サイエンストピックス
言語比較の数学的基礎 *

吉田知行 (数学専攻)

2005/05

目次：1. 数学は役に立つのか？2. 比較言語学．3. 日本語の起源問題．4. ポリアの検定法．5. オズワルトのシフト法．6. 群論の登場．7. 対称群を使うシフト法．8. 安本の二項検定法．9. 言語群同士の比較．10. 日本語の謎と不思議．11. ブートストラップ法．12. 系統樹．13. さらなる発展と課題．14. 数学の核．15. あとがき．参考書.

1　数学は役に立つのか？

「数学は役に立つのか？」とは，しばしば否定的な意味合いを込めて私たち数学者に投げかけられる言葉です.「二次方程式の解の公式はいらない」は有名ですが，私も「空間は3次元なので，3次元までのベクトル空間は学べばよい」とか「微分方程式の解の存在定理なんて役に立ったことはない」とか，はては「数式処理システムの使い方を知っていれば数学はいらない」と言われたことがあります．(自分にとって) 役に立たないものはいらないという考えは間違っていると思います.

数学には，スポーツや芸術，囲碁将棋のような文化の側面があるし，実際，和算はまさに日本の誇る文化だったと思います．数学は自然科学よりも人文科学の方に近いのと考える人も少なくないようですが，数学で使う方法はまさに自然科学の方法であり，またその普遍性と抽象性により現代では科学のための言語の役割を担っています.

功利的な面から見ても，数学は間違いなくいろいろな方面の役に立つし，科学技術に革命を引き起こしたこともまれではありません．相対性

*「数学の応用事例—比較言語学への応用」を改題

理論や量子論はあまりにも有名です．また，メンデルは数学と物理が得意で，メンデルの法則の発見に組合せ論が役に立ったそうです．集団遺伝学や分子進化の中立説で有名な木村資生 (もとお) も数学が得意で，相当高度の微分方程式や確率論，統計学を自分で学び，それを遺伝学に応用して革命的な理論を作り上げました (事故で亡くならなければノーベル賞の最有力候補だったでしょう)．近年では，符号や暗号の理論への代数幾何や整数論の応用，脳や人工知能へのカオス理論など非線形解析学の応用，ロボット制御への微分幾何の応用，電子顕微鏡・ＣＴスキャン・地中レーダーへの関数解析の応用，計算機への数学基礎論の応用，金融工学への確率微分方程式の応用，ヒトゲノム解読計画，フラーレンと準正多面体など，枚挙にいとまがありません．

ニュートンが力学の記述のために微分積分法を作ったように，多くの数学分野は，自然科学からの刺激を受けて発展してきました．数学者はそのような新しい数学が生まれると，それを使える形に整備し，数学内部でのいろいろな分野との関連を研究し，新しい数学的な結果や応用を見いだそうとします．例えば，常微分方程式の理論は，生まれ故郷のニュートン力学だけでなく，自然科学全般，はては軍事・商業 (ランチェスターの法則)，考古学の年代測定や贋作の鑑定のような分野にまで応用が広がっています．これも微分方程式論という抽象化された理論があったからこそ，分野を超えた応用が可能になったのです．

数学内部の，純粋に知的好奇心が発展のきっかけだった数学分野も多くあります．「二次方程式の解の公式はいらない」という有名な言葉がありました．実際に，5 次方程式の解の公式が見つかろうが見つかるまいが，私たちの生活に直接関係することはないでしょう．しかしこの問題を解決しようとする数学者の長い努力からガロアの理論が生まれ，この問題に決着 (５次方程式には解の公式がない) がつきました．時は流れて現代の情報社会．ＣＤにちょっとくらい汚れや傷が付いてもきれいな音楽が聴ける，LSI が宇宙や素材からの放射線の悪さに耐えて正しく動く，遠い惑星から鮮明な画像が送られてくる，こういった情報のやりとりにはすべて，誤り訂正符号という伝送方式が使われています．また，情報を守るために，大量の暗号が使われていることは周知のことです．しかし，現代社会を裏から支える暗号や誤り訂正符号に，ガロア理論が使われていることを知っている人はあまりいないでしょう．

ここではささやかながら，ちょっとめずらしい人文系の科学 (比較言語学) を巡る数学を紹介します．

2　比較言語学

比較言語学は，18 世紀末東洋学者ジョーンズ卿の研究に始まるといわれます．今から 2000 年以上前から伝わってきたインドのサンスクリット語 (梵語) の単語の中に，ギリシャ語やラテン語に似たものが数多く見出されたのです．例えば次のようなものです．

	二	三	十	父	母	兄弟
英語	two	three	ten	father	mother	brother
サンスクリット語	dvi	tri	dasa	pitar	matar	bhratar
ギリシア語	duo	treis	deka	pater	meter	phrater
ラテン語	duo	tres	decem	pater	mater	frater

表 1: 古典語の類似

その後多くの言語学者の努力によって，ヨーロッパ，イラン，インドのかなりの言語，さらにヒッタイト語やトカラ語までを含むインド・ヨーロッパ語族 (印欧語族ともいう) が確立しました．比較言語学では，言語の語族への分類，ふたつの言語が同系かどうか，言語がどのように分岐し変化するかが研究され，さらに祖語の復元までが試みられています．

