The University of Osaka
Institutional Knowledge Archive

Title	中国古典への〈N-gram分析〉応用に対する若干の疑問
Author(s)	井上, 了
Citation	中国研究集刊. 2006, 42, p. 29-38
Version Type	VoR
URL	https://doi.org/10.18910/61127
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

中国研究集刊 結号（総42号）平成18年12月 29 —38 頁

中国古典への〈N-gram 分析〉応用に対する若干の疑問

井 上 了

はじめに

近年、古典の使用語彙に対して統計的な調査を行い、あるいは「歴史言語学」的な技法を駆使して、その成書地域あるいは成書年代を定めようとする試みが積極的に進められている。この方法は、古くにカールグレンが行い鮮やかな成果を示したものだが、最近では吉本道雅氏や山田崇仁（たかひと）氏がこの方法によって多くの成果を上げている。とくに山田氏は、「N-gram モデル」なる手法で様々な文献を処理し、多くの文献の成立時期を特定している。

「N-gram」とは要するに“N字の連続”である。これは“N字の熟語”とは異なり、たとえば「天下国家」４字から 2 gram を取ると「天下」「国家」のみならず「下国」も抽出され３組の「共起」を得る。山田氏の方法は、文献全体から膨大な 2 gram を抽出して統計的に分析し、浮かび上がった文字／言語の使用傾向から文献の成書時期・地域を特定しようとするものである。『漢字文献情報処理研究』２（好文出版、2001年）に特集として「N-gram が開く世界」が組まれており、山田氏も「初めての N-gram 」として概説的な説明を寄せる。氏による具体的な作業については、氏の「『礼記』中庸篇の成立について」（『立命館大学東洋史学会叢書』四『中国古代史論叢続集』、平成17年）・「『周礼』の成書時期・地域について」（『立命館大学東洋史学会叢書』五『中国古代史論叢三集』、平成18年）など一連の論考を参照されたい。

筆者は、書誌学的な検討あるいは“思想史的”な検討のみならず、このように客観的な研究が進展することを大いに歓迎する者である。このような研究の最大の長所が、その反証可能性にあることは言うまでもない。誰にでも追検証できるデータと検証過程を公開することは、“人文科学”も“科学”である以上は当然必要な事であり、また追検証は行われなければならない。この点において筆者は山田氏の研究姿勢に最大の敬意を払っており、このような研究がひろく推進されることを願っている。

しかし筆者が追検証した範囲では、氏の方法および結論には賛同しがたい点も若干

認められる。本稿では、山田氏の「『礼記』中庸篇の成立について」を主たる例として筆者が疑問に感じた点を指摘したい。批判とは研究の進展の契機となり得るのであり、研究そのものへの批判は忌避されるべきではないと信じるが故である。以下における筆者の指摘に誤謬・誤解があれば、忌憚のない御叱正を賜りたい。

1　『中庸』分章のバラつきと分析母数の少なさ

山田氏は、『中庸』全体を分章した上で、各々について「共起頻度」を他文献と比較する。氏の分章は原則的に朱子と金谷治氏の分章に従っているが、このことによって各部分の長さにはかなりのバラつきが生じている。たとえば第20章ｃが250字、第33章が212字であるのに対し、第３章は13字、第５章にいたってはわずか８字であって、第５章「子曰、道其不行矣夫、」の 2 gram での共起は６組（「子曰」「道其」「其不」「不行」「行矣」「矣夫」）にすぎない[1]。

氏はこの６組を諸子書の用語と比較し、結果を「正規化」して、『論語』に対してはプラス0.18、『孟子』に対してはマイナス0.87という値を得るのだが（以下、「共起頻度」はすべて山田氏表２による）、これによってたとえば“『中庸』第５章の成書時期は『孟子』よりも『論語』に近い”などと言うことはできない。この分析結果は“該章のうち実に25％（共起数では約17％）を「子曰」が占める”という原因によるもので、さらに強いて言うならば、“『論語』全体に占める「子曰」の割合(約5.4％)が、『孟子』全体に占める「子曰」の割合(約1.6％)よりも多い”ということによる。むろん N-gram の原理上、孔子の発言としての「子曰」のみならず、「有子曰」「曾子曰」「季路見於孔子曰」「季康子問政於孔子曰」（『論語』）や、「孟子曰」「告子曰」「諸侯朝於天子曰述職」「老而無子曰独」（『孟子』）などが分析に影響を及ぼすことは言うまでもない[2]。

