の情報処理においても同様である。情報哲学を提唱しているフロリディ氏はこの問題について次のようにいう。

> ビッグデータに関する本当の認識論上の問題は小さなパターンにある。……巨大なデータベースの中のどこに本当に価値を付け加える新しいパターンがあるか、そこからいかに富を生み出し、人々の生活を改善し、知識を進歩させるために使うかを見出すことができるかが、圧力になっている。これはコンピュータの能力というよりも、知力の問題である。（ルチアーノ・フロリディ『第四の革命』先端社会科学技術研究所訳、新曜社、2017年、p. 18）

ここには膨大なデータから「小さなパターン」を取りだす「知力」の重要性が指摘されている。これは『論語』の「一言を以て之を蔽う」という思考に通じるところがあるのではないだろうか。

## 2．テクスト

古典には現代にいたるまでに伝承の歴史があり、長い歴史のなかで筆写されていくうちに多くの異本が作られた。この異本の間にはしばしば文字の差

異がある。よく知られている例でいえば、『老子』の「大器晩成」（41章）は、馬王堆帛書（乙本）や郭店楚簡などの出土文献によれば、前者は「免成」、後者は「曼城」に作り、いずれも「無成」に通じ、本来は「大いなる（無限の）器は完成することがない」の意味であった（蜂谷邦夫訳注『老子』岩波文庫、pp. 199–200）。また、『論語』学而の「貧而楽」（貧にして楽しむ）について、日本に伝わった本では「道」という文字がのこり、「貧にして道を楽しむ」と読む。しかし、定州漢墓竹簡『論語』には「道」の字はない（高橋均、2000年、p. 2）。加地伸行氏は十三経注疏本にしたがい、「貧而楽富而好礼」（貧にして楽しみ、富みて礼を好む）について、「かつて楽道（道を楽しむ）となっていたテキストもあったので、『楽（道）』と『好礼』はともに精神の向上とする」と注釈している（加地伸行訳注、p. 33）。

こうした異本の差異を検討し、本文を定めていくのが校勘という営みである。古典の注解はこの説明を示してある場合が多い。たとえば、金谷治訳注『論語』（1963年初版、1999年改訳）の場合、唐の開成石経（長安の国子監に建てられた12の儒教経典の石刻、837年完成）に由来する本と、明経博士清原家の証本や、梁の皇侃（488–545）『論語義疏』（中国では滅び、江戸時代に日本から中国へ輸出された）などを参照して、武内義雄（1886–1966）が定めた本文を基礎にしている。それは「中国と日本との最も由緒正しいテクストがここに統一されている」（「凡例」）とされる。このように古典の本文は自明のものではなく、多くの版本や史料を参照して緻密な作業をへて定められたものである。

現在、さまざまな形で公開されているデジタル・テクストにおいても、どのテクストによったかというのは重要な情報である。学術アーカイブは底本を示している場合が多く、底本の画像と対照できる場合もある。しかし、有名な古典の場合、世界中でさまざまな立場でデジタル・テクストが作られ、その間には相互に版本に起因する差異もあれば、音が似ている文字に変わっているものもあり、形が似ている文字に変わっているものもある。これがどのような原因によるのかは、底本が示されていないかぎり、簡単に判断をくだすことはできない。歴史上、文献の文字が変化をこうむることは、テクストが広まる過程でしばしば起こったが、デジタル・テクストにも「異本」の問題はあり、これを使って分析をする場合、一定の校正が必要になる。校正は何度かくり返すと見落としを発見することがある。

校正の問題は、テクストデータに数的処理をほどこす場合に、より重要になる。統計をとるためにテクストデータを作ることは、あとで文字や語を使われ

ている文脈から引き離して、トークン（表現型）からタイプ（抽象型・辞書型）に変換して集計するということである。日本語ではこれがはっきりしており、動詞や形容詞は「形態素解析エンジン」などで終止形にして統計をとる場合があり、西欧語でも動詞を原型や不定詞にもどして統計をとる場合がある。古典中国語（漢文）の場合は、データ上、トークンとタイプの区別はないようにみえるが、そのかわりに異体字の問題がある。たとえば、「群」と「羣」、「没」と「沒」、「歿」と「歾」などは字形が異なるのはもとより、情報処理上の実体である文字コードも異なるので、文字コードにしたがって単純集計を行うと別の項目として数えられる。デジタル・テクストを作るときに、異体字は原則的に底本にしたがうことになるが、避諱や欠画（貴人の名を憚って他の字にしたり、画を欠いたままにすること）など、一定の統一をしなければならない場合もある。

以上のような本文の問題にくわえ、文献内部の「まとまり」についても問題がある。テクストには篇・章・句などさまざまな段階の「まとまり」がある。現代に伝わる『論語』の場合、全体を20篇にわけるのは共通しているが、章については全487章・492章・513章などの分け方がある。たとえば、『論語』憲問の「作者七人」については、古注（魏・何晏『論語集解』）では前条とあわせて一章だが、新注（南宋・朱熹『論語集注』）ではわけている。つまり、章の分け方も自明のものではなく、本文の解釈を反映している。

このような「まとまり」をどのように扱うべきかという問題は、対象とする文献の「ありよう」にあわせて、比較的安定した「まとまり」を設定することになるだろう。だが、さまざまな分け方があるからと言って、先人の解釈によって示された「まとまり」をすててしまうのも惜しいと思う。文献にはいろいろな「分け方」があり、統計においてもどんな段階の「まとまり」が必要になるかは分からないからである。

