計量言語学の方法と実践

新井皓士

一 数詞と偶然

幾つかの事例を通じて、言語文化現象を計量的に探求する意義についてお話します。

初めに、文末にある資料の表（01）を見てください。すぐ気づかれるように、これは十四の言語における一から十までの数詞の表で、特殊文字の符号（変音記号など）を省いてアルファベット二六文字のみを使い、上から順に、英語、スェーデン語、デンマーク語、オランダ語、ドイツ語、フランス語、スペイン語、イタリア語、ポーランド語、ハンガリア語、ラテン語、ロシア語、ギリシャ語、日本（和）語を並べたものです。元の資料はポイヤ（POLYA, George）が『発見的推論』で論じ、安本（美典）が『言語の科学』で敷衍したものですが、[1] ラテン語以下四語は私が付け加えたものです。表を縦に（つまり同じ数詞ごとに）見ると、何となく似ているものとはっきり異なったものがあることは歴然としています。それもそのはず、このうちハンガリア語と日本語を除く十二語は、東はインド、西はヨーロッパに広がる印欧語族と呼ばれるグループに属し、その中ではさらに英語からドイツ語までがゲルマン語派、フランス語からイタリア語までがロマンス語派（ラテン語から発展したもの）、ロシア語とポーランド語がスラブ語派として下位グループを形成しています。どうしてそんなことが言えるのか、詳しいことはここでは省きますが、グリムなどに代表される十九世紀の言語学者による文法や音韻の研究成果で

あることは間違えありませんし、この印欧語族に関しては、サンスクリット、ギリシャ、ラテンといった古くからの文献記録が残っていて厳密な比較が可能であったことが、その系統関係を明らかにする好条件でした。

数詞は語彙としては文化語ともいえるもので、共通の祖語から変化したものか、他の有力言語から借用語として入ってきたものか、微妙なところがあります。日本語を考えてみても我々は普段「いち、にぃ、さん……」と数えますが、これは明らかに漢語の借用であり、少なくとも万葉時代から「ひとつ、ふたつ、みっつ……」という数え方（「つ」は数詞の語尾）の記録があり、また日常的には「ひぃ、ふぅ、みぃ」とも言っておりました。表ではスペースをつめる必要もあって、四まではこの日常的な数え方を入れてあります。というのも、ここではそれぞれの数詞の語頭字音に注目してみたいからで、その点ではどちらでもおなじだからです。首尾一貫しないようですが、語頭の音は、語中や語尾にくらべると、変化ないし摩滅しにくい、という一般的傾向の実例でもあり、また文字は比較的古い音を伝えていると考えられます。

しかしヒトが意味を伝達するため発音し判別理解できる音声（音韻）の枠はある程度限られています。たとえば世界の言語を見渡しても、母音の数は三種から七種位が最も多いのです。となると、異なる言語といえども偶然似通う単語があることは予想されますし、十個の数詞をアルファベットで表し頭文字にのみ着目すれば一致する可能性はさらに大きい、とも思われます。たとえば頭文字が英語の場合と一致するケースを数えてみると、スェーデン語やデンマーク語は十回のうち八回、イタリア語は四回、日本語でも一回あります。イタリア語がフランス語やスペイン語、ラテン語と九回一致するのはサスガですが、ドイツ語は元来近いはずの英語と四回しか一致しないようにみえるのに（実は歴史的な音韻推移を考慮した対応関係を加算すると八回になります）対して、ロシア語はスペイン語と六回も一致する反面オランダ語とは一致〇回、また他とは断然かけ離れて一致度が少ないハンガリア語も北欧語や日本語と二回は頭文字が一致しています。ギリシャ語は地中海に面する言語とは三、四回一致しますが、英独などとは一回しか一致しません。日本語やハンガリア語は印欧語族に属していませんし、

文化語としての借用関係もここでは考えられないので、いわゆる「記号の恣意性」からして[2]、偶然の一致がここには混在していると思われます。このような場合、一体十回のうち何回位までの一致が偶然に起きる可能性があるのでしょうか。

ポイヤはこの問題について数学的モデルを設定し確率論的に考察していますので、われわれもその方法をとり確かめてみましょう。数詞は十個で言語は十四ですから頭文字は総計百四十となります。いま二つの袋に百四十個ずつ玉を入れ、その玉には百四十の数詞の頭文字が頻度に応じて書いてあるとします（例えばAは三個、Sは二〇個という具合に）。そこで二つの袋から同時にそれぞれ一個ずつ玉を取り出し、その文字を記録し、元に戻す（復元抽出）試行を繰り返せば、「取り出された二つの玉の文字の一致は、二つの異なる言語で書かれた同じ数詞の頭文字の一致に擬せられる」というのが、そのモデルです。このモデルをもとに二項分布に基づいて、十個の数詞の一致する確率を順次計算するわけですが、そのプロセスは安本によって丁寧に紹介されています[3]。百四十個程度なら頭文字の分類を手作業で行ってもたいしたことはありませんが、本論末尾の資料（03、04）に、AWKという簡易言語（この言語は表計算にもテキスト処理にも使える、軽やかなフリーソフトです）を使って単語の頭文字を集計するスクリプトを示しておきました。資料（02）はその結果であり、資料（05）は一致の確率を求める式の例、資料（06）はこの確率に基づいて、十個の数詞のうちのX個以上の数詞の頭文字が一致する確率を計算する式の例です。