ふたつの言語が同系かどうかを判定する基準として，文法，音韻，語彙の類似が考えられます．英語を学んですぐ気がつくのは，日本語と英語の語順の違いです．日本語は SOV 型 (ここで S は主語，O は目的語，V は述語) であるのに対し，現代英語や中国語は SVO 型です．ではインド・ヨーロッパ語族に属する言語がすべて SVO 型かというと，そんなことはまったくなくて，肝心のラテン語は SOV 型 (古代英語や古代ギリシャ語もどちらかというと SOV 型) ですし，ウェールズ語やアイルランド語などのケルト系言語は VSO 型です．現代ドイツ語はどちらかというと SVO 型 (関係代名詞の中は SOV 型) です．世界の言語の中には OSV 型や VOS 型の言語もあるそうです．

ちなみに，古代英語は，名詞には男性女性中性の区別があり，名詞や動詞の語形変化も激しいなどドイツ語に似た言語でした．英語は，科学の世界で世界共通語の役割を果たしつつありますが，世界の言語の中ではかなり変わった言語です．子音 (24 種類の音素) や母音 (短母音 7 種類，そのほかに長母音，二重母音，三重母音) がやたらとあります．また綴り字と発音はかなりずれているし，さらに膨大な数の語彙 (数十万？) があります (これは日本語も同じ)．英語は，元のインドヨーロッパ語族に比べると，ある意味で退化した言語といえます．名詞の男性・女性・中性の

区別が完全になくなり，格変化もほとんど消滅し，複数形は若干の例外を除けば単に '-s' か '-es' を語尾につけるだけで良くなりました．ドイツ語やロシア語のような動詞の屈折変化も乏しくなりました．英語は，ゲルマン系言語の上に北欧の言語やラテン語，フランス語などがかぶさってできました．ごった煮のような状況では，名詞の性や格変化，動詞の複雑な屈折変化のような無駄は切り捨てられてしまったのでしょう．かわりに前置詞と助動詞が発達し，語順 (SVO 型) に厳しい制限が付くようになりました．このように，文法も変化するので，文法の類似だけでは同系の証明になりません．

比較言語学で同系の証明をどうするかというと，それは「**音韻対応の法則**」(または「音法則」) の成立によってなされます．例えば，英語とドイツ語では，drink ↔ trinken, think ↔ denken, day ↔ tag から直ちに見て取れる d ↔ t や t ↔ d のような規則正しい対応があります．ヨーロッパのいくつかの言語の数詞のリストをあげておきます (ただし文字表記で補助記号は省略)．

	英語	スウェーデン語	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	ハンガリー語
1	one	en	en	een	ein	un	uno	uno	jeden	egy
2	two	tra	to	twee	zwei	deux	dos	due	dwa	ketto
3	three	tre	tre	drie	drei	trois	tres	tre	trzy	harom
4	four	frya	fire	vier	vier	quatre	cuatro	quattro	cztery	negy
5	five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot
6	six	sex	seks	zes	sechs	six	seis	sei	szesc	hat
7	seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het
8	eight	atta	otte	acht	acht	huit	ocho	otto	osiem	nyolc
9	nine	nio	ni	negen	neun	neuf	nueve	ove	dziewiec	kilenc
10	ten	tio	ti	tien	zehn	dix	diez	dieci	dziesisc	tiz

表 2: ヨーロッパの言語数詞 (文字表記)

一見して明らかなように，多くの言語で数詞はよく似ています．ハンガリー語をのぞくすべての言語で，6, 7 は s 音で始まり，3 は t または d で始まっています．ポーランド語を除くすべての言語で 1 は 母音で始まっています．また，フランス語よりドイツ語の方が英語に近い印象を受けます．ドイツ語と英語が分かれたのは今から 2000 年前，フランス語とドイツ語が分かれたのが 5000 年前といわれています．ただしハンガリー語だけは他の言語との類似性が感じられません．実際，他の言語はすべてインド・ヨーロッパ語族なのに，ハンガリー語だけはウラル語族です．

3　日本語の起源問題

インドヨーロッパ語族やマライポリネシア語族などで，比較言語学は大きな成功を納めてきました．これに対して，日本語の系統関係は全く不明で，日本語との同系関係が証明された言語は沖縄方言だけです．日本語は○○語と同系であるとか，古代日本語が△△語で読めるとかいう本が結構出ていますが，まずは眉唾物と考えて間違いありません．当然ながら，そのような説は，少なくともひとつをのぞいて全部 (おそらく全部が) 間違っているはずです．単語の範囲を無制限に広げ，個々の単語の意味も曖昧にするなら日本語は英語と同系であるとさえ言えます．例えば，「斬る」と「kill」,「坊や」と「boy」,「そう」と「so」,「負う」と「owe」,「道路」と「road」,「字を引く書なり」と「dictionary」のように意味と発音の似た単語はたくさん挙げることができます (清水義範『序文』).

このような語呂合わせを防ぐために，比較言語学では，言語の同系の判別に「音韻対応の法則」という厳しい条件を課しているのです．また，比較言語学的方法で復元した単語の祖形を比べて，日本語は××語に同系だ，という議論もよく見られます．専門家のやることなので軽々しく批判は出来ませんが，何か一貫性のない名人芸に陥っているように感じを受けることがあります．祖語復元のアルゴリズムがあるなら納得できるのですが.