しかし氏は、39字しかない『中庸』第７章（しかも「子曰」と「中庸」とが10％以上を占める）に対してこのような分析を行い、『論語』に対してプラス0.51、『孟子』に対してマイナス0.14という結果を示した上で「(「共起頻度」が)『論語』は平均より多く、対して『孟子』は平均より少ない。従って第７章では『論語』の影響が強いことがわかる。」と行論するのである。「『論語』の影響が強い」とは、“「中庸」二字が『論語』に１例あり『孟子』には無い”ことを指すのだろうか？

氏は、朱子の章分けを採用した理由のひとつとして「比較的細かく分けられている事」を挙げる。しかし、小さな母数を対象とすればするほど分析の有意性が急速に薄

れていくのは、統計学上の常識であろう。むやみに「細かく分けられている」分章を採用するのではなく、より大きな分段や、あるいはたとえば“冒頭から五十字ごと”“百字ごと”といった機械的な分析などをも試みるべきではなかろうか。すくなくとも、このように小さい対象への統計分析によって「『論語』の影響」「『孟子』の影響」を云々することは危険である。

2　分析結果の恣意的な取捨

山田氏は『中庸』各章毎の『論語』・『子思子』・『左伝』・『孟子』・『荀子』・『墨子』（十論）・『荘子』・『韓非子』・『管子』（内篇）との「共起頻度」を一覧で示し、

> 現行本『中庸』が『韓非子』や『管子』がよく使用する言葉を使わず、儒家、特に『荀子』で使われる言葉を頻用する事が見て取れる。

という全体的な傾向を示す。また「共起頻度」による各章相互のグルーピングと断代を試み、『中庸』全体を４クラスターに分割してみた上で、たとえば

> 従来一括して扱われていた「中庸説」部分が、実は二つのそれぞれ異なる時期に成書されたという可能性を指摘し得る。
>
> （第3・7・8・9・11章は）現行本『中庸』の中で最も古い部分と想定される。
>
> （従来は一体として解釈されてきた第17～19章について）第17章は『左伝』より遡る前四世紀前半期のものであるとしてよい。……第18章は前三世紀の成書となろう。……第19章は『左伝』より遡る前四世紀前半期の成書であるとしてよいだろう。

などといった分析を『中庸』のほぼ全章に対して施していく。その行論は鮮やかで、氏の方法の有効性を具体的に示すもののごとくである。

しかし氏の作業については、細部に若干の疑義も提示され得る。

武内説のいわゆる「中庸説」に属する諸章について、山田氏が示す諸書に対する「共起頻度」は一般に低い。とくに第21～24章のそれは、第22章が『管子』に対してプラス0.03である以外はすべてマイナスの値を示す。このことは、これら諸章の用語が先秦諸子のそれとは異質であることを示し、『中庸』のうちこの部分の成立が『荀子』『韓非子』よりも新しいか、あるいは『論語』より古いことを（作業仮説上は）意味しよう。

しかし山田氏は、これら諸章に見える「至誠」「之性」などの文字列が『孟子』や『左伝』などにも見えることを個別に指摘し、これらの章が『孟子』『左伝』と同じく「前四世紀後半期の成書」だと結論づけるのである[3]。

むろん、統計分析を行った上で、補足的にこのような指摘を行って成書時期などを絞り込むことは認められてよい。しかし、ネガティブな結果を明らかに示すデータを特定の箇所についてのみ無視し、個別的な文字の一致のみを根拠として行論するのは、“得られたデータを(ポジティブなデータもネガティブなデータも)公平に取り扱う”という科学実験のルールからみて問題なしとしない。

