『連邦主義者』の著者問題と計量言語学の諸方法

新井皓士

一の一

『連邦主義者』とは、一七八七年十月二七日より一七八八年五月にかけて、ニューヨークの新聞紙上に「パブリアス（Publius）」なる匿名の論者によって、ほぼ三日毎（原則として火曜と金曜）に発表された七七篇の論説シリーズを原型とする、アメリカ合衆国連邦憲法擁護論である。発表紙は『インデペンデント・ジャーナル』『ニューヨーク・パケット』『デイリー・アドヴァイザー』の三紙であり、掲載紙面の都合もあろうか、同一テーマに関して同日に発表されたものが二紙に分割されている場合もある。各論の長さは二千語前後（最小九〇六語、最大三五五一語）、「ニューヨーク邦民（People of the State of New York）」に呼びかける形で、一七八七年九月一七日に制定公布された憲法（案）の批准を訴え、「カトー（Cato）」なる匿名のもとに憲法案批判を展開するクリントン知事などに反論し、憲法案の理を説くものであった。フィラデルフィアで制定された憲法案は、ニュー・ハンプシャー、マサチューセッツ、ロード・アイランド、コネティカット、ニュー・ヨーク、ニュー・ジャージー、ペンシルヴェニア、デラウェアー、メリーランド、ヴァージニア、ノース・カロライナ、サウス・カロライナ、ジョージアの十三邦のうち、九邦以上の批准をまって初めて発効することになっていたからである。

新聞紙上の連載がなお続く一方で、一七八八年三月及び四月に、これら一連の論説文の大半が、『ザ・フェデ

ラリスト』の書名を冠する二巻（三六篇＋三九篇）の書にまとめられて公刊されている。『連邦主義者』という名称はここから生じているが、現行の『連邦主義者』にみられる第七八篇から八五篇までは、定本とされる一七八八年「マクリーン版」刊行に際してつけ加えられたものである。

「パブリアス」という匿名の主が、アレクサンダー・ハミルトン（1757-1804）、ジェームス・マディソン（1751-1836）、ジョン・ジェイ（1745-1829）の三名であることが明示されたのは、一七九二年にパリで刊行された仏語訳本である。しかし全八五篇のどの部分が誰の筆になるものかは必ずしも直ちに明らかにはならなかった。ニューヨークを活動の場としたハミルトンが企画者ないし主唱者で多くの篇を書いたことは明らかであるが、合衆国財務長官を務めた彼は一八〇四年七月に決闘で倒れ、また後に大統領となるマディソンも晩年まで沈黙を守ったからである。その結果三人の分担についてベンソン・リスト、ケント・リスト、ギデオン・リストなど互いに食い違いを含むリストが発表され、論議と歴史的研究を刺激することとなった。今日ではダグラス・アデーア（Douglass Adair）などの考察をふまえた一応の定説として、ジェイが第二から第五までと第六四の計五篇、マディソンが第十、第十四、第三七から第四八までの計十四篇、第十八から第二十までの三篇はマディソンとハミルトンの合作、著者特定に関して従来論議が多かった第四九から第五八までと第六二、六三の計十二篇もおそらくマディソンの作で、残りの四三篇がハミルトンの作とされるが、「おそらくマディソン作」の部分についてはなお若干の異論が残っている。

政治史や政治思想の視点からのみでは最終的判定が困難な、この著者識別問題に対しては文体統計学の立場からも関心が寄せられたが、中でも一九四一年および一九六三年に共同研究の成果を発表し、当該の問題解決への寄与のみならず、計量言語学方法論の上でも注目されたのは、フレデリック・モステラーである。特にデーヴィッド・L・ウォレスとの連名でアメリカ統計学会誌に一九六三年六月に発表された論文「著者問題における推理――『連邦主義者』の帰属解明に応用された識別方法の比較研究」は、より詳細な統計資料等を網羅した研究書『連邦主義者――その作者問題と推論』（増補改訂版『ペ

イズ定理に基づく推論と古典的推論の応用法――連邦主義者の場合』）として翌年刊行され、問題の十二篇がマディソンに帰属することをほぼ決定的に論証するとともに、計量言語学的研究にとって極めて多くの有益な示唆と指摘を含む優れた方法論的成果であるといえよう。