## 3.「分け方」と「数え方」

統計をとる際に、ひとくちに文字や語を数えるといっても、どの範囲から何を数えるのかという問題がある。たとえば、『論語』に「仁」などの文字がどのように使われているかを問題にする場合、『論語』全体から数えることもできれば、何篇に使われているかという条件で数えることもでき、子路や子貢など、ある弟子の名がでてくる章だけに限って数えるということもできるだろう。このように、「分け方」と「数え方」は目的によって異なり、工夫次第で

多様である。

また、『論語』全篇から「有」という文字を数えれば、一般的に「有る」という意味で使われているほかに、「有子」「有若」「冉有」など人名の一部として使われる場合がある。「由」という文字についても弟子の子路（仲由）の名である場合もあり、「行不由径」（行くに径に由らず）のように動詞で使う場合もある。中国語は現代英語のように固有名詞を大文字で始めるという表記法がないので、統計の結果でてきた文字が人名や地名などの固有名詞の一部であるかもしれないことについて、注意を払う必要がある。この点について、現在、古典中国語を対象とする形態素解析（単語に区切ること）の研究が進んでいるようであるので、その進展に期待する次第である[11)]。

こうした「分け方」と「数え方」を使って、文献にふくまれる語に「重みづけ」をする方法が TF-IDF 法である[12)]。TF-IDF は「キーワード自動抽出」に用いる基礎的値であり、TF と IDF の二つの値をかけあわせたものである。以下、これについて説明を加えておきたい。

TF（Term Frequency）とは、ある一つの文書に特定の語（ターム）が何回出現したかということを示す値で、語の「局所的な重みづけ」として用いる。古典の研究でもテクストの内部に語が何回出現するかという点は言及されることがあり、最も基本的統計値であろう。

もうひとつの IDF（Inverse Document Frequency）とは、「文書頻度」（Document Frequency）の逆数である。「文書頻度」とは複数の文書をあつめた「文書集合」のなかで、タームが「いくつの文書に使われているか」ということを示す値である。この逆数である IDF は分母にタームの文書頻度、分子に全文書数をとり、この分数に対して対数をとる。IDF にはいくつかの計算式があり、情報検索の分野では分母に 1 を加えてゼロ除算をさけ、「文書集合」にないタームの抽出に対応する場合もある[13)]。この対応をしない場合、文書頻度が全文書数に等しい語は分数が 1 となり、対数をとる結果、IDF はゼロとなる。したがって、TF とかけあわせれば、TF-IDF 全体はゼロとなる。つまり、ある文書の中でどれだけ多く使われていても、「文書集合」のなかで他の文書すべてにある語であれば、TF-IDF 全体の「重み」はゼロである。いっぽう、「文書集合」のなかで一つの文書にしか使われていない語は IDF が全文書数に応じて最大となる。つまり、IDF は語を文書集合という比較的大きな範囲から「大局的」に重みづけた値であり、IDF が高いほど、そのタームは与えられた文書集合のなかで「珍しい」という性質がある。したがって、TF-IDF は「局所的」な頻度と「大局的」な「珍しさ」をかけあわせて、一つの文書内にある

語の「重みづけ」を求めるものである。

以上にみたように、TF と IDF は本来「語」に対して用いる。中国語においても「文字」と「語」は異なるが、漢字文献の基礎統計として、文字（ユニグラム）の統計をとる場合にも「分け方」と「数え方」にもとづいた「局所的重みづけ」と「大局的重みづけ」の考えは利用できるであろう。この場合、TF は CF（Character Frequency）とすべきであろうが、本稿では新奇な用語をさけ、TF のままとする。

小論で分析した現行本『論語』は20篇からなり、篇名は冒頭の 2 字または 3 字によってつけられている。これによって篇の区切りははっきりしている。つまり、『論語』は20篇の文書集合としてとらえることができる。であれば、篇のなかの頻度によって「局所的」に文字の頻度を計ることができ、20篇のうち何篇にあるかを数えることで「大局的」に分布を知ることができる。

また、中国語テクストの最小部分である漢字は表意文字の側面がある。アルファベットを文字ごとに数えても、それは音を表しているだけで意味を表さないが、漢字を文字ごとに集計すると、テクストにふくまれる意味の標識をある程度把握できる。そして、ある文書に使われている文字を字種ごとに分類して数えると、「異なり字種」の値が得られる。これは「その文書に何種の文字が使われているか」ということである。これを比較することで、文書の大まかな差異を知ることができる。

しかし、同じ程度の長さの文書で同じ程度の「異なり字種」の文書であっても、その分布が同じになるとは限らない。使用頻度が多い文字に偏っている文書もあれば、使用頻度が比較的なだらかに分布している文書もありうる。このような「集中度」をはかる値がウドニー・ユール（1871–1951）の「K 特性値」（以下、K 値）である[14]。K 値は、かりに文書のなかの文字がすべて異なる場合はゼロとなり、一種の文字しか使われていない文書があれば最大になる。つまり、K 値が高い文書は使用頻度上位の語がくり返し使われているのに対し、K 値が低い文書は使用頻度上位の「山」が低いかわりに、中位や下位の語が比較的多く使われているということになる。いいかえれば、前者では決まった言い回しや内容がくり返されている可能性があり、後者は決まった言い回しが比較的少なく、内容も分散し、多くの語が使われている可能性がある。