試算したところ、三個以上一致する確率は五%、四個以上一致する確率は〇・七%となりました。つまり、二個が偶然一致することは案外起こりうるのですが（確率二一%強）、三個以上が偶然一致するのは百回に五回程度、四個以上となると百回に一回もないということになります。逆にいえば、四個、あるいは三個以上の一致は偶然とは考えられず、それらの言語間に強い関連性がある、ということになります。先にも申しましたように、数詞は借用の可能性のある文化語の一面がありますから、これだけでたとえば印欧語の系統や相互関連性を言うことはできませんが、日本語やハンガリア語がこれに属さないということも含めて、すでに知られている言語史的

事実がこの僅かな資料でも検証され確認されうるといえるでしょう。

印欧語に関する計量言語学的なアプローチの手法として安本によって詳しく説明されているものに、スエーデンのエレゴルド（ELLEGÅRD, Alvar）の提唱する相似性係数（coefficient of similarity）があり、その有意性検定や、言語年代学への応用可能性もたいへん興味深いものです。たとえばロス＝ブリトンによる印欧語の語根に関する調査データに基づく相似性係数において[4]、一番数値の低いのはアルメニア語とアルバニア語のそれですが、ブリトンの調べた一八六〇個の印欧語の語根のうち、前者は四四二個、後者は二九〇個の共通語根が残存しているとされ、これは他の印欧語の場合にくらべ、かなり少ないようです。それでもアルメニア語は他の言語との相似性係数がそう極端に低いわけではないので、印欧語への帰属が疑われるわけではありません。しかしアルバニア語の場合は他の言語との相似性係数も低いので、もしこの両者の相似性係数が偶然に基づくとすれば、印欧語への帰属に多少とも疑いが生じます。そこで試しに相似性係数の有意性を検定してみることにしました。

この検定における帰無仮説は、「二つの言語間にみられる相似性係数は偶然に基づく（母相似係数は〇）」となります。ここで残存する共通語根の有無と両言語を縦横の分類とする二・二分割表により検定統計量を計算し、相似性係数が検定統計量（絶対値）より大きければ、仮説を棄却することになります。つまり二つの言語間の相似性（一種の相関性といえるでしょう）は偶然のものとはいえない、換言すれば両者は無関係とはいえない、ということになります。実際計算してみると、相似性係数三六に対し、検定統計量は二七・七となりましたので、アルメニア語とアルバニア語はやはり無縁とはいえないわけです。ついでに他の相似性係数データについても同様の検定をしてみましたが、いずれも帰無仮説は棄却され、印欧語族であるかどうかを改めて疑わせるものはありませんでした。（資料07）

以上述べたことは、結論を求めるというより、方法の正しさを確かめる意味が大きいのです。研究の蓄積が大きい印欧語はその点で格好の相手だといえましょう。主観性が働き共通の土俵を欠く飛躍した議論が行われるよ

うな場合に、誰が行っても同じ結果が出る、つまり追試可能性のある方法が有効であることはいうまでもありません。

二 言語年代学との接点

次に紹介する服部四郎氏の、日本語の系統に関する「語彙統計学的〝水深測量〟」は、思いつき的なものも含めて熱い議論が闘わされがちな日本語の起源論争に関し、いま述べた意味において重要な問題提起をし、少し古いとはいえ、参照するに足る基礎データを提供したものです。服部は、言語年代学の創始者スワデシュ（SWADESH, Morris: 1909-1967）が提唱した基礎語彙百項目について、日本語と「四周」十八の言語の比較を試みる「水深測量」を行い、こう述べています。「どの言語も例外なしに、日本語との間に、五乃至一〇項目程度の類似を示すと言ってよい。これは、これらの諸言語が、互いに遠い親族関係を有するためであるか、或いは借用関係（最も広義で、底層或いは上層からのそれを含む）に起因するものか、或いは偶然の類似であるか、……」(5)