3　共起単語の個別指摘

山田氏はさらに N-gram 分析を無視して、単語・用語の個別の一致を求める。

たとえば『中庸』第28章の「今天下、車同軌、書同文、行同倫。」は古くから秦始皇の統一に関するものと見なされ、『中庸』の成書を秦代以降とする根拠と考えられてきた。しかし氏は、第28章の「共起頻度」(すべてプラスだが、『論語』『孟子』『墨子』に対してはそれぞれプラス0.75、プラス0.94、プラス0.61と高く、『荘子』『荀子』『管子』に対してはそれぞれプラス0.27、プラス0.13、プラス0.03と比較的低い。第31章が『論語』『左伝』『孟子』に対して大きなマイナス値を示すのとは対照的) を無視し、第28章の用語が『墨子』『荘子』『荀子』『穀梁』など「前三世紀の成書が想定される諸文献」にも見えることを根拠として、第28章が「諸子が天下統一という近未来を見据えた前三世紀中頃までに盛んに唱えていたスローガン的な言葉を、積極的に取り入れたものなのである。」とする。

しかしこの行論は成り立たない。氏は『中庸』第28章「行同倫」から「同倫」のみを切り出し、この二字が『穀梁』にも見えると指摘するのだが (ただし『穀梁』の「同倫」は名詞であって、『中庸』の「倫を同じうす」とは異なる語)、もしもこのような断片的な一致によって『中庸』第28章の成立時期を決定するのであれば、たとえば「同軌」が『左伝』(隠公元年) や『韓非子』(内儲説上・八経・八姦) に見えることによっても『中庸』第28章の成書時期を決定できなければならない[4]。『中庸』第28章が『穀梁』と同じく前三世紀の成書であり、また『左伝』と同じく前四世紀の成書であるというのは矛盾であって、このような矛盾の発生は、行論過程に誤謬があることを示唆しよう。筆者の単なる印象としても、もし仮に『中庸』が「近未来を見据えた」いわば予言であったとしても、いまだ天下が統一されていない時点で「今天下、車同軌、書同文、

行同倫。」と書き、これを教団の内外に示すといった事態は想定しがたいと感じる。

またたとえば、『中庸』第16章に見える「微之顕」について氏は

> 第16章は「微之顕」を第32章(井上注:「第33章」の誤)と共有する。この「○之顕」は『荀子』儒效・王霸・君道、『韓非子』顕学、『呂氏春秋』当染・勧学・必己にそれぞれ用例が見られる。これらの文献は何れも前三世紀後半の成書にかかり、……少なくとも第一クラスター(井上注:基本的に前四世紀中期の孟子より以前の成書とされる)とは関係が薄いことはこの点からも明らかである。

と指摘し、最終的には第16章の成書時期を「前三世紀中期(下限 『荀子』『呂氏春秋』以前)」とする。しかし一方で氏は、『中庸』第33章に見える「微之顕」については

> 「微之顕」は『左伝』昭公十二年に「微而顕」と類似する文句が見え(井上注:『左伝』昭公十二年に「微而顕」なる句は見えない。成公十四年や三十一年の「春秋之称、微而顕、」を誤り引いたものか)、これは「表記」の「君子隠而顕」を踏まえていると想定され、従ってこの言葉のみから第33章の時代を伺(ママ)うことはできない。

と述べる。もっとも氏は、第33章から『荀子』『韓非子』などと共通する語彙をいくつか指摘し、結果的には「第33章も前三世紀のものであるとしてよい。」とするのだが、それにしても同じ「微之顕」について、一方では『荀子』との関係を云々し、また一方では「表記」を踏まえているとするのは“場当たり的”との印象を禁じ得まい[5]。

4 現行本を対象とする N-gram 分析そのものの限界

山田氏が分析対象とするテキストは、いうまでもなく現行本である。現存しないテキストを分析対象とすることは何人にも不可能であり、現行本を用いること自体は当然である。しかし、現行本を分析して得られるデータはあくまでも現行本の特徴を示すものにすぎない。“重層的な成立”が想定される『周礼』を例にとって、このことに注意すべきを述べてみたい。

山田氏は『周礼』の 2 gram での共起頻度を調査し、五官の間に用語のバラつきが小さいことなどを示して「検討の結果、『周礼』は他の先秦諸子の如き編纂の重層性は見られず、一人乃至は少数の人数で短期間に編纂されたとする結論を得た。」と述べる。このような飛躍は、にわかには承認しがたい。