また、文書の「分け方」を利用して、文字や語が「どのグループに偏って出現しているか」ということを示す値もある。これが言語統計で使われる「カイ二乗値」である[15]。この値はグループごとの語の総頻度の比から、「理論頻度」（「期待値」）を算出し、この「理論頻度」と「実際頻度」（「観測値」）の偏りを

もとめた値であり、有意水準 p 値に対応するカイ二乗分布の棄却域と比較することで、有意な偏差のある語を抽出することができる。

古典中国語にふくまれる文字（ユニグラム）の統計をとった場合、「分け方」によって生じるグループの「長さ」の比によって「理論頻度」をみちびく。この「理論頻度」との偏りによって、あるグループに偏って使用されている文字をみつけることができる。これは文書じたいから得られる情報であり、キーワード検索のように人間の関心に依存することがない[16)]。この点は重要であり、読解を裏づける場合もあれば、読解で捨象された文字の偏りを抽出することもあるかもしれない。なお、カイ二乗値は偏りの程度をあらわすだけで、どのグループに偏っているかは頻度を参照する必要がある。

以上、「分け方」と「数え方」にもとづいた基礎的方法について述べた。その算出法に興味がある方は、参考文献や文末の附録を参照していただきたい。大きな文献でないかぎり、どの値も表計算ソフトで算出することができる。

## 4．基礎統計

以上に述べてきたところにしたがい、『論語』の基礎的統計をとった[17)]。底本は校勘情報にくわしく、一般的であるという点を考え、岩波文庫本（武内義雄＝金谷治本）によった。これによってデジタル・データを作り、校正を行った。なお、句読点は削って章の分け方はのこした。統計表（表 2）は一文字（ユニグラム）のものである。以下の記述はこの条件のもとであるという制約があるが、全体から部分に基礎統計から読みとれることを確認したい。

まず、『論語』全体はおよそ 1 万6000字の長さがあり、1350種ほどの異なり字種で成りたっている。そのうち、全篇で使われている文字（20種）の頻度を合計すると、およそ40パーセント（6300字余）であり、一つの篇だけで使われている文字の頻度を合計すると、 5 パーセント程度である。のこる55パーセントが複数の篇で使われているが、全篇に共通するわけではない文字である。

一般に情報検索では、総頻度が上位の語はそれがないと文章が構成できない語で、文法上の機能語であることが多い。文書の含意を決める内容語は、総頻度では中位以下に属する。

『論語』において全篇にあらわれている文字の内容は以下である（カッコ内は総頻度）。

①【子】(982)　②【曰】(763)　③【之】(614)　④【不】(584)

表２　『論語』（岩波文庫）の基礎統計

| no | 篇 名 | 篇 長 | | 異なり字種 | | 集中度 | | IDF＝0 | | IDF＝最大 | | 特徴 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 字数 | 偏差値 | 字種 | 偏差値 | K値 | 偏差値 | 合計 | 割合 | 合計 | 割合 | |
| 1 | 学而 | 498 | 37－ | 180 | 33－ | 145 | 48 | 213 | 43%＋ | 4 | 1%－ | **** |
| 2 | 為政 | 581 | 40－ | 214 | 41－ | 153 | 51 | 232 | 40% | 14 | 2%－ | *** |
| 3 | 八佾 | 691 | 45 | 241 | 46 | 142 | 47 | 275 | 40% | 49 | 7%＋ | * |
| 4 | 里仁 | 503 | 37－ | 168 | 31－ | 197 | 69＋ | 222 | 44%＋ | 5 | 1%－ | ***** |
| 5 | 公冶長 | 877 | 53 | 276 | 54 | 183 | 64＋ | 356 | 41% | 38 | 4% | * |
| 6 | 雍也 | 823 | 51 | 282 | 55 | 163 | 56 | 325 | 39% | 38 | 5%＋ | * |
| 7 | 述而 | 892 | 54 | 296 | 58＋ | 174 | 60＋ | 383 | 43%＋ | 33 | 4% | *** |
| 8 | 泰伯 | 617 | 42－ | 244 | 47 | 119 | 38－ | 238 | 39% | 30 | 5%＋ | ** |
| 9 | 子罕 | 813 | 50 | 284 | 56＋ | 130 | 43 | 290 | 36%－ | 31 | 4% | ** |
| 10 | 郷党 | 645 | 43 | 280 | 55 | 116 | 37－ | 187 | 29%－ | 99 | 15%＋ | *** |
| 11 | 先進 | 1073 | 62＋ | 292 | 57＋ | 164 | 56 | 420 | 39% | 42 | 4% | ** |
| 12 | 顔淵 | 1000 | 59＋ | 266 | 52 | 141 | 47 | 384 | 38% | 34 | 3%－ | ** |
| 13 | 子路 | 1041 | 60＋ | 268 | 52 | 163 | 56 | 454 | 44%＋ | 30 | 3%－ | *** |
| 14 | 憲問 | 1349 | 74＋ | 364 | 73＋ | 164 | 56 | 552 | 41% | 68 | 5%＋ | *** |
| 15 | 衛霊公 | 911 | 55 | 266 | 52 | 177 | 61＋ | 406 | 45%＋ | 23 | 3%－ | *** |
| 16 | 季氏 | 869 | 53 | 270 | 53 | 107 | 33－ | 293 | 34%－ | 44 | 5%＋ | *** |
| 17 | 陽貨 | 1021 | 59＋ | 309 | 61＋ | 147 | 49 | 396 | 39% | 44 | 4% | ** |
| 18 | 微子 | 626 | 42 | 238 | 46－ | 112 | 35－ | 211 | 34%－ | 47 | 8%＋ | **** |
| 19 | 子張 | 848 | 52 | 252 | 49 | 165 | 56＋ | 344 | 41% | 30 | 4% | * |
| 20 | 堯曰 | 373 | 31－ | 171 | 31－ | 115 | 37－ | 125 | 34%－ | 20 | 5% | **** |
| 1–10 | 上論 | 6940 | | 984 | | 136 | | 2721 | 39% | 341 | 5% | |
| 11–20 | 下論 | 9111 | | 1005 | | 139 | | 3585 | 39% | 382 | 4% | |
| 1–20 | 全体 | 16051 | | 1355 | | 137 | | 6306 | 39% | 723 | 5% | |
| | 最大 | 1349 | | 364 | | 197 | | 552 | | 99 | | |
| | 最小 | 373 | | 168 | | 107 | | 125 | | 4 | | |
| | 平均 | 803 | | 258 | | 149 | | 315 | | 36 | | |
| | 標準偏差 | 231 | | 47 | | 25 | | 103 | | 21 | | |