日本語と他言語との同系関係の有無は，100 年も前から比較言語学の専門家が研究し続けて来ました．それでも定説といえる同系関係を見い出せなかったのなら，もはやこの地球上に日本語 (含沖縄方言) と同系の言語は存在しない可能性の方を考えるべきでしょう.

なら，日本語は全く孤立した言語なのかというとそんなことはありません．そもそも人類 (ホモサピエンス) の祖先がアフリカを出たのが 15 万年前，アジア人とヨーロッパ人が分かれたのが 5 万年前といわれています．日本語や新大陸の言語を含む東アジアの言語は，2 万年ほど前に分岐したと考える言語学者もいます．そうすると日本語も 2 万年前にさかのぼれば，東アジア人の共通言語にたどり着くはずです．しかし残念なことに，分岐の時期がいつであれ，音韻対応の法則を見いだす方法では (いや，おそらくいかなる方法でも)，これほど古く分岐した言語の同系を証明することは不可能でしょう．音韻対応の法則を見出すという方法で，5000 年以上前に分岐した言語の同系関係の証明や祖語の復元が可能だそうです．しかしそれも，比較出来る多くの言語が周囲にあり，それらの古い記録が残っているという好条件に恵まれてのことです．現代の英語とネパー

ル語だけを比べて，それらが同系であると証明するのは困難と思います．同じインドヨーロッパ語族に属しながら，ネパール語の語順などは英語より日本語に似ている印象を受けます．

日本語が他の言語から分かれたというなら，その年代として，弥生時代の初め (2400〜3000 年前?) と縄文時代の初め (12000 年以上前) が考えられます．その年代が弥生時代の開始時期で，日本語の元になった言語の子孫がどこかに残っていたら，言語学者が見つけていないはずはありません．反対に，日本語の祖語が 1 万年以上前に他の言語から分岐し，日本列島内で現代の日本語に変化していったというなら，もはや同系関係を証明するのは不可能でしょう．結局，日本語と同系な言語はもう残っていないか，残っていても分岐年代が古すぎるなどの理由で，比較言語学の方法では同系の証明ができない，と考えるのが妥当でしょう．特に中国語以外のアジアの言語に，古い文字資料の少ないことが致命的です．

やや絶望的な状況ですが，まだ探求の道はあります．日本列島周辺に分布している言語と比べると，日本語は，中国語以外のアジアの諸言語の中ではきわめて標準的で平凡な音韻や文法体系を持っています．特に朝鮮語やモンゴル語の文章は，単語を対応する日本語の単語に置き換えるだけでほとんどそのまま日本語の文章になるほどです．つまり日本語ときわめてよく似た文法体系を持つ北方系の言語があるのです．そのため，日本語も朝鮮語もアルタイ系 (トルコ語，モンゴル語など) というのが有力な説です．日本語と朝鮮語が同系だったとして，その分岐年代は今から 6700 年ほど前，という言語年代学の計算結果もあります．文法は北方的だが，音韻や語彙は，南方系だという言語学者も多いようです．伝達の経路がどうであれ，稲作は明らかに南方系の文化要素なので，稲作とともに南方系の言葉が伝わったと考えるのはきわめて自然でしょう．

また，クレオール語 (混成言語) 説や流入説もあります．日本列島にはいろいろな時代にいろいろな民族がいろいろな文化を持ち込み，そのたびに新しい言葉が流れ込んで来ました．やまとことばを核として出来た日本語には，千年以上の間に膨大な量の漢語が加わり，現代でもカタカナ語として大量のヨーロッパ系の言語が流れ込んでいます．日本語が他言語の単語を特別受け入れやすい言語なのかは知りませんが，有史以前からこのような状況だった考えれば，流入説の方にも説得力があります．いずれにせよ，定説のなさは，日本語の起源問題の難しさと，この問題が従来の比較言語学の方法の射程外にある可能性を示唆します．問題解決には新たな方法が必要なようです．

4　ポリアの検定法

高校時代，ハンガリーの数学者ポリアの著作は私の愛読書のひとつでした．特に組合せ論を専門のひとつにすることになったのは『帰納と類比—数学における発見はいかになされるか１』(丸善) の影響が大きかったと思います．その続きに次の本があります．

ポリア『発見的推論　そのパターン—数学における発見はいかになされるか２』丸善 (1959)

ポリアはこの本の中で，言語同士の近さを数学的に測定するする方法を提唱しています．例として，前に使ったヨーロッパの言語の数詞でやってみます．英語とスエーデン語の数詞の語頭の文字を比べてゆくと，1 と 8 で語頭の文字が異なります (1: one – en, 8 : eight – atta)．残りの 8 個の数詞では語頭の文字がすべて一致します．したがって，英語とスエーデン語の間の一致数は 8 となります．このような語頭の文字の一致数をすべての言語の組について行うことによって次の表が得られます．一致数が大きければふたつの言語は近く，小さければ遠いと言えるでしょう．