一の二

いわゆる人工言語や数理言語は別として、一般に自然言語を計量的に扱う場合、その目的と方向は大雑把にいえば二つに分けることができる。第一は言語の諸要素をあえて数量的に処理し一般的法則を導きだそうとするものであり、第二は従来の人文学的方法のみでは解決できなかった歴史上あるいは文体上の諸問題に統計的手法を応用する試みである。『連邦主義者』の著者問題は後者の事例であるが、この方向ではユールに代表される記述統計学の諸方法、カイ自乗一様性検定などを応用する推測統計学の諸方法、それにモステラー等が示したベイズ推論の応用が主な分析手法のグループをなしている。前者の領域では、まず文字と音節に関する分析があり、語彙の法則性に関する帰納的な研究と推論があり、言語の系統に関するスワデシュ（Swadesh）等の言語年代学があり、最近ではコンピュータに関連し言語統計模型を構築する為の様々な大規模分析がなされている。構文解析（パージング）あるいは生成文法に関連した研究も多少とも計量的側面をもつといえるかもしれない。

文字や音節に関する定量的研究は、暗号解読、速記術、モールス信号、タイプライターの機能的文字配列などとも関連して発展したが、対象言語が表音文字、表意文字、音節文字のいずれを用いるかによって、方法も結果も異なるのは当然であろう。アルファベット系ではたとえばgがごく僅かならばスペイン語と推測できるし、英語でも独語でもeの使用度が一番高いが二番目となると英語ではt、独語ではnとなる。八ヶ国語を分析したW・フックス（Fuchs）は文字より音節（シラブル）を単語等の分析単位とすべきであるとし、ドイツ系の学者はそれにならっているが、英米系では文字を単位とするのが普通である。同一シラブルをもつ単語間の間隔を調べる研究がある一方、プーシキン作品における母音と子音の交替に着目したマルコフ・プロセスという研究成果も知られており、またシラブル構造のエントロピーを求

める研究もある。「(単語は) 音節数が増すにつれ音素数は相対的に減少する」というメンツェラートの法則 (Menzerath's law) と称されるものは、ドイツ語以外の言語にどの程度あてはまるか検討を要するだろうが、少なくとも音節文字を使用する日本語に直接あてはめることはできまい。ちなみに、七つの母音を有した奈良朝以前はともかく、平安時代以後について五十音図のア段、イ段、ウ段、エ段、オ段に分けて和歌集や古文を資料に出現文字頻度を調べてみると、ア段が三割近くを占め以下オイウエの順になるが、漢字かなまじりが主流になる鎌倉期以降の日本語分析にも、音節文字（かな）のみで表記可能な特性を考慮し統計資料として活用すべきであろう。

語彙に関しては、単語の使用度数と順位に関するジップの法則 (Zipf's law) と称されるものがあり、情報理論の立場からの検証や、精度を高めるためにB・マンデルブロや水谷静夫その他による修正案も提案されている。順位情報を過度に重視すると一般に結果が厳密なものにはなりにくい面があると思われるが、その点「一著者が一文書で用いる異なり語数は、その文書の延べ語数の平方根に比例する」とするモリスとチェリィ (Morris & Cherry) の一見単純なテーゼは順位情報に左右されず、今後様々な実例に即して再検討する価値があるかもしれない。P・ギロー (Guiraud) はこれに対して一定のサンプルから語彙総量を推定する式を立ててランボー等について検証し、D・R・マクネイル (McNeil) はジェームス・ジョイスを例にとって作者の語彙総量を推定する試みを行っている。いずれも今後様々な例に当てはめてその定式の一般性を検証する必要があろう。

単語の定義ないし語の認定もいささか注意を要する。右の段落でも異なり語数と延べ語数という表現を用いたが、水谷は更に単位語と見出し語という分類を定義し、見出し語の集まりを語彙、その数を異なり語数と呼び、これに対して文中に出現したままの語を単位語とし、その数を延べ語数と呼んでいる。つまり見出し語は単語の原形ないし基本形にあたり、単位語は活用形にあたるといえよう。これは実際にはかなり厄介な問題で、英語のように語形変化の少ない言語でも厳密には同型異義語に細心の注意を払わねばならないし、たとえばドイツ語のように語尾変化ばかりか、形容詞や動詞の名詞化などが

頻繁にみられ、そのうえ分離動詞などが存在する言語では、統計量の算定にあたり単位語（出現形）で考えるか見出し語（基本形）で考えるかによって、集計作業も結果も相当に異なるものとなる。語彙を検証する場合は当然見出し語が基本となり、変化形などは基本形にもどして処理する必要があるが、それにつけても名詞化したものの扱いなど問題が残るのである。日本語の場合は単語の区切りそのものにも機械的には処理しきれぬ面があることはしばしば指摘されるところである。