注１）＋は上位五位、－は下位五位、特徴は＋と－の個数
　２）上論・下論の共通文字の合計頻度：14905（634種）93%
　３）小数点第一位を四捨五入

⑤【也】（581）　⑥【而】（344）　⑦【其】（271）　⑧【人】（221）
⑨【以】（211）　⑩【有】（197）　⑪【矣】（187）　⑫【於】（173）
⑬【為】（171）　⑭【乎】（161）　⑮【君】（160）　⑯【如】（158）
⑰【與】（144）　⑱【無】（132）　⑲【言】（131）　⑳【問】（121）

じつに『論語』全篇の約４割がこれらの文字である。内容をみると、③④⑤⑥⑦⑨⑪⑫⑭⑰などは文法上の機能語（助字）として使われているらしいことがわかる。しかし、とくに①と②は『論語』の書き方にも関わる文字であるから、やや説明をくわえておきたい。

①【子】はもちろん単独で「先生」の意味で用いるほかに、「孔子」「曾子」「有子」などの一部としても使われ、「子路」「子貢」などの固有名詞の一部でもある。そして、⑮【君】と組みあわせて【君子】としても使う。【君子】という熟語も全篇に共通して106回使われている。解釈上、理想の人物を指すとされ、「学徳のできあがった人」（吉田賢抗注）、「紳士」（吉川幸次郎訳）、「徳のできあがった人」（金谷治訳）、「為政者としての適格者」（木村英一注）、「学問と志ある人」（藤堂明保注）、「教養人」（加地伸行訳）、「上層階層に属し、美的・倫理的に修練を積んだ人物を指すが……階層を超え、より広がりをもつ理想的人間像」（井波律子訳）などの定義がある。

こうした「君子」の内容を記述する指摘がある一方で、「君子」に代名詞的機能があるという指摘もある。宮崎市定（1901–1995）『現代語訳論語』に「身分のある男子が原義であるが……時には第二人称の諸君の意に用いる」（p. 3）とし、「諸君は器械になってもらっては困る」（同前「君子不器」p. 28）、「諸君は正義に敏感であってほしい」（同前「君子喩義」p. 65）と訳す。また「他者が孔子のことを指して君子と呼ぶ箇所もあり、孔子が自分のことを君子と考えているらしい文章も見られた」（前掲、湯浅邦弘、p. 163）という指摘もある。使用頻度が多いからといって、「君子」に代名詞的機能があるという根拠にはならないであろうが、「君子」という語は機能語のような分布で全篇にあらわれ、機能語に次ぐ使用回数があると言うことはできる。

②【曰】は前に①【子】や固有名詞をとって、発言をあらわす文字である。対話で発言者が明らかな場合は、前置部分を省略することもある。そして⑳【問】と共起する部分をみると、問答・対話という『論語』の基本構造の一つを示している。この「構造」を使って、くり返し使用されている「問〜曰…」の用例を抽出すると、次のような用例が得られる（カッコ内は頻度）。

問〔政〕子曰（6）
問〔政〕於孔子、孔子對曰（2）
問〔孝〕子曰（4）
問〔君子〕子曰（3）
問〔知〕子曰（3）
問〔曰何如斯可謂之士矣〕子曰（2）
問〔政〕於孔子曰（2）
問〔仁〕子曰（5）
問〔聞斯行諸〕子曰（4）
問〔於子貢〕曰（3）
問〔其次〕曰（2）
問〔之〕曰（2）

また、「問」の後の一文字を取りだすと、次の用例を加えることができる。

問〔事〕（2）：「問事鬼神」「問事君」の一部

問〔禮〕(2)：「問礼之本」「大哉問、礼……」の一部

これらの用例から構造をとりさると、〔政〕・〔仁〕・〔孝〕・〔君子〕・〔知〕・〔事鬼神〕・〔事君〕・〔礼〕など、『論語』で問題となっている概念の一部を抽出することができる。前述のように、一般に頻度が中位以下の語が文書の内容を決めるが、古典中国語で書かれたテクストから、人間が読解の際に問題にする内容語を、頻度だけによってくまなく自動抽出することは困難である。しかし、文書を書きあらわしたのも人間であるから、人間が問題にする点の標識がないわけではない。『論語』の場合、この標識の一つが⑳【問】と②【曰】で示される構造であろう。要するに、テクストには「問い」もふくまれていることがあるのだから、これを情報検索に利用することはできるだろう。