数詞の例でも見ましたように、親族関係はなくても一割程度の偶然の一致が生じる可能性を考慮すれば、服部の慎重な発言は一般に今でも貴重なものといえますが、我々はそこで提示されている諸言語の、日本語と形の類似した基礎語彙の中から、更に語頭音が一致または対応すると思われるものをピックアップして、試しに言語年代学の定式に当てはめてみましょう。朝鮮語をはじめとする十八の言語と日本語の間に、厳密な「音韻法則」は成り立たない、とされますから、これは決して親族関係を前提に「分離した年代」をはかるものではありませんが、仮に遠い関係があったとしても、少なくとも算出された年代より以後の分離はありえないだろう、という意味での目安になると思われます。

ところで、ここにいう言語年代学の定式とは、（たとえば甲乙）二つの言語に共通する基礎語彙の数を基礎語彙総数で割り、それを甲乙両言語固有の「共通残存語率r」とし、広範囲な多言語的調査から求めた千年あたりの一般的残存率を「基礎語彙残存率s」とする時、rの対数をsの対数で割り、更に定数cで割るものです（資料08）。rは実際に調べることができますが、問題になるのはsとcの値で、スワデシュはcを二（二つの言語

が分離後ほとんど接触しないという前提）とするのに対し、服部は色々試した結果一・四を提唱しています。またrについては、英独仏西など十三言語の検証結果から平均〇・八〇五、最大〇・八五四などの数値があげられています。調査すべき基礎語彙の数についても、スワデシュ自身が百と二一五の項目をあげるなど揺れが見られ、また身体各部や天然自然の名称や基本的動作をあらわす言葉など、どの民族にも当てはまる語彙項目の選択についても多少の異論はあるようです。

ここではしかし、スワデシュ、服部両氏に敬意を表し、sについては〇・八五四と〇・八〇五の二種、cについても一・四と二の二種をとることにしましょう。つまり四つの組み合わせができるわけですが、このうち〇・八五四と一・四の組み合わせが最大値、〇・八〇五と二の組み合わせが最小値を与える事にすぐ気づきます。朝鮮語に関しては服部自身の計算例があり、京都方言と京城方言を比較すると対応する九三の基礎語彙のうち十個から十八個が類似するので、cを一・四としsを〇・八五四とすると一〇・〇六（千年）から七・四一（千年）、「念のため一・四の代わりに二とし、rを八〇・五%として計算すると四・九〇千年乃至三・六二千年となる」としています。検算してみると少数二桁目が多少違うようですが、勿論このような数値は「厳密に受け取ることは、あらゆる意味で不可能」ですから、多少の誤差は無視していいでしょう。「それにもかかわらず、それが、両者が親族関係を有するとしても非常に古く分裂したものであること（四千年まえ以後に分裂したものではあり得ないこと）、を物語るものと解釈することは、決して不適当ではなかろう」という氏の考え方に、筆者も賛成したいと思います。上に述べた四つの組み合わせのうち、最大値と最小値を与える式に服部データに基づく数値（頭文字の一致または対応と考えられるもの）を代入した結果は資料（09）に示しておきました。

詳しく触れることはできませんが、この問題に関しても安本氏は、相似性係数との関連をはじめ色々有意義な検討を行い、日本語の基礎語彙を確定し起源を探る試みをしています。日本語の系譜に関しては大野晋氏のタミル語同系説における基礎語彙の検証も含めていずれ改めて取り組んでみたいと思っています。

三 コーパス言語学と計量言語学

話題を少し変えて私のもう一つの専門であるゲルマニスティク（広義のドイツ文学研究）と計量言語学の接点、ないし応用可能性に移りましょう。データと統計的分析手法の関係でいうと、ニューズ記事と重回帰分析、ゲーテに関するデータの主成分分析と分散分析などです。これらはいずれもテキスト・データベースを用いて統計的分析を行うという意味ではコーパス言語学（Corpus Linguistics）と最近呼ばれるようになったものと同じ方向にあるといえるでしょう。本来コーパスとは文献集成を意味し、大規模で体系的なものをいうことが多いので、個人的に随意作成したテキスト・データベースを安易にこれと同一視することはできませんが。

昨年の夏の始め頃、ドイツの超特急が大惨事を起こしたことを御記憶でしょうか。私はそれを契機に一ヶ月ほどインターネットで毎日ニューズを受信し、プレイン・テキストに変換して文体分析資料としてみました。入手したのはウルム大学で発信している「ジャーマン・ニューズ（GN）」という文字情報主体のものですが、これを読んでいるうちにふと思ったのは、報道記事には一定のスタイルがあるゆえ、分析次第では自動翻訳や機械的通訳の基礎資料になるということです。というのもドイツ語は英語などと比べて倒置文が非常に多く、述語（動詞群）がひとかたまりにならず文成分として二番目の位置と文末に分かれて出現します。いわゆる「枠構造」や否定詞の位置の関係で、文意の予測がしにくいところがあるのです。文型上のこの特徴は、ノーマン・コンクェストのような社会変動を経験しなかったドイツ語世界が、「曲折」要素を強く残していることと関係すると思われますが、いずれにしても文頭（Vorfeld）に必ずしも主語が立たないとすれば、他にどんな要素ないし語句がその位置を占めるか、またそれによって文の長さや構成に一定の傾向がみられるか、などということを調べることは無駄ではないでしょう。