アルファベット系の言語の場合、単語の長さに着目し、文字数あるいは音節数による階級（クラス）別の頻度を調べることは、T・C・メンデンホール（Mendenhall）が光のスペクトルに模して「単語のスペクトル」と呼んだ頃には膨大な作業と注意力を要するものであったが、コンピュータの登場と進歩のお陰で今日では簡単なプログラムを組むことにより個人にも可能な作業となった。得られたデータに基づきカイ自乗検定やt検定など推計学の手法を応用する研究が一九六〇年代から盛んになり、著者不明問題の解明（たとえばマーク・トゥエインとQCS文書）や盗作問題（たとえばショーロホフの『ドン・コサック』）の究明などに応用されているが、その一方で単語長の利用に関する方法論的反省もあり、ジャンルやテーマの影響を考慮すべきこと（たとえば四文字語の多いシェークスピアと三文字語の多いベーコンの差はむしろジャンル差であり、作者の個性ではないこと）、無意識裏に使われ統計処理になじむ程度に頻度の高い前置詞や代名詞などを中心に考えるべきことなどの提言もなされている。

G・U・ユール（Yule）はこれに対して一作者ないし作品の語彙集中度を測る指標としてK特性値と呼ばれるものを提唱した。これはまず各語の出現頻度を調べ、語の出現回数についてその一次モーメント、二次モーメントを基に、全文がすべて異なる語から成り立つ場合は値が〇、ただ一個の語から成り立つ場合は値が九九九九となるように工夫し定式化したものである。このK特性値はしかし特に十八世紀頃に多く見られる型にはまった文章には有効ではないとして、ある単語について、ある著者の書中の相対出現頻度と同時代同ジャンルの書中から採取した百万語中の相対出現頻度との比を識別指標と定義し、その値が一（すなわち時代平均）より相当大き

くなる語（プラス・ワード）、〇に近い語（マイナス・ワード）がその著者の特徴を示すと考えたのが、A・エレゴール（Ellegård）である。実際にはそのような特徴をもつ語の出現頻度自体が十分大きくならない場合が生ずるため、必ずしも個々の単語の識別指標が直ちに有効に働くわけではなく、グループ化などの工夫が必要になるが、何よりもまずバランスのとれた百万語の対照語群を作成することが前提である。百万語といっても今ではさして難事業とは思えないが、十九世紀末にW・ケーディング（Kaeding）は千九一万余語を数え上げ、ドイツ語単語の頻度辞典を作りあげている。それによれば冠詞、接続詞、代名詞など出現頻度上位三十語だけで全体の三一・八パーセントを占め、上位二〇七語をとると五四パーセントを越えるという。しかもこの場合の単語は先にふれた単位語であり、冠詞などの変化形は別語として数えられているから、見出し語にすればその比率は更に高まることになる。

単語の次元では右に述べた語彙や出現頻度を精査し定式化する試みの他に、特定の品詞や連語に着目する研究がある。この方向は特にギリシャ語文献を対象として比較的古くからなされているもので、H・レーダー（Raeder）は偽書説のあるプラトンの『第七書簡』に関し、またA・Q・モートン（Morton）は『パウロの書簡』に関する研究で、いわゆる不変化詞の頻度平均や位置を調べている。前置詞、接続詞、冠詞、そして特定の副詞などはテキストの内容にあまり関わりなく無意識に使われることに着目したのである。この視点はモステラー等の研究では、統計分析上危険をはらむ「テーマ依存的」(contextual) な語を避け、「埋め語」(filler words, function words, non-contextual) を重視する考えにつながるものである。反対にたった一度しか出現しない稀少語（Hapaxlegomena）に注目する立場もあるが、この場合はテキスト周辺のさまざまな状況証拠と組み合わせて特徴をとらえようとする、いわば定性分析的な考えであるといえよう。

文字、音節、単語と並んで、文の次元でも計量的研究は早くから行われている。たとえば文の長さに着目すれば、語の区切りが明白なアルファベット系の言語の場合少なくともコンピュータ出現以前は、単語の数で文章長を計測する方が、単語長を計測し分類するより、単純な