次に、前半10篇と後半10篇の基礎統計をみてみたい。前半は「上論」といわれ、後半は「下論」といわれる。この分類は伊藤仁斎（名は維禎、1627–1707）や崔東壁（名は述、1740–1816）以来の説で、武内義雄『論語之研究』（岩波書店、1939年）には、下論の「季氏・陽貨・微子の三篇を刪（けず）って、先進・顔淵・子路・憲問・衛霊公・子張・堯曰の七篇を以て又別種の論語と考えたい」と述べ、「此等七篇はその内容と用語とによって判断すると恐らく斉に伝わった論語であるらしい」としている（p. 151）。こうした方法論について、湯浅邦弘氏は出土文献の調査にもとづいて批判をくわえているが、上論と下論については「いくぶんその色彩が違うのに気づく」とも述べている（前掲、湯浅邦弘、p. 80）。

基礎統計から言えることは、まず、下論が上論に比べて30パーセントほど長いことは確かである。篇の長さに注目すると、上位５位までに属する篇（憲問・先進・子路・陽貨・顔淵）はすべて下論にあり、短い篇といえる下位４位までに属する篇（学而・里仁・為政・泰伯）は上論に属する。堯曰篇は特殊であり、最も短く下論に属して全篇の最後にある。異なり字種に注目すると、前半と後半に共通する字種は634種あり、この使用回数を合計すると１万4900字余であり、全篇の長さのおよそ93パーセントにあたる。これは一つの目安ではあるが、もちろん、文字の組合せは異なり、のこり７パーセントの文字とも組み合わさるので、これをもって一概に上論と下論の類似度とすることはできない。

上論で使われている異なり字種は984種、うち634種が共通するので、上論のみで使われている字種は350種である。同様に下論のみで使われている字種は371種である。上論に使われている字種の25パーセントは上論のみで使用

され、下論に使われている字種の27パーセントが下論の範囲で使われている。全篇に共通する字種の割合は上論・下論ともにおよそ40パーセントで、一篇にのみ使用される字種の長さの割合も大差はなく、およそ 4 〜 5 パーセント前後である。

カイ二乗検定によって、上論と下論に偏ってあらわれる文字を抽出すると、以下の文字を抽出できる。有意水準を0.001とすると棄却域は10.82となり、これ以上が有意な偏差である[18)]。

①【孔】上論頻度（12）　下論頻度（63）　　カイ二乗値　22.77
②【至】上論頻度（17）　下論頻度（ 3）　　カイ二乗値　14.23
③【孝】上論頻度（16）　下論頻度（ 3）　　カイ二乗値　13.01
④【夫】上論頻度（28）　下論頻度（77）　　カイ二乗値　11.83
⑤【衣】上論頻度（13）　下論頻度（ 2）　　カイ二乗値　11.54
⑥【知】上論頻度（69）　下論頻度（49）　　カイ二乗値　11.25
⑦【言】上論頻度（38）　下論頻度（93）　　カイ二乗値　10.90

それぞれ興味深いが、①【孔】、③【孝】、⑥【知】について指摘しておきたい。

①【孔】については、公冶長篇（上論）にみえる「孔文子」（1例）（衛の国の大夫）以外は孔子にかかわる。憲問篇（下論）にみえる「孔氏」（2例）は「孔の家」の意味であり、微子篇（下論）にいう「孔丘」（丘は名、3例）は、孔子の弟子と隠者の対話にある。のこる69例はすべて「孔子」と使われる。この分布をみると、上論では合計11例、いずれも 1 〜 3 例である。下論では合計58例である。つまり、「孔子」という用例は下論に偏って使われるのであるが、木村英一（1906–1981）訳注に季氏篇の「孔子曰」（14例）について「孔門外の世間の伝誦から出ている」とし、「斉における子張後学の編集かと思われる」（p. 429）と指摘している。定州簡『論語』では「孔子」と「子」の使い方は厳密ではない（前掲、湯浅邦弘、p. 63）。なお、孔子の字、「仲尼」は子張篇（6例）にだけみえる。孔子の呼び名は④【夫】にもいえ、これも下論に偏って出現している。「夫」にはさまざまな用例があるが、孔子を指す「夫子」（40例）が最も多い。上論では合計13例、下論では合計27例である。

③【孝】については、上論に偏って出現している。上論では学而（4例）・為政（10例）・里仁（1例）・泰伯（1例）であり、為政篇に集中して出現することがわかる（うち2例は『書』の引用）。学而篇にみえる4例のうち2例は「孝弟」であり、この例はほかの篇にはない。下論に「孝」は先進・子路・子張の各篇に1例ずつある。為政篇について、武内義雄『論語之研究』に「孔

子の教が家庭内における道徳孝を本として、之を社会全体に推しひろげることにあることを説いたものである」、「孔子の教が孝に出発している」（前掲、pp. 115–116）という。

⑥【知】については多様であるが、二文字では「不知」が最も多く、上論19字・下論 9 字で上論に偏る。孔子はしばしば「知らず」と言う人であり、「禘」という礼を問われた時は「知らざるなり」（八佾）と答え、三人の弟子について「其の仁を知らざるなり」（公冶長）と答え、自ら「老いの将に至らんとするを知らざるのみ」（述而）といい、「知らざるを知らざると為せ」（為政）と教える。また、『論語』の冒頭は「人知らずして慍みず、亦た君子ならずや」（学而）で終わり、学而篇は「人を知らざることを患う」で終わり、『論語』末尾はいわゆる「三不知」で、「言を知らざれば、以て人を知ること無きなり」（堯曰）で終わる。この冒頭と末尾の対応に、荻生徂徠（双松、1666–1728）は「編輯者之意」（『論語徴』癸、堯曰）を指摘している。