報道記事の六要素はふつう「誰が、いつ、どこで、何を、なぜ、どのように」したか、だと言いますから、主語は文頭に立ちやすいと思われますが、実際GNで調べてみると、平均して五五%前後でした。残りを占める要素で多いのは、前置詞句と副詞であり、だいぶ差がつい

て目的語や副文、そして接続詞となります。文の長さは平均一五語ほどで、これはFAZのような特に長文の多いことで知られる新聞を別にすれば、報道記事としてまず平均的なものといえましょう。こういうことはいわば記述統計学に属する分析ですが、私はここで文章の長さ（Y）と、固有名詞を含む名詞の頻度（E）、一度しか現れない語の頻度（H）、それに正置文か否か（S）、という三つの要素の間に一定の（線的）関係がみられないか、推測統計学的な分析を試みてみました。つまり、文章長を目的変数（従属変数）とし、他の要素を説明変数（独立変数）とする重回帰分析を試みたのです。

その為まず、一ヶ月分のテキスト（タイトル等を除く）を三、四日毎にまとめた八つのファイルを作りました。一日毎のファイルでは文数が少ない上、大事件などがあると記事に偏りが生ずるので、それを避ける必要があるからです。独立変数のEとHは千語当たりの頻度に換算し、Sは百分率にしたうえで、Sのみを説明変数とする一変数モデル、説明変数を二つずつ組み合わせた二変数モデル、および三変数モデルについて、それぞれ回帰式を算出しました。これらの式と元のデータとから残差平方和が得られるので、情報量基準（AIC）、決定係数、分散比を計算して比較したところ、やや意外にも三変数モデルより、EとSを説明変数とする二変数モデルの方が目的変数をよりよく説明することが判明したのです。というより、このモデルの場合だけ、有意水準五％で回帰は有意となり、EとSがYの予測に役立つことになったわけです。この問題は今後なお色々な角度から検討する必要がありそうですが、とりあえず資料（10）に結果を示しておきました。最近IBM社が発表した「ヴィア・ヴォイス」という音声認識ソフトは、音響的特徴をとらえるのに隠れマルコフモデル（HMM）を用いる一方、言語的特徴をとらえるのには、ある単語の次には何が来る確率が高いか、Nグラム・モデル（N-gram）に拠る統計的分析を重ねた成果を利用しているようですが、(6) ドイツ語における文頭の前置詞句などは三グラム程度の分析をさまざまなジャンルについて行えば、面白い結果が生ずると思われます。

次に紹介するのは、ゲーテ及びその同時代人ニコライの小説を資料にした文体分析で、統計手法としてはカイ

二乗検定や主成分分析を使ってみました。

二五歳のゲーテ（Johann Wolfgang Goethe）が一七七四年に発表した書簡体小説『若きウェルテルの悩み』は、身分制度と啓蒙悟性主義で感情のはけ口を失っていた当時の若者の心をとらえ、ナポレオンが何度も読み直したと伝えられるほどヨーロッパ規模のセンセーショナルな成功をおさめました。その一方、主人公を真似た服装や自殺がはやるという社会現象も生じ、一部の識者は眉をひそめておりましたが、これをうけて悟性派作家兼出版者の代表をもって任ずるベルリンのニコライ（Friedrich Nicolai）は『若きウェルテルの喜び』というパロディ小説を書いて迷える若者を善導しようと試みました。彼はまたほぼ同時期に『修士ゼバルドゥス・ノタンカーの生活と意見』という長編小説も発表しています。いずれも善意こそ溢れてはいるが毒にも薬にもならない通俗的啓蒙小説といっていいでしょう。ゲーテにはまた三十年以上後に書かれた『親和力』という小説があり、これも理屈では律しきれない心の動きをテーマとしたものですが、『ウェルテル』が典型的なシュトゥルム・ウント・ドラング（疾風怒濤、嵐と襲撃）期の作品であるのに対し、『親和力』は古典期を経てロマン主義にかかる時期の作品です。

ゲーテとニコライの『ウェルテル』については単語の長さを基準にした分割表によるカイ二乗検定を試みたことがかつてあります。[7] その頃は活字の読みにくい初版本の復刻をもとにテキストを手作業で入力したのですが、最近はインターネットを通じてテキストを入手し分析用に加工（SEDなどを使ってプレーンテキストにする）することができるので、少し規模を拡大して再度文体分析を試みてみました。目標は、ゲーテの青年期と壮年期の間に文体の変化がみられるか、ゲーテとニコライという別人の間に相違がみられるか、統計的に確かめることにあります。二部からなる『ウェルテル』については第一部は初版を、第二部は再版からテキスト・データをとりましたので、都合五個のテキスト群を用意したわけです。これをそれぞれW1（ウェルテル一）、W2（ウェルテル二）、WV（親和力）、NW（ニコライ・ウェルテル）、SB（ゼバルドゥス）とし、単語の長さ（文字数による。固有名詞は除去）を基準にしてクラス（階級）分けした分割表によりカイ二乗検定（一様性仮説検定）