観測誤差を犯す可能性が小さい。「単純な」というわけは文の長さを測る場合に引用をどう扱うかによって結果がかなり左右されるからで、しかも引用の仕方には個人差やヴァリエーションがあるゆえ、包括的指針をたてることが案外難しいのである。日本語の場合は単語の切れ目が視覚的にも明らかでなく、といって分節を単位とすれば多少とも主観の入る余地や文語と口語、「です、ます」体か否かなど、別種の問題が生ずるので、ふつう漢字かなまじり文をそのまま文字単位で測ることが多いようである。当然そこに漢字の使用頻度という指標が考えられ、また前川守によると代表的現代文学作品では漢字一文字は仮名に直してせいぜい二文字に相当するというが、日本語における使用では音声の要素より視覚的抽象的要素が強く意識される漢字を、一度すべて日本語特有の音節文字「かな」に置き換えて観測することも必要ではないかと思われる。もともと古典文学の代表とされる『源氏物語』など漢字の使用度は極めて低かったのであり、また文章語でも音読を意識して書くかどうかは文体に当然反映するからである。なお日本語では句点（。）もさることながら、読点（、）の使用に特徴がみられる、とする村上征勝の指摘は、アルファベット系文章のコンマについても検証する価値があると思われるが、句読点のルールは洋の東西を問わず一般には比較的新しいものであり、古くは筆写や植字のくせ、新しくは編集方針の影響を無視できないことも一応考慮する必要がある。

ユールが『キリストに倣いて』の作者問題に関連して、文の長さの平均値、中央値、四分位範囲を調べ、同一作家の作品ではこれらがほぼ一定しているというテーゼを立てたことは広く知られている。しかしこのテーゼは、複数の作者が互いに意識的に似たような文を書く場合や、手本あるいは特定の流行スタイルに倣って文を書いている場合には、有効性を失ってしまう。十八世紀はそのような例が多いようで、前述の識別指標を提唱したエレゴールが、「ジュニアス（Junius）」なる匿名の著者をサー・フィリップ・フランシスであるとほぼ断定した『パブリック・アドヴァタイザー』紙上の「ジュニアス書簡」（1769-1772）がその一例であり、『スペクテイター』紙の文体に倣うといわれる『連邦主義者』もその例に数えられよう。文の長さにはまた時代的変化もあるようで、英文学では時代が下がるにつれて短くなりつつあるとさ

れ、ドイツでも一般に十七世紀十八世紀は文が長いが、個人次元でも年とともに文が短くなる一般的傾向とは別に、ゲーテなどは年齢とともに文が長くなることがK・グロース（Groos）によって指摘されている。また理由は不明だが、クリストファー・マーローの場合十一語から成る文が極端に少ない事がL・ウーレ（Ule）によって確認されている。

語順が比較的安定している現代英語にあっても文の先頭に主文の主語が置かれるか他の要素がくるか観測することは必ずしも無駄とは限らないが、ドイツ語の場合は主文と副文いずれが前に置かれるかで語順にも影響があるので、一応注目する必要がある。同様に純粋な機械的処理は不可能だが、ドイツ語などでは副文における動詞群の語順、あるいは受動文の比率なども、場合によって調べる必要があろう。因みに現代ドイツ語では、一般のテキストにおける受動文の比率は十五パーセント強だが、官庁語では二六パーセントにのぼり、そのうえ更に代替受動に属する文が十四パーセント加わるとされている。

一の三

一九四一年に当時まだプリンストン大学の大学院学生だったF・モステラーが、F・ウィリアムズに誘われ一緒に『連邦主義者』全文の文章長を数え上げた時、二人はむろん手作業でそれを行い、数え上げの困難という経験則を身をもって知り「フラストレーション」を味わったという。一九五九年に再びこの作者問題解明に取り組んだとき彼は「高速コンピュータ」の威力と少なからぬ人々の援助を享受することができた。すなわち一方ではパンチカードにテキストをタイプする作業、他方では専門家による単語数え上げのプログラム作成が行われ、機械処理によって打ち出された文書毎の単語の頻度（および出現箇所）を計算器による手作業で最終集計したのである。更に作者問題の的になっている文書十二篇は手作業によるデータ集積も並行的に行われた。すなわちローリング・ペーパーに一行一語の形でタイプし、後でそれを一語ずつ切り放し、「紙片が飛び散らぬよう息をひそめるようにして」並び替え、集計する作業である。コンピュータによる結果と手作業による結果を突き合わせて検査する為でもあった。