そもそも氏が示したバラつきの小ささは、五官すべての共起頻度の上位に「二人」「四人」「下士」「中士」など序官にかかわる語が来ていることに起因する。これは『周礼』が最終的に現在のそれに近い形へと編纂された、その際における形式統一の徹底を示すものである。この作業が「短期間」に行われたことは、用いられた素材の「重層性」を否定しない。氏は「『周礼』は……パターン化された方式で記述されてはいる」としつつ「各官の個別具体的な職掌の記述に関しては……互いに矛盾を来すものもある。」と認める。そもそも一人が短期間で（ゼロから）作文したのであれば、「互いに矛盾を来す」ことはあり得まい。この「矛盾」について氏は、「『周礼』の重層性ではなく、編纂に際して複数の材料が取り込まれている事に起因するものである。」と述べるのだが、「編纂に際して複数の材料が取り込まれている」とはすなわち「『周礼』の重層性」そのものではなかろうか？

すくなくとも、氏が 2 gram 分析によって示す『周礼』の用語の「バラツキの少なさ」は、『周礼』が取り込んだ「複数の材料」相互の間にバラつきが小さいことを示し得ない。現行本『周礼』の全体にちりばめられた「二人」「四人」など「序官」にかかわる用語が 2 gram 分析に対して甚大な影響を与えるため、単なる 2 gram 分析では「複数の材料」間の「バラツキ」の有無を見ることなど出来ないのである[6]。

5　文献の「成書時期」ということ

山田氏は、「書物を記述する際の言葉は、それを記述する者（著者・編者・記録者）の意識的無意識的な認識や用語の選択を反映しているものである。」とし、「従って、言葉の傾向を分析する事で、著作物の時代や地域を弁別する（少なくともそれぞれの先後関係を明らかにする）事が可能となる。」とする。前半部分については筆者も同意するが、後半については無条件に認めることはできない。著者がまったくオリジナルな書物を新たに作る場合であれば、その書物の用語は著者の時代や地域を完全に反映するであろう（たとえば、『周礼』が先行する「材料」を全く取り込んでいないか、あるいは編纂時に「材料」の語彙を漏らさず編纂者の語彙に書き換えたという場合）。しかし現実に我々が手にし得る書物は、決してそのようなものではない。『周礼』は「複数の素材」を取り込んで素材相互の矛盾を存しており、またたとえば『史記』は『左伝』（の藍本）などに取材して語彙の一部を自己のものに改めつつ一部は素材の語彙をそのまま用いる。山田氏の方法で『史記』の使用語彙を分析すれば、『史記』の「成書」は戦国時代だと

結論づけられはしないだろうか。

たとえば、古く東方でほぼ成書していた『中庸』に、秦儒が「今天下、車同軌……」句などを挿入し、漢儒が全体を「整理」したとすれば、用語の分析によって得られる「成書時期」「地域」とは、どの段階でのそれを指すのであろうか。現行本『中庸』の「国」字が漢諱「邦」を改めた可能性は山田氏も認めるが、ならば現行本『中庸』の用字を分析した結果導かれるべき「成書時期」は漢代以降でしかあり得まい。

そもそも筆者は、一般に用いられている「文献の成書時期」という概念に対して、本質的な疑問を抱いている。たとえば“『周礼』の基本的な成書は（あるいは、素材の大部分の成書は）戦国期だ”とする主張と、“『周礼』の最終的な完成は漢代だ”とする主張とは、まったく矛盾しない[7]。大部分の先秦文献の「成書」を前漢末（劉向校書）とする主張は充分に成り立ち得るし、さらに言えば、現行本『中庸』（『礼記』中庸篇）の編者は顔師古であって、その「成書」は唐初だとさえ主張し得よう。

幾重にも整理が施された伝世文献に対して「歴史言語学的手法」による分析を行うことには、常に一定の危険がつきまとうのである。

6　文献の年代の基準となる文献の年代

山田氏は『中庸』全体の成書を戦国期とし、この結論に殉じる形で第28章の「今天下、車同軌、書同文、行同倫。」を統一以前の儒家が「近未来を見据えた」予言だとする。また氏は『山海経』海内南経を戦国期の成書としつつ、「匈奴」という秦代以降の言葉が「追加」されたとする[8]。これらの原因は筆者のみるところ、氏が基準としている文献をすべて戦国期のものとみなしている点にある。