以上、上論と下論について一定の差異を指摘した。もちろん、すでに先学が指摘していることであるが、これを基礎統計によっても確認できる。

各篇の特徴にうつると、最も特徴的なのは里仁篇である。里仁篇は短い篇に属し、最も異なり字種が少なく、全篇で使用されている字種の合計頻度は全体よりやや高く、この篇でのみ使用される文字の割合は学而篇に次いで低い。特筆すべきは集中度（K値）が最も高いことであろう。これは頻度上位に属する字の「山」が高いことを示す。つまり、里仁篇はその頻度上位に属する文字を使った比較的一定の形式で文が成りたち、一定の内容を多く言及するということである。一言でいえば、「里仁篇は整っている」のである。この点について、各篇の成立を説明している木村英一訳注を要約すると、最後の章以外はすべて「子曰」を冠した孔子の「格言集」で、１～７章までがすべて「仁」の格言、８～９章が「道」、10～11章が「君子」、12～14章が「君子」の行動や態度に関する格言であるという（pp. 76–77）。

次に特徴的なのは、学而篇・微子篇・堯曰篇である。まず、微子篇と堯曰篇について記述しておきたい。この２篇も短い篇に属する。堯曰篇は『論語』のなかで最も短く、微子篇は平均よりやや短い程度である。異なり字種については２篇とも少ない篇に属する。特徴は字種の集中度にあらわれており、両篇ともに低い。字種の集中度が低いということは、一定の言い方が少ないということである。したがって、両篇とも全篇に共通する文字の比率も低く、微子篇は他篇にない文字が多い。このような特徴から２篇は特殊な篇であるといえる。微子篇には「乱世に身を処する君子」（木村英一訳注、p. 477）、つまり隠者が

でてきて孔子や門弟と対話している部分がある。これを「孔門の言行録ではない」という指摘もあり、これに意味をみとめる指摘もある。堯曰篇には古記録の断片があり、「附録という性質が残存している」（木村英一訳注、p. 525）と指摘されている。

学而篇は短い篇に属し、異なり字種は少ない。字種の集中度と、全篇に共通する文字の合計頻度はともにやや高いが平均的である。最も際だった特徴はこの章にだけ出現する文字の合計頻度が全篇で最も低いことである。つまり、学而篇はほとんど他篇にある文字で構成されており、独特な文字を使うことが最も少ない。この点で全体に親和的であり、『論語』を代表しうることが、基礎統計からも確認できる。学而篇のみで使われる文字の内容は以下である。すべて1例であるから『論語』全体の総頻度によっても抽出できる。

①【汎】1例　②【良】1例　③【磋】1例　④【琢】1例

③と④は子貢が『詩』を引いた部分にみえる文字であり、いわゆる「切磋琢磨」（『詩』衛風・淇奥）の一部である。これは『論語』以外の書物の引用であるから、引用がくり返されないかぎり、ほかの部分にみえなくても当然といえる。①は孔子（子曰）の言葉で「汎く衆を愛して仁に親しむ」の部分であり、副詞的に使われている。

興味深いのは②【良】であろう。これは子貢が、孔子その人について「温・良・恭・倹・譲」と5字で述べている部分にある。「良」は『論語』において他に例がないので、解釈上問題とならざるを得ない。

『論語』の「良」について、朱熹（朱子、1130–1200）は「易直なり」という（『論語集注』）。この「易直」はおおむね「まっすぐ」なことを指し、『礼記』楽記・祭義にみえる「易直子諒之心」にもとづく。孔穎達（574–648）らの『礼記正義』（653年頒布）に「易」は「和易」、「直」は「正直」、「子」は「子愛」、「諒」は「誠信」とする（土田健次郎訳注1巻、p. 100参照）。荻生徂徠は朱熹の注について、「大いに字義を失う」と批判し、「股肱良哉」（『尚書』益稷）・「良相」（『左伝』成公七年）・「良馬」（『周易』大畜等）・「良工」（『孟子』滕文公下等）・「良医」（『左伝』成公十年等）・「三良」（『詩』秦風・黄鳥）などの例をあげ、「みな材の良を以てこれを言う。良に豈に易直の義あらんや。温はその容なり、良はその材なり」（『論語徵』甲）という。これについて、本居宣長（1730–1801）の弟子で、徂徠の孫弟子にあたる鈴木朖（あきら）（名古屋の人、1764–1837）は「温良は相反せり。温は温和なり。良はしっかりと性のよいと云心にて、物の用に立べき材智器量をたしかに内に持（もて）るを云。さる人は多くは

温和にては得あらず。温和なる人には良材なき者多きに、温にしてしかも良なるを貴しとするなり」(『論語参解』原文カタカナ）と解説している。

なぜ、このようにさまざまな議論があるのかという点を考えれば、それはまず『論語』のなかにほかに用例がみえないからであろう。つまり、その篇だけにある文字は解釈を行うときに他の書物に開いていかざるを得ない文字である。統計では大きな部分が問題になることが多いが、漢字文献の統計上、頻度の少ない文字には注釈にひろがりがあるとみられる。したがって、頻度の少ない文字をみつけることも決して無駄ではない[19)]。