を行ってみると、五%、一%、〇・一%のいずれの有意水準においても帰無仮説保留（棄却されない）となるのはW1とW2の組み合わせのみで、WVとSBの組み合わせは〇・一%水準では仮説が保留されました。次に単語の長さの平均について（母分散不明と仮定し）アスピンの検定を行ったところ、ここではW1とW2の組み合わせのみならず、NBとW1、W2の組み合わせや、SBとWVの組み合わせまで、平均が等しい、とする仮説が棄却されない結果となりました。この検定はこういう場合あまり検出力が高くないようです。

次の作業としてテキストに含まれる全ての単語（レンマ・見出し語ではなく出現語形、固有名詞や数字などを省く）の出現頻度を調べ、その中から頻度の高い不変化詞（前置詞、接続詞、副詞など）二五を抜き出して、千語当たりの頻度に換算しました。私の場合こういう作業はすべてAWKを道具として活用していますが、選んだ単語は次の二五語です。

und nicht zu so in mit was wie nur
auf doch von an wenn da dass man als
auch nun noch nach wohl denn aber

これらはコンテクストとあまり関わりなく無意識に使われる事が多いものですが、この二五語を変数とみなし、相関係数を算出し、それをもとに主成分分析を行った結果、四つの主成分に縮約されました。資料（12）はその第一主成分を横軸に、第二主成分を縦軸にとった付置図で、『ウェルテル』の一部と二部がY軸をはさんで近接し、『親和力』がやや離れた第三象限、『ゼバルドゥス』は更に離れ、ニコライの『ウェルテル』は他とまったく離れて第四象限に位置しています。五作品の関係をほぼ正確に反映しているといってよいでしょう。また、この第一主成分では前置詞はほとんど例外なくマイナス値をとっており、傾向としては前置詞グループと心態詞中心グループの要素が対立していると分析できます。なお、この二五語の頻度に基づき、ノンパラメトリック検定の一つであるクラスカル＝ワリス検定を試みましたが、有意差はみられませんでした。これまでの経験からしても一般に文体統計では順位度にもとづく検定はあまり効果的ではないように思われます。

最後に『ゲーテ・シラー往復書簡』を資料とした実験

的試みの結果をお話しましょう。十歳の年齢差のある二人の出会いはドイツ文学の黄金期を招来し、ワイマールに在って公国閣僚として繁忙をきわめるゲーテと、大学町イェーナで病身をおして広義の文学活動を続けるシラーとは、一七九四年からシラーの没する一八〇五年まで実に頻繁な実りある文通をしています。今なら車で一時間もかからない距離ですが、冬場など病身のシラーは次第に外出もままならないようになるので、その分手紙に託して意見の交換をしていたのです。一七九四年と九五年の往復書簡の分析は以前に一度試みたことがあり、たとえば文章の長さなど、最初かなり違いのあった両者が次第に似通ってくる傾向などがみられました。[8] 今回は「バラードの年」と呼ばれる一七九七年について予備調査をするうち、特に年度後半に奇妙な現象がみられたので、少し集中的に検討してみたものです。

というのは八月頃のゲーテの書簡文が異常に長いのです。そこで六月十日から年末までの二人の書簡を、ゲーテは五個（G1、G2、G3、G4、G5）、シラーは四個（S1、S2、S3、S4）のテキスト・ファイルに格納し、さしあたり文章長（語数）の分散分析とカイ二乗一様性検定、および単語長のカイ二乗一様性検定を試みました。文章長について単純に平均値を比べて異常に長いのはG2で、八月九日から八月二四日までの期間ですが、ゲーテの他のグループは二五語前後であるのに、この部分だけは三二語弱になっています。まずゲーテについてのみ、一因子分散分析（一元配置分散分析）でのF検定を行ったところ、分散比は五・三〇となり（自由度四及び六九四）、平均値の一様性仮説が棄却され、この結果は五%トリム・データを使っても変わりませんでした（分散比六・三三、自由度は六二四）。一方、シラーについては最初S1が不備であったためS2、S3、S4の三群で分散分析を行ったのですが、分散比は〇・八〇七（自由度は二及び四五五）となり、予測されるように、平均値の一様性に関する帰無仮説は棄却されません。そこで今度はG2を除いたゲーテ四群について同様の検定を行ってみると、分散比は〇・〇八八となり、平均値が等しいとする帰無仮説は棄却されないのです。G2が例外的であるのは明らかですが、この結果は文章長を五語刻みの階級にわけ、カイ二乗検定を行っても変わりませんでした。すなわちG2と他のファイルでは、