こんなことをあえて紹介したのは、計量言語学的研究

にあっては、テキスト・データの入念な準備が極めて大きな比重を占めるからであり、またそれがコンピュータの進歩によって格段に迅速に行いうるようになってきたからである。筆者自身六年ほど前『カルストハンス』の匿名作者を究明すべくカイ自乗検定を応用したとき、パーソナル・コンピュータに正確にテキストを入力したり、PCフォートランでプログラムを組むのに苦労しつつも、先人の苦労を思えばなんとありがたい時代になったものかと感じたものであるが、今ではそれすら一昔前の笑い話になりかねない。というのは今回筆者は『連邦主義者』のテキスト・データを揃えるに際し、主としてインターネットを利用することができたし、不足する一部はCDディスクに収められたテキストから転写し、一部は実験的に光学読取機で「百五十周年記念版」テキストを読みとらせデジタル化することができたからであり、こうして得たテキスト・データをいわば修正ないし編集加工するのみで筆者自身の作業は事足りたからである。

とはいっても、九九パーセント正確なデータも残り一パーセントを訂正するには慎重な検証が必要で、スピード・アップと肉体労働軽減がさしあたって進歩による直接の恩恵である。インターネットのゴーファー・サーバを通じて筆者はミネソタ大学のプロジェクト・グーテンベルク『フェデラリスト・ペーパーズ』の大半を入手したのであるが、ここにこの恩恵に対し謝意を表するとともに、通信を通じて得たテキストには思わぬ制御記号などが紛れ込んでいる可能性もあることを指摘しておかねばならない。またこのテキスト・データベースは逐次ないし複数の手で並行的に入力されたものらしく、『連邦主義者』に関しては文書配列が通常のものとは異なっているので、定本と照らし合わせて確かめる必要がある。

OCRにより読みとったテキストもむろん完璧ではないし、CDテキストは著作権保護上簡単にコピーできないようプロテクトがかかっているのが普通であるから、少なくとも学問的利用の為にはインターネットの存在は大変貴重であり、また我々自身も無償のテキスト・データを積極的に提供するようにし、インターネット本来の学問的相互協力に参与すべきであろう。

テキストの加工や集計に関しても、元はユニックス上で開発されパソコン用に移植された、いわゆるフリーソフトのプログラム言語やエディタでほとんどのことが可

能である。筆者は今回はもっぱら「セド（sed）」「オーク（igawk）」「グレップ（cgrep）」を利用し、先人の発表しているスクリプト（プログラム）を見よう見まねで少しずつ変えては、その時々の目的に適った加工や集計を行って『フェデラリスト』の文体解析に利用した。モステラーとウォレスのほとんど完璧な研究成果があるにもかかわらず、念のため『フェデラリスト』中のマディソン文書、問題の十二篇、ハミルトン文書の一部（第九、十一、十二、十五―十七、三五、三六、五九―六一、六五―六九）を標本とし、文章長、単語頻度、K特性値その他の基本的統計値を確認し、できれば既に得られている成果を更に補うような作者問題解明の手がかりを探りたかったからである。

一の四

「パブリアス」の正体は三人の分身から成るが、そのうち問題の十二篇を担当した可能性はマディソンとハミルトンに絞られ、この点では諸家に異論がない。ただこの両者の文体はいわゆる「スペクテイター・スタイル」で互いに酷似しており、また問題の十二篇は全体を一つのブロックとみなせば文の数は七百を越え単語数も一万八千を越えるが、その各篇は文の数にして最大八五、単語の数にして最大三〇二〇にすぎない。この標本規模はたとえば単語長や文章長の頻度分布によるカイ自乗一様性検定を適用した場合、よほど顕著な差がなければ本来異なるものでも一様性の仮説は保留されることが経験的に明らかで、標本は文章では最低百以上、単語数なら一万位がないと信頼できる結果を得られない。実際マディソンとハミルトンの文体は極めて似通っており、文章長に関してはウィリアムスとモステラーによれば平均が三四・五九および三四・五五、標準偏差が二〇・三および一九・二（筆者の標本による計算ではそれぞれ三四・一七、三四・五九、二一・五九、一九・六七）となり、彼らの第二の「フラストレーション」の因であった。単語長に関しては筆者の標本ではマディソンが平均値四・八九、ハミルトンが四・八〇、標準偏差はそれぞれ二・九〇と二・八八である。文章長と単語長をそれぞれ十五の階級に分けて一様性検定を行ってみると、文章長については危険率（有意水準）五パーセントおよび一パーセントで一様性仮説が保留されるが、単語長については仮説