	英	ス	デ	オ	ド	フ	ス	イ	ポ	ハ
英		8	8	3	4	4	4	4	3	1
ス	8		9	5	6	4	4	4	3	2
デ	8	9		4	5	4	5	5	4	2
オ	3	5	4		5	1	1	1	0	2
ド	4	6	5	5		3	3	3	2	1
フ	4	4	4	1	3		8	9	5	0
ス	4	4	5	1	3	8		9	7	0
イ	4	4	5	1	3	9	9		6	0
ポ	3	3	4	0	2	5	7	6		0
ハ	1	2	2	2	1	0	0	0	0	
計	39	45	46	22	32	38	41	41	30	8

表 3: 数詞についてのヨーロッパの言語同士の語頭文字一致数

ふたつの言語の一致数がどのくらいなら，このふたつの言語は近い，あるいは遠いと言えるのでしょうか．ふたつの言語で，意味を同じくする単語同士の語頭音や語頭文字が一致する理由は大まかに二通り考えられます：

(A) 偶然一致した．

(B) 元は同じ単語だった (共通祖語から受け継いだ，あるいは借用).

ただし，よほどの好条件に恵まれていない限り，どれが一致の理由かを決定することはできません．例えば,「名前」は英語では ‘name ’，日本語では ‘namae ’ とよく似ています．地球を半周する距離の遠さから，この類似は偶然と判断するのが常識でしょう (そうでないという説もある). しかし，それを証明するには，他の可能性 (同系か借用) を排除しなければなりませんが，中間地域での古い時代の資料がない現状でそんなことは原理的に不可能です.

そこで，ふたつの言語の数詞で，語頭文字の一致数の大きさが偶然かどうかという問題を考えてみましょう．これは，個々の単語の類似が同系か借用かの議論をやめることを意味します．まず推測統計学の検定方式にしたがって次の仮説を立てます：

帰無仮説 H_0 : 数詞の語頭文字の一致数の大きさは偶然によるものである.

考えているふたつの言語で，対応する単語の対の語頭文字が等しい確率を p とします．このとき帰無仮説 H_0 のもとで，n 個の単語の組のうち，r 組で語頭文字が等しくなる確率は,

$$\binom{n}{r}p^r(1-p)^{n-r}, \quad \text{ここで } \binom{n}{r} = {}_nC_r = \frac{n!}{r!(n-r)!} \text{ は二項係数}$$

で与えられます．よって，x 組以上で語頭文字が等しい確率 (**上側確率**) は

$$P(x) = \sum_{r=x}^{n}\binom{n}{r}p^r(1-p)^{n-r}$$

となります．ふつうの統計学の検定の基準では $P(x) < 0.05$ なら，帰無仮説を棄却し，x 組以上の語頭文字が等しいのは偶然ではないと判断します．ここで，0.05 を**有意水準**といいます．もっと厳しい判断基準が必要なときは，有意水準として 0.01 などを用います.

上にあげた 10 の言語について，10 個ずつの数詞，合計 100 個の語頭文字があり，その中には，a, b, $\cdots$, z が現れる回数は順に，3, 0, 5, 11, 6, 7, 0, 4, 0, 1, 2, 0, 0, 10, 6, 1, 2, 0, 16, 16, 3, 3, 0, 0, 0, 4 となっています．そうすると，偶然ふたつの頭文字が一致する確率は

$$p = \frac{3^2 + 0^2 + 5^2 + 11^2 + \cdots + 4^2}{100^2} = 0.0948$$

で見積もることができます．この p の値を採用すると，右のような上側確率 $P(x)$ $(x = 0, 1, \cdots, 10)$ の表が得られます．例えば $P(3) > 0.05$ なので，3 個の一致数では偶然の範囲を超えていません．

x	$P(x)$
0	1.000000
1	.630644
2	.243823
3	.0615239
4	.0106122
5	$.0^2128135$
6	$.0^3108711$
7	$.0^5637072$
8	$.0^6246186$
9	$.0^8565645$
10	$.0^{10}586251$

一致数が 4 だと 5%水準，5 だと 0.5%水準，6 だと 0.05%水準で，それぞれ偶然ではないと結論されます．ここでは，一致数が 5 以上のとき，偶然でないと判断することにします．こうすれば，上の表により，偶然の一致なのに，間違って偶然でないと判断する可能性はほぼ 1000 回に 1 回です．これは，ほぼ 10 回連続してコインの表が出続けるほどめずらしい出来事です．

結局数詞の語頭文字に関しては，次のような結論が得られます．

(a) ハンガリー語の数詞と他の言語の数詞との語頭文字の一致は偶然によるとして矛盾はない (一致数 2 以下).
(b) 北方系の英語，スウェーデン語，デンマーク語は強いまとまりを見せる (一致数 8 以上)．このような一致数が偶然得られる確率はきわめて小さい (有意水準 1 千万分の 5 以下).
(c) 南方系のフランス語，イタリア語，スペイン語も強いまとまりを見せる (一致数 8 以上).
(d) ドイツ語は，北方系言語のグループに近い.
(e) ポーランド語は，南方系言語のグループに近い.
(f) オランダ語は，北方系言語のグループに近いが，南方系グループとはかなり離れ，全体的に孤立している.

従来の比較言語学によれば，英語・スウェーデン語・デンマーク語・ドイツ語・オランダ語はゲルマン系の言語であり，フランス語・イタリア語・スペイン語はラテン系の言語，ポーランド語はスラブ系の言語，ハンガリー語はウラル系の言語です.