G4との組み合わせ以外はすべて一様性の帰無仮説が棄却されます。他のゲーテのファイルとシラーのファイルを組み合わせても棄却されないように、文章長の場合カイ二乗検定をしても仮説保留となることが一般に多いので、G2は際立っています。そこで今度はG1からG5まですべて合併したものと、G2のみ除いた合併ファイル、及びシラーの合併ファイルをつくり、単語長による分割表に基づくカイ二乗一様性検定を行ったところ、ゲーテの全合併ファイルとG2を除いた四ファイルは、シラーと組み合わせてみると当然ながら仮説棄却、すなわち両者に一様性は認められない、ということになりますが、ゲーテ四ファイルとG2は帰無仮説が棄却されません。つまり単語長に関してはゲーテの書簡に一様性があると考えられますから、矢張りG2、八月の手紙の文章は（語数で計る限り）異常に長い文が多いということになります。

これはどういうことなのでしょうか。この年のゲーテの行動をざっとみると、五月から六月にかけてイェーナに滞在し、シラーとの間で「バラード対話」とでも呼ぶべき文学論が交わされています。七月九日には何を思ったのか、一七七二年から一七九二年までに受け取った手紙を一切焼却（日記に"Briefe verbrannt"と簡潔な記録あり）。七月十一日から十八日まではシラーがワイマールのゲーテ家に宿泊し、ゲーテの内縁の妻クリスティアーネとも会ったと推定されます。七月二四日には遺言書を書き換え、クリスティアーネと二人の子アウグストの権利をはっきり認めたうえで、七月三〇日にワイマールを発って故郷フランクフルトへ向かい、クリスティアーネ達「家族」も後からフランクフルトに着いてゲーテの母にまみえるということになるのです。つまり件の書簡はこのフランクフルト滞在期に書かれたもので、ワイマールの閣僚でありながら庶民の娘を内縁の妻とし、貴族社会および教会から公然とは言わぬまでも非難の視線を浴びていたゲーテが、はっきりと自分の意思を表明した時期にあたると言えましょう。ゲーテはこの後、単独でスイスに旅し、ヴィルヘルム・テルゆかりの地やゴッタルト峠に登ったりし、それをこと細かく盟友シラーに書き送っています。病弱で遠く旅することのできないシラーのいわば目となり、シラーはゲーテを通じて代理体験をしている、といえるかもしれません。シラー最後の

戯曲の傑作『ヴィルヘルム・テル』はこの代理体験ぬきに考えられませんが、同時にゲーテもこの頃シラーの強い慫慂に応じて、一度筆を絶っていた畢生の大作『ファウスト』にあらためて取り組むことになったのでした。

文体の分析は実用性という面からいえば、不明文書の作者解明や、場合によっては犯人究明に応用することも可能です。ここにはほんの一端を紹介しました。

（1） 安本美典『言語の科学』（一九九五、朝倉書店）。

（2） “arbitrariness”。ソシュール（Ferdinand de Saussure）の唱えた言語観で、言語記号とそれが表す対象（意味内容）との関係は、（擬音語などを別にすれば）物理的必然性はなく、恣意的（arbitraire）とする。たとえば、「いぬ、犬、dog、Hund」は同じものをあらわす。

（3） 安本、二一頁～二八頁。

（4） 安本、七八頁。

（5） 服部四郎『日本語の系統』（一九五九、岩波書店）二二六頁。

（6） 西村雅史・伊東伸泰「音声認識の最新技術」（一九九八『Bit』八頁～十三頁。

（7） 新井皓士『近世ドイツ言語文化史論』（一九九四年、近代文藝社）一〇二～一一五頁。

（8） 新井「文体統計論的にみた『ゲーテ・シラー往復書簡集』の特性」（一九九六、『ゲーテ年鑑』第三八巻）一五五頁～一七四頁。

（一橋大学言語社会研究科教授）

(01) polya14.dat

one	two	three	four	five	six	seven	eight	nine	ten
en	tva	tre	fyra	fem	sex	sju	atta	nio	tio
en	to	tre	fire	fem	seks	syv	otte	ni	ti
een	twee	drie	vier	vijf	zes	zeven	acht	negen	tien
eins	zwei	drei	vier	funf	sechs	sieben	acht	neun	zehn
un	deux	trois	quatre	cinq	six	sept	huit	neuf	dix
uno	dos	tres	cuatro	cinco	seis	siete	och	nueve	diez
uno	due	tre	quattro	cinque	sei	sette	otto	nove	dieci
jeden	dwa	trzy	cztery	piec	szesc	siedem	osiem	dziewiec	dziesiec
egy	ketto	harom	negy	ot	hat	het	nyolc	kilenc	tiz
unus	duo	tres	quattior	quinque	sex	septem	octo	novem	decem
odin	dva	tri	chetire	pyat	shest	semi	vosem	devyat	desryat
ena	duo	tria	tessera	pente	exi	epta	okto	ennea	deka
hi	hu	mi	yo	itutsu	mutsu	nanats	yatsu	kokonots	toh