はめでたく（?）棄却される。しかしこれは単語数がいずれも三万を越えるブロックとしての結果であるから、問題の十二篇それぞれを対象に同様の手法を試すことは無意義である。

モステラーとウォレスの研究も、この著者候補両名の文体の類似性と識別すべきテキスト（標本）の規模の認識から始まっている。と同時に彼らはこの著者問題をして統計学を用いた識別の諸方法を体系的に比較する絶好の「ケース・スタディ」とみなし、とりわけトーマス・ベイズ（1702-1761）を祖とするベイズ推定法の応用とR・A・フィッシャー（1890-1962）が基礎づけた「古典的」方法との対比研究を試みたのであった。二十世紀半ば頃から再評価されつつあるベイズ定理を「データから得られる証拠を事前情報と結合させる数学的デヴァイス」と呼ぶ彼らは、（著者）判別に用いる単語の選定および使用率推定にベイズ推定法を応用して得られた結果と、線形判別関数による「古典的」方法の結果を対比させ、更にベイズ定理に基づく「ロバストな」アプローチをも試みて、問題の十二篇はいずれも「確信といえる程度に」マディソンを著者とするという総括的結論を得ている。

一般に何か特別な識別子が存在すれば著者問題は簡単である。たとえばある著者は必ずある語ないし句を余人とは異なる形でさりげなく使い、かつそれが一定の頻度で定期的に出現し、しかも他人がそれを模倣したり手を加えたりすることができない場合である。マディソンとハミルトンに即して言えば、whilstという語はマディソン独特で、ハミルトンはもっぱらwhileを用いることに歴史家アデーァが気づき、モステラーに知らせているし、モステラー達も手始めの分析でuponとenoughはハミルトン語でマディソンはほとんど使わないことに気づく。しかし千語あたりに直して約三回弱の出現比率となるuponはまだしも、他の語は出現の頻度が低い為これだけをもって直ちに判定材料とするわけにはいかない。uponにしてもマディソンが絶対使わないわけではないから、問題の十二篇中にそれが現れても、それをもってその篇をハミルトン作とすることはできないのである。モステラー達はそこで両者の使用率が異なる語を相当数集めることによって、「潜在的識別子」のプール、換言すれば変数のセットを用意することとした。むろん

それは単語と限る必要はないのだが、量がある程度多いこと、観察（数え上げ）しやすいこと、そして分布に関する仮説が一般的に認められやすいことが、単語を選ぶ理由である。「各語の出現率は変数とみなしうるゆえ、語は数千の変数のプールを供給する」と彼らは言う。

有効な語のプールを作成するために、英語一般の機能語に関する既成の研究や『フェデラリスト』およびそれ以外の二人の著作からとられた延べ語数二十万強（約六七〇〇の見出語リストなど）の標本データがふるいにかけられ、識別子となる可能性をもつ語は最終的には一六五に絞られる。この一六五語は大別すれば三つのグループになり、そのうちの一つは冠詞、前置詞、接続詞、関係詞など統語的役割をもっぱら担う機能語である。他の一グループは出現頻度の高いもの、および頻度はそれほど高くなくともハミルトンまたはマディソンに特徴的な標識語（marker word）である。いずれにしても、テキストの内容に左右されない（非文脈的 non-contextual）ことが選定の重要な基準であり、数次のふるい（screening）にかけられ精選される。またベイズ推定法では分布の型が重要な役割を占めるから、この作業と並行するかたちで、これら非文脈的単語分布は統計モデルとしてはポアソン分布より負の二項分布に近いことが一応確かめられている。

但し正確なデータ分布とそのパラメータは不明である。ベイズ推定法を使用するにはパラメータの事前分布が明確でなければならない。モステラー達はそこで分析を二段階に分けベイズ定理を二度使うことによって問題を解決した。第一段階ではハミルトンまたはマディソンのいずれか著者が明らかな文書を分析し、単語使用比率について得られた事前情報を変換してこれらパラメータの事後分布とする。これが第二段階におけるパラメータの事前分布とみなされ、著者不明文書の分析が行われて全データに基づくパラメータの事後分布が求められる。但しモステラー達は計算と直観的把握を容易にする為に、著者をハミルトンとする仮説とマディソンとする仮説の事後確率の比をオッズ（final odds）の形で示すので、パラメータの事後分布まで算定する必要はないとしている。この手順の理論的根拠は『ベイズ定理に基づく……』の第四章において詳細に論証され、また先にふれた一六五語のプールはここでは三〇の「最終リスト」に縮約され