一致数を数えるという方法は，音韻対応の法則を見出すという古典的な比較言語学の方法とは全く考え方が違います．2 番目以降の文字の情報を捨て去って，語頭の文字や音だけを比較するおおざっぱさもあります．その他諸々の理由で，専門の言語学者から批判されがちです．しかしそこでは個々の単語の比較でなく，単語のセットが客観的方法で比較されるのであり，わずかの単語の選定や比較のミス，個人の恣意的判断が結論を動かすことはありません．計算機が打ち出す結果は常に同じです．

5　オズワルトのシフト法

ポリアの検定法と同じく語頭音 (または語頭文字) の一致による検定法として，オズワルトのシフト法があります．ふたつの言語 LA と LB の語彙リスト (数詞や基礎２００語等) を用意し，それぞれの言語で意味の対応する単語には 1 から n までの番号を付けておきます．同じ番号の単語はふたつの言語で同じ意味を持たなければなりません．リストの単語を比べていって，語頭音の一致する組が何組あるかを数えます．この一致数 x_0 を**粗点** (gross point) といいます．問題は，粗点の大きさが偶然の一致数を超えているかどうかで，そのために偶然の一致数の分布を知る必要があります．そこでオズワルトはうまい方法を見つけました：

R.L.Oswald, The detection of remote linguistic relationships, *Computer Studies in the Humanities and Verval Behavior*, **III** (1970), 117–129.

すなわち各 $i = 1, \cdots, n-1$ について， LA の k 番目の単語と LB の $k+i \pmod n$ 番目の単語を $k = 1, \cdots, n$ まで比べ，一致数 x_i を求めます．LB の単語の順番をずらすと LA の単語の意味とあわないので，こうして得られた一致数 $x_1, \cdots, x_{n-1}$ (**背景点**という) は偶然による一致数のはずです．計算を簡単にするために x_0 も背景点に入れておきます．

さらに，これら n 個の背景点の平均 m と分散 s^2，偏差値 z を求めます：

$$
\begin{aligned}
m &= \frac{1}{n}(x_0 + x_1 + \cdots + x_{n-1}) \\
s^2 &= \frac{1}{n}\{(x_0 - m)^2 + (x_1 - m)^2 + \cdots + (x_{n-1} - m)^2\} \\
&= \frac{1}{n}(x_0^2 + x_1^2 + \cdots + {x_{n-1}}^2) - m^2 \\
z &= \frac{x_0 - m}{s}
\end{aligned}
$$

項目	意味	日本語	朝鮮語
1	all	mina	motĕn
2	ash	нанi	tʃĕi
3	bark	кана	kəptʃir
4	belly	нara	pɐi
⋮	⋮	⋮	⋮
99	woman	me	kyətʃip
100	yellow	kï	nurŭ
1	all	mina	(motĕn)

表 4: ひとつずつずらしながら比較する

そこで，例えば $z > 2.33$ なら，粗点 x_0 の値は偶然による一致数を (5 パーセント水準で) 越えていると判定します．一般に偏差値 z から正規分布の上側確率は次の式で与えられます．

$$Q_n(z) = \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-t^2/2} dt. \tag{1}$$

特に $Q_n(-\infty) = 1, Q_n(0) = 1/2$ です．その他の特別の値として，

z	1.654	2.326	2.576	3.090
$Q_n(z)$	0.05	0.01	0.05	0.001

これにより例えば $z > 3.090$ なら，そのような現象の起こる確率は 0.001 未満，すなわち 0.1 パーセント水準で有意と判定します．この判定法が使えるには，一致数の分布が正規分布に似た形 (中央が高くすそが低い) である必要があります．

奈良時代の日本語と，ハングルの発明された 1400 年代中頃李氏朝鮮時代の朝鮮語の基礎 200 語でヒストグラムを描いて見ると，次のような図が得られます．ただし似た音はまとめてあります．ごくおおざっぱな判

定法として，粗点が背景点の中の最大値になれば，両言語での語頭音の一致は 0.5 パーセント水準で偶然でないと判断できます．つまり粗点が偶然最高位になるのは 200 回に 1 回ということです．日本語と朝鮮語の場合，統計学的には粗点 $x_0 = 53$ は偶然では得られないと判定されます ($p = 0.000663$).