(02)

a	c	d	e	f	h	i	j	k	m	n
3	6	18	10	7	6	1	1	3	2	12

o	p	q	s	t	u	v	y	z
9	3	4	20	21	4	4	2	4

(03)　awk -f fchar.awk c:¥polya14.dat

(04)　fchar.awk

```
# first character frequency of words
{
      for(i=1; i<=NF; i++){
            fchar = substr($i, 1, 1)
            list[fchar] ++
      }
}
END{
      for(fchar in list)
            if(fchar !="¥t") printf("%2s %3d¥n", fchar, list[fchar])
            else printf("¥¥t %4d¥n",list[fchar])
}
```

(05)

$$p=\frac{3^2+6^2+18^2+10^2+7^2+6^2+1^2\cdots+9^2+3^2+4^2+20^2+21^2+4^2+4^2+2^2+4^2}{140\times140}$$

(06)

$$\pi=\sum_{i=x}^{10}({}_{10}C_i)\,(0.08735)^i\times(0.91265)^{10-i}$$

(07) 相似性係数(coefficient of similarity) .

1）言語ＡＢについて、たとえば７４項目の言語学的特徴の有無を調べ２×２分割表をつくる。

		言語A	
		有り	無し
言語B	有り	a	b
	無し	c	(d)

双方にその特徴がない場合の(d)については、さしあたり考慮しない。

2）相似性係数（実際には $100r_n$ を用いる）：$r_n = \dfrac{a}{\sqrt{(a+b)\times(a+c)}}$

3）相似性係数の有意性検定

H_0： ２つの言語間に認められる相似性係数r_nは、偶然にもとづく。

$100r_n > |y|$ なら、仮説棄却。但し

$$y = \frac{100\,\varepsilon}{\sqrt{(a+b)(a+c)}} \qquad \varepsilon = a\times\left[\frac{a+b}{a+b+c+d} - 1\right]$$

4）Armenia/Albania の $100r_n$ は３６ (安本 p.78)

a = 130; b = 312; c = 160; d = 1258 ゆえ、上式より

$|y| = 28.682$ $\quad\therefore\quad 100r_n > |y|$

仮説棄却。アルメニア語とアルバニア語の相似性係数は偶然にもとづくということはできない。

(08) 分離年代推定式

$$t = \frac{\log r}{c \log s}$$

但し、 r ：共通残存語率
s ：基礎語彙残存率（千年当り）
0.805 or 0.854(max)
c ：定数(1.4 or 2)

(09)服部データに基づく試算。（頭文字一致は最大限広義にとらえている *印)

言語名	対応数	類似数	頭文字一致数	$\log r/1.4*\log 0.854$	$\log r/2*\log 0.805$
朝鮮	93	20	12*	9.27	4.72
満州	93	15	5*	13.23	6.74
蒙古	100?	16	7*	12.04	6.13
タタール	100	14	7*	12.04	6.13
アイヌ	100	22	12*	9.60	4.89
ギリヤーク	100	9	6*	12.73	6.49
シナ（上古）	100	14	7	12.04	6.13
チベット	100	7	3	15.87	8.08
タイ	100	3	3*	15.87	8.08
ヴェトナム	100	7	6	12.73	6.49
カンボジャ	100	5	5*	13.56	6.91
マライ	100	5	3	15.87	8.08
タガログ	100	4	2	17.70	9.02
高砂族	86	10	7	11.35	5.78
モトゥ	86	7	3	15.19	7.74
サモア	93	6	4*	14.24	7.25
カロリン	98	4	4*	14.48	7.34
ボグー	96	12	9*	10.71	5.46

(10)　"German News"　1998.06.03 - 07.04

ファイル	文数	語数	文章長平均	名詞等	頻度1の語	正置文
1(03-05)	160	2324	14.64	300.77	277.97	55.6
2(06-09)	162	2260	14.03	333.19	358.41	60.5
3(10-12)	140	2131	15.17	331.30	366.30	56.4
4(13-16)	159	2326	14.54	331.47	364.57	56.6
5(17-19)	139	2168	15.60	331.64	394.37	50.4
6(20-23)	176	2494	14.23	255.41	383.72	54.5
7(24-28)	162	2484	15.02	319.24	358.70	58.0
8(02-04)	147	2275	15.27	318.68	369.67	52.4