## おわりに

小論の方法は情報検索の面では、とくに珍しくも新しくもなく、むしろ古典的である。用いたツールも基礎的なプログラミング言語と、正規表現などのパターン検索、集計は表計算ソフトであり、ありふれたツールにすぎない。また、キーワード検索にその役割をゆずった『引得』類を使っても確認できることも多いかもしれない。

ともあれ、基礎統計は抽象度が低く、直接にテクストの性質を示している。したがって、基礎統計は「変形されたテクスト」といえるかもしれない。この「テクスト」を読み解くことで、もとのテクストを読む手がかりを得ることはできるだろう。長い研究の歴史がある古典について、統計によって新たな論点を発見することができるとは限らないが、すくなくとも事実としてのテクストの姿を写した基礎統計は、これによって疑問を起こし、先学の研究に親しむ出発点になるのではないだろうか。

ふえつづけるデジタル・テクストの時代に生きる一人として、筆者は「小さな関心」のもと、このようにデジタル・テクストを使っている。今後多様な関心に応じて、さまざまな「分け方」と「数え方」ができる公開がすすめば、デジタル・テクストの利用の幅もより広がるのではないだろうか。また、小論を書くにあたって先学の著作を参照したが、著作権保護期間が満了した古典研究も、いっそうデジタル化がすすめば、「温故知新」に大いに役立つと思う。

### 注

1）高度な情報抽出には「潜在的意味解析」や「自己組織化マッピング」などをあげられるだろう。「潜在的意味解析」は語の共起により次元を縮めてトピックをまとめ、直接キー

ワードをふくまない文書を抽出する情報検索に利用される。「自己組織化マッピング」に使われるニューラルネットワークは人工知能の学習モデルに使われるが、判断の根拠がブラックボックスになるという問題が指摘されている。ほかにも多変量解析の結果をしめす樹上図（デンドログラム）や多次元尺度構成法など、文章など高次元の構造を平面に描く方法があり、このような視覚化法はさまざまな文書で試みられている。ネットワーク図を描くツールには Pajek・GraphViz・Gephi などがある。これらのツールには媒介中心性（Centrality of Betweeness）などを算出できるものもある。拙論「『老子』の聖人と玄徳」（『漢字文献情報処理研究』第 8 号、2007年）に『老子』のテキスト構造を示すネットワークの一部を抽出してみた。

2）計量文献学の歴史は、村上征勝『真贋の科学―計量文献学入門―』（朝倉出版、1994年）、同氏「計量文献学の歴史と課題」（『計算機統計学』 9 巻 1 号、1996年）、土山玄「文学作品の計量分析―その方法と歴史―」（『情報処理学会報告』CH107、7 号、2015年）を参照。

3）後藤義乗「数理文献学的方法による無量寿経類漢訳者の推定」（『印度学仏教学研究』26 巻 2 号、1978年）、李賢平「『紅楼夢』成書新説」（『復旦学報』社会科学版、1987年第 5 期）。前掲、村上征勝（1994年）pp. 94–96 に李氏の手法の解説がある。

4）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年）を参照。なお、氏は「過去の言語体系に対しては、研究者の内省だけでは予想以上に理解の及ばない点のあることを認めねばなるまい」という（p. 221）。

5）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、2001年）。

6）谷本玲大「曖昧検索性を持たせた N-gram サーチの手法―『新撰萬葉集』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、2001年）。

7）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『東洋学へのコンピュータ利用』2002年）、「大規模仏教文献群に対する確率統計的分析の試み」（「中国宗教文献研究国際シンポジウム報告書」2004年）。なお、師茂樹氏は文献の距離について「ばねモデル」で視覚化する方法も提起している（『漢字文献情報処理研究』第 5 号、2005年）。

8）土橋喜『情報視覚化と問題発見支援―問題構造の可視化による仮説生成―』あるむ、2000年。

9）山田崇仁「中国戦国期の語彙量について― N-gram とユールの K 特性値を利用した分析」（『漢字文献情報処理研究』第 5 号、2004年）、「N-gram 方式を利用した漢字文献の分析」（『立命館白川静記念東洋文字文化研究所紀要』第 1 号、2007年）。

10）秋山陽一郎「『老子』傅奕本来源考―『項羽妾本』介在の検証―」（『漢字文献情報処理研究』第 4 号、2003年）。

11）安岡孝一・ウィッテルン＝クリスティアン・守岡知彦・池田巧・山崎直樹・二階堂善弘・鈴木慎吾・師茂樹「古典中国語（漢文）の形態素解析とその応用」（『情報処理学会論文誌』2018年）、また、ライデン大学では MARKUS が公開されており、中国語文献に人名・地名・官名などを自動マークアップできるようである（2019年、ベータ版）。

12）1990年代に TFIDF を改良した重みづけ法 BM25 が提案されており、短い文書から得られる情報について補正を行い、経験的に最もよい「重みづけ」とされている。松本裕治編『言語と情報科学』（朝倉出版、2011年、p. 67）参照。