氏が『中庸』との「共起頻度」比較に用いた『論語』『子思子』から『韓非子』『管子』までの九種、および個別的に共通の単語を指摘した『公羊』『穀梁』『呂氏春秋』『商君書』などは、氏によるとすべて前五世紀末から前三世紀末の間の成書とされる。この範囲の内でいくら類似や一致を指摘しても、結果として得られる「成書年代」はやはり前五世紀から前三世紀の間のいずれかとしかなるまい。分析対象が必ず戦国期の成書だという確信（先入観）を抱いているのでなければ、戦国期の文献に対してのみならず、漢代の文献に対しても同様の比較検討を行わねばならぬはずであろう。

また山田氏は、『左伝』の成書を前四世紀、『公羊』『穀梁』を前三世紀だとしてこれらを断代の重要な指標とする。しかし現行本『左伝』『公羊』などが漢代のすくなく

とも整理を経ていることは常識であり[9]、とくに『穀梁』全体を戦国中期とする説は承認しがたい。坊記や檀弓を前四世紀初とする説は必ずしも広く承認されているとは言えず、『墨子』の時期についてはなお議論が絶えぬし、『韓非子』には韓非以前の文献が混入している可能性も指摘されよう。

戦国期だと断代した文献との類似を根拠として別の文献を戦国期と断代し、これらを基準としてまた別の文献を戦国期と断代する。これを繰り返しておれば、あらゆる文献が戦国期の成書と推定されていくのは当然で、ついには統一以前の儒家が「今天下、車同軌、書同文、……」と述べたという奇妙な結論に至ってしまうのである。

おわりに

「歴史言語学的」な分析は、比較的大部な文献の成立時期等を考えるためには有効な方法であると筆者は考えている。しかし、零細な文献あるいは細かすぎる分章に対しては、これが有効な方法であるとは思えない。上掲「特集」において山田氏自身が「N-gram は万能ではない。N-gram は、ある特定パターンの文字列が頻出するテキストの分析が得意であり、本文中に紛れ込んだ注釈の文章を見つけだすような作業には向かない」と述べている。現行本の中から成立時期の異なる部分を「見つけだす」作業、とくに『中庸』のごとき短い文献に対しては、N-gram 分析は「向かない」のではなかろうか。これが疑問の第一点である。

山田氏は、いわゆる「中庸説」部分にあらわれているような明瞭な偏差であっても、自説にとって都合の悪い(と筆者には見える)結果であればこれを無視し、恣意的に選択した少数の文字の一致によってこれらの部分の成立年代を決定している。氏が「補正」と称するこのような資料操作・方法に対する疑問が第二点である。

現行本を対象とした分析によって示されるのは、現行本が「完成」した時期とは限らない。また山田氏のいう「素材」と「編纂」、あるいは「成立」と「追加」との関係は、(すくなくとも筆者にとって) なお充分には明らかでない。疑問の第三点である。

文献の使用語彙の相互比較による前後の判定は、どこまでも相対的なものである。『孟子』『荀子』などといった成立時期について比較的コンセンサスの得られている文献のみを指標とするならばともかく、『論語』→『礼記』(坊記・緇衣・曲礼等) →『左伝』→『孟子』→『公羊』→『穀梁』→『荀子』という成立順を前提とする分析には危惧を禁じ得ない。第四点である。

さて、あまりに細かな傷まで求めすぎたが、目録ではなく文献そのものの内容から成書時期や地域を決定するという作業は魅惑的であり、「文献に使用されている用語の統計的分析」はこのための強力な武器となり得る。筆者がその可能性を認めていることはあらためて言明しておきたい。しかしこの方法は危うさを秘めてもおり、「熟語」レベルの統計ではなく N-gram による機械的な処理を行うことによって、この危険は飛躍的に増大するように見える。筆者は「歴史言語学」や N-gram 統計などについて必ずしも詳しくないが、この方面における研究の先端を走る山田氏であれば、上記の危険についてはおそらく諒解されるであろう。上記のような危険を回避し得る手段・方法が講じられた上で、さらに客観性・説得力に富む研究が進められることを、心から期待したい。