	変数	回帰式
a)	S	$\hat{y} = 21.3636 - 0.117931$
b)	E, S	$\hat{y} = 18.8137 + 0.01068X_1 - 0.13265X_2$
c)	H, S	$\hat{y} = 20.9642 + 0.00073X_1 - 0.11545X_2$
d)	E, H	$\hat{y} = 11.1273 + 0.00774X_1 + 0.00347X_2$
e)	E, H, S	$\hat{y} = 19.0219 + 0.01075X_1 - 0.00041X_2 - 0.13415X_3$

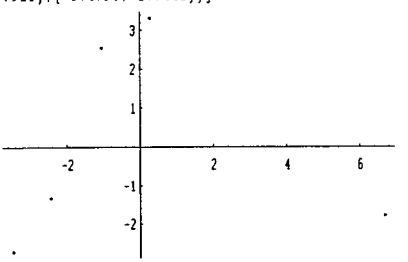
		AIC	決定係数	分散比
a)	S	-11.990	0.473	2.358
b)	E, S	-15.739	0.739	7.079 *
c)	H, S	-10.02	0.475	2.259
d)	E, H	- 6.726	0.207	1.630
e)	E, H, S	-13.71	0.745	3.527

(11) Nicolai：ウェルテルの喜び(NW)、マギステルＳＢ(SB)、
Goethe：ウェルテルの悩み（初版）(W1)、同（再版）(W2)、親和力(WV)

	NW	SB	W1	W2	WV
単語長平均	4.924	5.366	4.882	4.844	5.246
標本分散	15.640	28.922	28.216	25.829	28.776
不偏分散	15.650	28.930	28.222	25.836	28.784
総数	3775	12881	21662	12545	12732

(12)不変化詞２５の頻度に基づく主成分分析付置図

```
ListPlot[{{0 2787,3.3172},{-1.0412,2.5495},{-2.4466,-1.3316}
{6 6885,-1.7915},{-3.4796,-2.7441}}]
```



	（X軸、	y軸）
	第１主成分	第２主成分
NW	6.6855	-1.7915
SB	-3.1796	-2.7441
W1	0.2787	3.3172
W2	-1.0112	2.5495
WV	-2.1166	-1.3316

固有値	(1)	12.7308	12.7308
	(2)	6.0812	18.812
	(3)	3.4354	22.2474
	(4)	2.2916	24.5390
	etc		累積

(13)ゲーテ・シラー往復書簡　１９９７年後期

ファイル名	期間（日付）	行数	文章長平均	
G1	'97. 06. 10-07. 29	116	25. 39	
G2	08. 09-08. 24	147	31. 86	
G3	08. 30-09. 26	174	25. 11	
G4	10. 14-11. 29	142	25. 27	
G5	12. 02-12.	120	24. 41	総平均　26. 46
S1	'97. 06. 18-08. 17	215	25. 05	
S2	08. 30-09. 15	121	24. 82	
S3	09. 22-10. 30	129	22. 99	
S4	11. 22-12. 29	208	22. 87	総平均　23. 42

分散分析1 (G1, G2, G3, G4, G5)

変動要因	平方和	自由度	平均平方	分散比
標本間	5499. 0	4	1374. 75	F=5. 30 **
標本内	181000. 3	694	259. 31	
全体	186499. 3	698		

分散分析2 (G1, G3, G4, G5)

変動要因	平方和	自由度	平均平方	分散比
標本間	58. 4	3	19. 47	F=0. 088
標本内	120661. 5	548	220. 19	
全体	120719. 9	551		

分散分析３ (S2, S3, S4)

変動要因	平方和	自由度	平均平方	分散比
標本間	324. 1	2	162. 06	F=0. 807
標本内	91373. 2	455	200. 82	
全体	91697. 35	457		

文章長：カイ二乗検定

	G1	G2	G3	G4	G5	S1	S2	S3	S4
G1	*	10. 994	5. 268	3. 732	9. 784	10. 752	5. 971	4. 644	5. 644
G2		*	19. 560	12. 903	15. 961	31. 744	14. 930	23. 745	29. 567
G3			*	10. 816	9. 823	8. 818	10. 911	5. 988	9. 679
G4				*	11. 650	14. 069	6. 952	7. 750	8. 078
G5					*	4. 752	9. 753	6. 673	9. 685
S1						*	9. 601	4. 342	4. 871
S2							*	6. 968	10. 228
S3								*	3. 078
S4									

[X^2(8, 0. 05) = 15. 5073]
[X^2(8, 0. 01) = 20. 0902]
[X^2(8, 0. 001]= 26. 1245]

単語長：カイ二乗検定

(G1, G2, G3, G4, G5)	: (S1, S2, S3, S4)	X^2 = 47. 314 *, **, ***
(G1, G3, G4, G5)	: (S1, S2, S3, S4)	X^2 = 40. 537 *, **, ***
(G1, G3, G4, G5)	: (G2)	X^2 = 10. 796