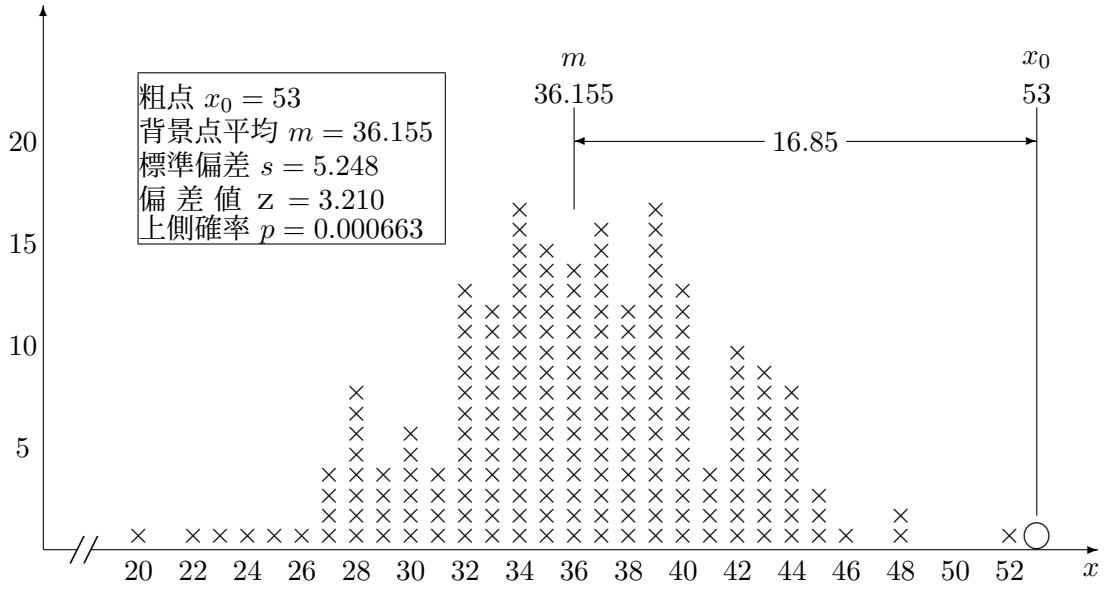

図 1: 「上古日本語」と「中期朝鮮語」のシフト検定 (基礎 200 語)

オズワルトのシフト法は，音韻対応が見いだせないほど離れたふたつの言語の関係でも検出できるきわめて強力な方法です．しかし，検出力が上がった反面，個々の単語同士の類似が借用か同系かの判別や音韻対応についての情報が犠牲になったのはやむを得ません．さらに問題点もいくつかあることが分かってきました．

(a) 基礎 200 語でふたつの言語を比べるには，全部で $200 \times 200 = 40000$ 回の単語の比較が必要となり，人手では困難である．
(b) 背景点の分布が本当に正規分布で近似できる保証がない．実際に，語彙をアルファベット順に並べた場合は，印欧語同士の比較でおかしな分布になる．
(c) 単語の順番によって，分布の形が変わる．

このうち手間の問題 (a) はコンピュータが解決してくれます．しかしそれを含め，すべての問題点は群論という代数学を使って解決します．

6　群論の登場

対称群については，大学1年の線形代数で学びますが，簡単に定義を述べておきます．n 個の文字 $1, 2, \cdots, n$ の**置換** π とは，集合 $[n] = \{1, 2, \cdots, n\}$ から $[n]$ 自身への写像で，$\pi(1)\,\pi(2)\cdots\pi(n)$ が $1, 2, \cdots, n$ の順列 (並べ替え) になっているものをいいます．このような置換 π を

$$\begin{pmatrix} 1 & 2 & \cdots & n \\ \pi(1) & \pi(2) & \cdots & \pi(n) \end{pmatrix}$$

と表します．また $\pi(a_1) = a_2, \pi(a_2) = a_3, \cdots, \pi(a_{r-1}) = a_r, \pi(a_r) = a_1$ であるような置換 π (ただし他の文字は動かさない) を**巡回置換**といい，

$$\pi = (a_1, a_2, \cdots, a_r)$$

と略記します．

n 文字の置換全体のなす集合を**対称群**といい，S_n で表します．S_n に含まれる置換の個数 (**位数**という) は $n!$ です．

$\sigma, \tau \in S_n$ の**積**を写像の合成 ($\sigma\tau$ と書く) で定義すれば，対称群 S_n はいわゆる**群**になります．すなわちこの積は次の性質を持ちます:

(G1) 結合法則：$(\sigma\tau)\rho = \sigma(\tau\rho)$ $(\sigma, \tau, \rho \in S_n)$;
(G2) 単位元の存在：恒等写像 ϵ について，$\epsilon\sigma = \sigma\epsilon = \sigma$ $(\sigma \in S_n)$;
(G3) 逆元の存在：$\sigma \in S_n$ の逆写像 σ^{-1} について，$\sigma\sigma^{-1} = \sigma^{-1}\sigma = \epsilon$.

部分集合 $G \subseteq S_n$ が S_n の**部分群** (あるいは $[n]$ 上の**置換群**) であるとは,

(S1) $\sigma, \tau \in G$ なら $\sigma\tau \in G$;
(S2) $\epsilon \in G$;
(S3) $\sigma \in G$ なら $\sigma^{-1} \in G$

が成り立つことをいいます (G が空集合でなければ，条件 (S2), (S3) は不要です)．例えば長さ n の巡回置換 $\gamma = (1, 2, \cdots, n)$ の積をくり返して得られる部分集合

$$C_n = \{\epsilon, \gamma, \gamma^2, \cdots, \gamma^{n-1}\}$$

は位数 n の置換群になります (以下巡回置換群).