て負の二項分布、ポアソン分布と組み合わされ利用されている。そしてこのような複雑な手順を経て得られたハミルトン文書、マディソン文書、著者不明文書の「対数オッズ」が、問題の文書はほとんど疑問の余地なくマディソンによるものであることを示したのである。

一の五

モステラーとウォレスの研究の優れた点は、ベイズ定理の応用をはじめとして識別問題に関する統計学的アプローチの諸方法を駆使した体系的規模と精密さにあることはいうまでもないが、ハミルトンおよびマディソンが書いたことの明らかな資料を常に二分して、その半分（選定セットscreening set）をもって識別子または手法を確立し、まず残りの半分（調整セットcalibrating set）にこれを適用して補正した上で、最終的に著者不明文書に適用する慎重さも見事であり、もって範とすべきものである。それは（未知の平均値と分散の代わりに中央値と範囲を用いて算出した）ウェイトと使用頻度の積の総和から成る線形判別関数（linear discriminating function）による「古典的」方法の場合も同様で、そこでは最終的に二十語を選別して利用している。この手法および二分法を基礎とする「ロバストな」ベイズ推定法（ここでは三一語が選定される）は、大規模な計算器使用や複雑な手順も少なく、統計学にさほど詳しくない一個人でも十分応用可能であると思われ、特に「計算尺でも可能」とされる後者は方法論的にも魅力的である。むろんこれらから得られる結果は先述の方法による結果にくらべ精度はやや落ちるようではあるが。

方法としての簡潔性ないし経済性という視点からモステラー達は更に、分割表（contingency tables）を使って識別子となる語（六三）を選定し各語のログ・オッズを定めて、著者不明文書を分析する手法も試みている。この結果も他の三手法と同様に、もともと共同執筆の可能性があり語数も少ない論説第二十と、問題の十二篇中でも「下院議会」の定数を論じる特殊性の為か他の分析でも判別度が多少弱い論説第五五以外は、マディソン執筆者説を明確に支持している。またシーザーなる匿名で発表され一説によればハミルトンが真の著者であるとされる「シーザー文書」（引用をのぞき二六九八語）について、ベイズ推定法を応用した結果、むしろマディソン

の可能性こそあれ少なくともハミルトンの可能性はない、という推定がなされている。この場合は『連邦主義者』の場合のように二者択一ではなく他の著者の可能性もあるので、必ずしも断定はできないが、用いられた識別法の威力と精度は十分発揮されている

筆者はこれまで三千語程度の文書の著者識別は、僥倖に類する手がかりがある場合以外、一般的、体系的にはほとんど不可能に近いと感じていたが、モステラー達の研究は隘路を打開する有力な方法論として光明を与えるものである。もちろんこの場合はハミルトンおよびマディソンに問題の文書以外に多くの言語資料があり、かつ二人の内いずれかが真の著者であると想定できるという事情があって、他の資料がほとんどなかったり著者候補が多数想定されうるような場合とは違うわけだが、ここに開示された厳密な手法は計量言語学にとって一つの有力な指針となろう。このような研究を前にして些細なことにふれるのは些か気がひけることだが、最後に筆者が気づいた点を幾つか挙げてこの稿を閉じることとしよう。

まずユールのK特性値について。『連邦主義者』中の問題の文書（Dと略す）、マディソン文書（Mと略す）、ハミルトン文書抜粋（Hと略す）について、それぞれのK特性値を暫定的に求めてみると、Dは一八三・六、Mは一八八・九、Hは一九〇・二となった。また前置詞だけをとりあげてK特性値を計算すると、Dは二〇一四・一、Mは二〇六九・八、Hは二三二八・八となった。値の桁が違うのはK特性値は語彙の集中度をはかるものであるから、二〇ほどの前置詞グループにおけるそれと、すべての種類の語を含んだ場合との違いであるが、いずれの場合も集中度の順位が同じである点が注目される。前置詞の中ではofの使用頻度が高く、次いでtoとinとなるのは予測されるところであるが、四番目のbyについてはHの使用度が相対的に低い（Dは六・六%、Mは七%、Hは四%）のが気になるので二・二分割表によるカイ自乗検定をおこなったところ、DとMは差異なしとする仮説が保留されるが、DとH、MとHは仮説が棄却される。byの使用は受動態の文の頻度とも関係すると思われるので、これを調査しかけたが、意外と時間がかかり現在進行中である。また関係代名詞としてのthatの使用にも差異がみられ、Hの使用率が高いように思われるので、関係代名詞としてのthatとそうでない機能