13）酒井哲也「ランクつき検索」（『言語処理学事典』共立出版、2009年、p. 296）参照。
14）ユールのK特性値については、前掲、村上征勝（1994年）p. 59を参照。
15）カイ二乗値については『計量国語学』（1958年6号、pp. 29–39）。また、伊藤雅光『計量言語学入門』（大修館書店、2002年、pp. 104–111）を参照。
16）キーワード検索は知りたいことを検索することはできるが、興味のあることしか検索できないという問題点をのこす。直接に関係はないが、グーグルなどの検索エンジンの個人化についてはイーライ・パリサー『閉じこもるインタネット―グーグル・パーソナライズ・民主主義―』（早川書房、2012年）を参照。
17）宮崎市定『論語の新しい読み方』（岩波書店、1996年、初出1969年、p. 107）に「ごく大体のこと」と断った上で、「仁」が「九十七回」、「礼」が「七十五回」、「孝」が「十八回」で「忠」と同じとする。今回『論語』全篇の使用字数を調べた結果、「礼」と「忠」は同じで、「仁」は109字（里仁篇の篇名の「仁」を除く）、「孝」は19字という結果を得た。中央研究院「漢籍電子文献」の「上古漢語語料庫」で確かめた結果も今回の統計と一致した。
18）拙論「偽古文尚書の賢と官」（『漢字文献情報処理研究』第6号、2005年）にカイ二乗値を用いた同様の試みをした。
19）たとえば「異端」（為政篇）については、2字の熟語とみれば、『論語』にほかに用例がないが、「異」は総頻度11、文書頻度8、「端」は総頻度3、文書頻度3である。

## 参考文献

井波律子『完訳 論語』岩波書店、2016年
小川環訳注『論語徵』平凡社、1999年
荻生徂徠『論語徵』（関儀一郎『日本名家四書注釈全書』9巻、東洋図書刊行会、1926年）
加地伸行『論語』講談社学術文庫、2004年
金谷治『論語』岩波文庫、1963年初版、1999年改訳
木村英一訳注『論語』講談社文庫、1975年
朱熹『四書章句集注』中華書局、1983年
鈴木朖『論語参解』鈴木朖学会、1981年
髙橋均「定州漢墓竹簡『論語』試探」(2)、『大妻女子大学紀要・文系』32巻、2000年
武内義雄『論語之研究』岩波書店、1939年
土田健次郎訳注『論語集注』平凡社、2013年
藤堂明保『論語』学習研究社、1981年
宮崎市定『現代語訳 論語』岩波書店、2000年（初出1974年）
宮崎市定『論語の新しい読み方』岩波書店、1986年
湯浅邦弘『論語―真意を読む―』中公新書、2012年
吉田賢抗『論語』明治書院、1960年
吉川幸次郎『論語』朝日文庫、1978年（1959年序）

## 附録

1．重みづけ法：TF-IDF 法

$$\text{TF-IDF} = TF_i \times \log_2 \frac{N}{DF_i}$$

$N$：全文書数
$TF_i$：ある文書内のターム $i$ の頻度
$DF_i$：ターム $i$ の文書頻度

2．ユールのK特性値

$$S_1 = \sum x_i f_i \qquad S_2 = \sum x_i^2 f_i \qquad K = 10^4 \left( \frac{S_2 - S_1}{S_1^2} \right)$$

$x_i$：文書中の出現回数
$f_i$：$x_i$ 回出現した語の個数
$10^4$ は値全体が小さくなりすぎることをふせぐための係数

例：『論語』学而のK特性値

| x | f |
|---|---|
| 1 | 99 |
| 2 | 34 |
| 3 | 10 |
| 4 | 9 |
| 5 | 6 |
| 6 | 8 |
| 7 | 5 |
| 9 | 2 |
| 10 | 1 |
| 15 | 1 |
| 19 | 1 |
| 20 | 2 |
| 21 | 1 |
| 29 | 1 |

$S_1 = 1 \times 99 + 2 \times 34 \cdots\cdots 29 \times 1 = 498$
$S_2 = 1^2 \times 99 + 2^2 \times 34 \cdots\cdots 29^2 \times 1 = 4082$
$K = 10^4 \times (4082 - 498) \div (498^2) = 144.513\ldots$

3．カイ二乗検定

例：『論語』前半十篇（上論）と後半十篇（下論）の「孝」について

表1　使用頻度（観測値）のクロス統計表

| | 孝 | その他 | 合計 |
|---|---|---|---|
| 上論 | 16 | 6924 | 6940 |
| 下論 | 3 | 9108 | 9111 |
| 合計 | 19 | 16032 | 16051 |

理論頻度（期待値）の算出

・上論「孝」の理論頻度＝19×6940÷16051＝8.215…
・下論「孝」の理論頻度＝19×9111÷16051＝10.784…
・上論「その他」の理論頻度＝16032×6940÷16051＝6931.785…
・下論「その他」の理論頻度＝16032×9111÷16051＝9100.215…

カイ二乗値

$$\chi^2=\sum\frac{(\text{実際頻度}-\text{理論頻度})^2}{\text{理論頻度}}$$

$$\chi^2=\frac{(16-8.215..)^2}{8.215}+\frac{(3-10.784..)^2}{10.784}+\frac{(6924-6931.785..)^2}{6931.785}+\frac{(9108-9100.215..)^2}{9100.215}=13..01..$$

表2　自由度1のカイ二乗分布の棄却域

| 有意水準p | カイ二乗値 |
|---|---|
| 0.05 | 3.84 |
| 0.01 | 6.63 |
| 0.001 | 10.82 |

「孝」のカイ二乗値13.01は有意水準0.001の棄却域10.82より大きく、この水準で有意な偏りがあると判断できる。

なお、カイ二乗分布の棄却域や帰無仮説を設ける点については紙幅がないので、注15)の文献や統計を説明しているウェブページを参照していただきたい。