ここまでくるとオズワルトのシフト法と巡回置換群 C_n の関係が見えて来ます．オズワルトのシフト法で生じた問題点は，C_n の代わりにより大きな置換群，特に対称群 S_n を用いれば解決するのです

比較したい言語を LA, LB とし，LA の n 個の単語 $V_1, V_2, \cdots, V_n$，およびそれらに対応する意味を持った LB の単語 $W_1, W_2, \cdots, W_n$ を用意します．V_i の語頭音を $f(i)$ で，W_i の語頭音を $g(i)$ で表せば，f, g は集合 $N = \{1, 2, \cdots, n\}$ から音の集合 Q への写像と見なせます．

$$f, g : N \longrightarrow Q.$$

このとき粗点 x_0 は次のように表せます．

$$x_0 = \sharp\{i \in N \mid f(i) = g(i)\}.$$

($\sharp$ は集合の元の個数)．置換 $\pi \in S_n$ によって LB の単語をかき混ぜ，それによる一致数を $x(\pi)$ とすれば，この一致数は偶然による一致数 (背景点) のはずです：

$$x(\pi) = \sharp\{i \in N \mid f(i) = g(\pi i)\}.$$

置換の供給源としては適当な置換群 G を使うことにします．このとき，背景点の**平均値** m と**分散** s^2 は次で定義されます．

$$m = \frac{1}{|G|}\sum_{\pi \in G} x(\pi), \quad s^2 = \frac{1}{|G|}\sum_{\pi \in G}(x(\pi) - m)^2.$$

そこで偏差値 $x = (x_0 - m)/s$ が 2.33 以上のとき，粗点 x_0 の大きさは (1 パーセント水準で) 偶然によるものより大きい，すなわち LA と LB は関係ありと判断することになります．オズワルトのシフト法は，ちょうど G が $\gamma = (1, 2, \cdots, n)$ で生成される巡回群 C_n の場合です．

置換群 G を使ったシフト法において，語彙リストの順番を変えるということは，ある $\sigma \in S_n$ によって，$\{x(\sigma\pi\sigma^{-1}) \mid \pi \in G\}$ を背景点の集合 (重複を許す集合) とすることを意味します．つまり背景点の分布が語彙リストの順番に依存しないためには，どんな $\sigma \in S_n$ と $\pi \in G$ に対しても $\sigma\pi\sigma^{-1} \in G$ でなければなりません．「群論」の用語を使えば，次のようになります：

群 G は，対称群 S_n の正規部分群でなければならない．

オズワルトのシフト法で使った巡回群 C_n は，$n \geq 4$ では S_n の正規部分群になりません．単位元だけからなる自明な部分群 $\{\epsilon\}$ と，S_n 全体は，S_n の正規部分群です．また偶数符号の置換全体の集合 (**交代群**)

$$A_n = \{\pi \in S_n \mid \mathrm{sgn}(\pi) = 1\}$$

も対称群の正規部分群です．実は次の定理が知られています：

定理：$n \neq 4$ のとき，対称群 S_n の正規部分群は，$\{\epsilon\}$，A_n，S_n に限る．

このことから，4 次以下の対称群は，巡回群を積み上げて得られる (可解群という) のに対し，5 次以上の対称群はそうでない (非可解群である) ことが分かります．

[**ガロアと群論**] ガロア (1811–1832) は，5 次以上の代数方程式に解の公式がないことを，上の定理に帰着させて証明しました．可解，非可解という言葉もそこから来ています．ガロアが考えた群の概念は，対称性を測るための代数系として，今や数学はもちろん，物理や化学などに広く使われています．生物学や人類学や心理学でも使われたことがあるそうです．昔はやったルービックキューブというパズルを解くのにも群論が役立ちました．

全体と単位群だけしか正規部分群を持たない群を**単純群**といいます．素数位数の巡回群は単純群ですし，5 次以上の交代群も単純群です．有限位数の単純群の完全な分類は，ごく最近になって完全に終了しました．巡回群と交代群の他に，有限体上の線形群 (一般線形群，直交群，ユニタリ群など) から得られるたくさんの単純群がありますが，なぜか有限単純群にはこれら以外に 26 個の例外 (散在型単純群) があります．例外的な単純群の中の最大のものはモンスター群とよばれていて，位数は

$$2^{46} \cdot 3^{20} \cdot 5^9 \cdot 7^6 \cdot 11^3 \cdot 13^3 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 41 \cdot 47 \cdot 59 \cdot 71$$

です．モンスター群についてはいろいろ不思議なことが知られていて，モンスター群の発見こそ有限群論最大の成果という人もいます．

7　対称群を使うシフト法

シフト法に使える群は，対称群と交代群だけであることが分かりました．ここからは対称群 S_n を使うシフト法 (完全シフト法) を考えましょう．この方法だと，シフト法の問題点 (b) と (c) は解決します．しかし今度は，オズワルトのシフト法で解決したはずの計算の手間がまた問題になります．つまり，対称群の位数 $n!$ は大きすぎて，コンピュータでも扱えないのです．実際，**スターリングの公式**

$$n! \approx \sqrt{2\pi} n^{n-1/2} e^{-n} \tag{2}$$

によれば

$$100! \approx 9.33 \times 10^{157}, \quad 200! \approx 7.9 \times 10^{374}$$

となって，これほどの回数の単語の比較は計算機でもどうしようもありません．交代群を使っても五十歩百歩です．

そこで群論の出番です．先ほどと同じく，$f, g : N \longrightarrow Q$ (ここで $N = [n] = \{1, 2, \cdots, n\}$ で Q は音の集合) を，言語 LA, LB の i 番目の単語の語頭音を対応させる写像とします．このとき $\lambda \in Q$ の逆像 $A_\lambda = f^{-1}(\lambda), B_\lambda = g^{-1}(\lambda)$ は，それぞれ言語 LA と LB で音 λ で始まる単語の (番号の) 集合です．さらに $a_\lambda = |A_\lambda|, b_\lambda = |B_\lambda|$ とおきます．対称群を使ったシフト法の場合，背景点 (偶然による一致数) の集合は $\{x(\pi) \mid \pi \in S_n\}$ です．驚くことに次の結果が成り立ちます：