の that の頻度、および関係代名詞 which と that の頻度に一様性がみられるか検定したところ、これはMとHでは有意水準五%、一%とも仮説が棄却されるが、DとM、DとHでは仮説は保留される。当時は関係代名詞としては which の使用が圧倒的に多く、Dでは関係代名詞 that の用例が二桁そこそこの少数であることも影響しているのかもしれない。いずれにしてもこれらは問題文書を一つのブロックとして扱ってのことであり、その意味では各篇の帰属問題解決にはほとんど寄与することができない。また be had という比較的珍しい（と筆者には思われる）受動態パターンがあったので念のため調べてみると、Mには一例も見られず、Hに四例、Dに一例があった。句では as well as がMに比較的多くHの倍の比率であり、as far as は逆にHがMの倍で、Dはいずれもほぼ中間になる。ざっと調べた限りでは他の句に大きな違いはみつからなかったが、いずれにしてもモステラーが指摘しているように、句の次元では出現頻度が低すぎて厳密な統計的な判別には簡単には利用できないようである。なお単語についてもモステラー達は、great plan と Great Britain、代名詞の I とローマ数次の I を特に区別せず、実際最終的な識別にこれらは無関係であるからそれでよいようなものの、一般的には固有名詞や数詞は別扱いすべきであろう。

文字の出現頻度、二文字連接の頻度、三文字連接の頻度についても念のため調べてみたが、今のところ格別な発見はない。ついでながら「コンピュート（算定する Compute）」間連語をあたってみたところ、『連邦主義者』全編で八回出現し、そのうち二回は問題の第五五に含まれていた。

宮野教授の退官を記念する号なので、ひょっとしてハイネの文中に、ハミルトンあるいはマディソンの名が出ぬか調べたいと思ったが、意余りて時足らず、代わりに偶然ぱらぱら覗いていたゲーテ往復書簡集に、マディソンの名をみつけた。一八一七年九月七日付けで、エヴェレット（Edward Everett）という男がゲッティンゲンからワイマールのゲーテに送った英文の手紙である。もしゲーテから依頼があれば、モンロー、アダムズ、ジェファソン、マディソンの書物を喜んで集めます、というのである。代わりに（当地）大学図書館を充実させる為

に是非御著書を御寄贈されたく云々、とある。ヴィルヘルム・マイスターではないが、容易に政治思想を語らぬゲーテも新興アメリカ合衆国、連邦主義と分邦主義、自治思想などに多少とも関心があったのか、あるいは世界文学を唱えた彼一流の漠とした関心の一端か、当方の想像も朦朧としている。

テキストとして次の二点のみあげておく。

＊ *The Federalist*—A Commentary on THE CONSTITUTION OF THE UNITED STATES being A Collection of Essays written in Support of the Constitution agreed upon September 17, 1787, by The Federal Convention (Sesquicentennial edition, with an Introduction by Edward Mead Earle, Washington)

＊ Applied Bayesian and Classical Inference—The Case of *The Federalist* Papers (Frederick Mosteller & David L. Wallace, 2nd Edition of *Inference and Disputed Authorship: The Federalist*, New York 1984)

筆者がこの問題に興味をもったのは、『数学セミナー』（一九八九・二）の村上征勝氏の記事がきっかけである。あえて記し謝意を表しておきたい。L・ウーレの『クリストファー・マーロー全作品コンコーダンス』には作品間の距離を測る興味ある方法論が示されているが、ここには紹介しえなかった。また筆者がヴェルツブルク大学（ドイツ）で一時的に教壇（そのものはなかったが）に立っていた時、博士論文として提出され回覧されたもので、トーマス・マンの作品を中心に「作家の文体的指紋」をつきとめる意気込みの試みがあった。出来映えそのものは必ずしも満足できるものではなかったが、方法論的批判検討をまじえて紹介しようと思いつつ果たさなかった。怠け者の常として、他日を期して筆を置くことにしよう。

（一橋大学教授）