注

(1) 山田氏は「句読点を跨ぐ「共起」は集計対象から外している」と述べ、『中庸』本文は岩波文庫本を用いたとするので、「曰道」は排除されることとなる。ただし氏は、『中庸』以外の諸子書については「句読点」を採用した底本を示していない。

(2) このような分析を行う際には、機械的な N-gram 分析よりはむしろ孔子の発言としての「子曰」のみを抽出して計量すべきであろう。近藤泰弘氏・近藤みゆき氏は、日本語の N-gram 分析が「単語や複合語や言い回しを調査することができる」のに対し、漢文におけるそれが「形態素や単語の連続を採取したこととなる」ことに注意を喚起している（上掲「特集」）。

(3) たとえば「能尽人之性。」の「之性」と「自誠明、謂之性。」の「之性」とを区別しないままで、これらと「天地之性」（『左伝』）・「聖人之性与中人之性」（『成之聞之』）などとの一致を指摘するのは、あまり意味がないように筆者には感じられる。

(4) 「同倫」二字が『穀梁』に見えることを指摘する以上は、「同軌」二字が『左伝』に見えることを、少なくとも指摘だけはしておくべきである。これを伏せたままで「一部『孟子』で使われる以外は、何れも前三世紀の成書が想定される諸文献で使用されている。」とする山田氏の行論は認めがたい。

(5) 後出の文献が古い文献を「踏まえて」作文される可能性を認めるならば、個別的な一致・類似によって「成書時期」が近いとする山田氏の論法そのものが殆くなろう。

(6) また山田氏は、「（『周礼』内部の）バラツキの少なさは、従来言われている冬官が

散逸して他の篇に紛れ込んでいるという指摘も否定するものである。」と述べる。しかしもしも冬官が他の五官と同様に「二人」「四人」など序官に関する文字列を豊富に含むものであったならば、これが他の篇に紛れ込んでいたとしても、 2 gram 分析によって検出されるバラつきはやはり小さなままとなるだろう。

(7) 筆者が「閩」字の使用を『周礼』秦漢期成書説の根拠とする（本誌30号「『周礼』の構成とその外族観」2002年。ただし「成書」の概念が山田氏と筆者との間で異なるようなので煩瑣を厭わず指摘しておくならば、「井上氏の『周礼』秦漢交代期成書説」とは山田氏の表現であって、筆者は「編纂に際して戦国以前の素材が利用された可能性は指摘されているが、現行本に近い形の『周礼』が成立したのは秦漢期だとするのは、現在ではほぼ定説だと言ってよいだろう。」と表現している）ことに対して山田氏は、『呂氏春秋』に先行する『山海経』海内南経にも「閩」字が見えると指摘し、「閩」は戦国期には既に用いられていたとする（筆者は「(海内南経の) この部分は……秦代以降に整理されたものであろう」と述べた）。しかし海内南経を戦国期とする氏が、一方では海内南経に見える「匈奴」の語を「漢代以降に係る他篇への追加部分が紛れ込んだもの」と処理し去るのは理解しがたい。海内南経の戦国期における「成書」と漢代以降における「追加」との関係を氏はどのように考えるのだろうか。また海内南経の「閩」字が（直後に見える「桂林」「番隅」とともに）秦代以降に「追加」された可能性はいかに処理されるのだろうか。

(8) 上記注(7)参照。

(9) 吉本道雅氏「左伝成書考」(『立命館東洋史学』25、2002年）は『左伝』の事後予言の中否を検討し、たとえば文公六年「君子是以知秦之不復東征也。」を秦献公による東方進出再開（氏は前364年とする）より以前に書かれたと判断すべきことなどを指摘して、「このことは、『左伝』が前364年以前に基本的に成書したことを示すとともに……それ以降の「附加」が、前364年以前に成立した『左伝』の基本的な枠組みに大幅な変更を加えない、正に部分的な「附加」に過ぎなかったことを雄弁に物語っている。」とする。重要な指摘だが、しかしこの現象は『左伝』の素材となった説話（の一つ）が当該時期に成立したことを示すもので、現行本『左伝』全体の成書や用語傾向について規定するものとは言えない。現行本『左伝』が“外れた予言”をそのまま存していることについては、「附加」（いわゆる「劉氏宣揚」問題も）への評価を含めて更なる議論が必要であろう。