定理 (モメント公式) $\dfrac{1}{n!}\displaystyle\sum_{\pi \in S_n} \binom{x(\pi)}{k} = \frac{(n-k)!}{n!} \sum_{\sum k_\lambda = k} \prod_\lambda \binom{a_\lambda}{k_\lambda}\binom{b_\lambda}{k_\lambda} k_\lambda!$

ここで，右辺の和は $\sum k_\lambda = k$ となる整数 $k_\lambda \geq 0$ の組全体を動きます．$\displaystyle\prod_\lambda$ は λ に関する積を表す記号です．

系． 背景点の平均 $m = \dfrac{1}{n!}\displaystyle\sum_{\pi \in S_n} x(\pi)$ と分散 $s^2 = \dfrac{1}{n!}\displaystyle\sum_{\pi \in S_n} (x(\pi) - m)^2$ は次の式で与えられる：

$$\begin{aligned} m &= \frac{1}{n}\sum_\lambda a_\lambda b_\lambda \\ s^2 &= \frac{1}{n-1} m(m+n) - \frac{1}{n(n-1)} \sum_\lambda a_\lambda b_\lambda (a_\lambda + b_\lambda) \end{aligned}$$

直積集合 $X \times Y$ とは，ふたつの元の対 (x, y) (ただし $x \in X, y \in Y$) 全体のなす集合です．部分集合 $A \subset X \times Y$ と $x \in X, y \in Y$ に対し，

$${}_xA = \{y \in Y \mid (x, y) \in A\}, \quad A_y = \{x \in X \mid (x, y) \in A\}$$

とおけば，

$$|A| = \sum_{x \in X} |{}_xA| = \sum_{y \in Y} |A_y|.$$

この等式を使って集合のサイズを計算することを「**二通りに数える方法**」といいます．ここでは系 1 をこの方法で証明してみましょう．

(系 1 の証明) 集合

$$A = \{(i, \pi) \in N \times S_n \mid f(i) = g(\pi(i))\} \subset N \times S_n$$

とする．このとき二通りに数える方法により

$$\begin{aligned} |A| &= \sum_{i=1}^{n} \sharp\{\pi \in S_n \mid f(i) = g(\pi(i))\} \\ &= \sum_{\pi \in S_n} \sharp\{i \in N \mid f(i) = g(\pi(i))\}. \end{aligned}$$

ここで第一の和を計算すると，

$$\begin{aligned} |A| &= \sum_{i=1}^{n} \sharp\{\pi \in S_n \mid f(i) = g(\pi(i))\} \\ &= \sum_{i} \sum_{\lambda} \sharp\{\pi \in S_n \mid f(i) = g(\pi(i)) = \lambda\} \\ &= \sum_{\lambda} \sum_{i} \sharp\{\pi \in S_n \mid i \in f^{-1}(\lambda) = A_\lambda, \pi(i) \in g^{-1}(\lambda) = B_\lambda\} \\ &= \sum_{\lambda} \sum_{i \in A_\lambda} \sum_{j \in B_\lambda} \sharp\{\pi \in S_n \mid j = \pi(i)\}. \end{aligned}$$

ところで，任意の $i, j \in N$ に対し，$\pi(i) = j$ となる $\pi \in S_n$ の個数は，i, j によらず $(n-1)!$ なので，

$$\begin{aligned} \therefore\ |A| &= \sum_{\lambda} \sum_{i \in A_\lambda} \sum_{j \in B_\lambda} \sharp\{\pi \in S_n \mid j = \pi(i)\} \\ &= (n-1)! \sum_{\lambda} |A_\lambda| \cdot |B_\lambda| = (n-1)! \sum_{\lambda} a_\lambda b_\lambda. \end{aligned}$$

第二の和については

$$\begin{aligned} |A| &= \sum_{\pi \in S_n} \sharp\{i \in N \mid f(i) = g(\pi(i))\} \\ &= \sum_{\pi \in S_n} x(\pi) = n!\, m. \end{aligned}$$

両者を比べて平均値の公式 $m = (1/n) \sum_\lambda a_\lambda b_\lambda$ を得る．(証明終)

8　安本の二項検定法

シフト法で残った最大の問題点は，x_0 以上の粗点が偶然得られる確率 $P(x_0)$ の正確な値の計算が困難なことです．この状況は対称群を使ったシフト法でも同じで，粗点が x_0 以上になる正確な確率 $P(x_0)$

$$\begin{aligned} P(x_0) &= \sum_{\sum x_\lambda \geq x_0} \prod_\lambda p_\lambda(x_\lambda) \\ p_\lambda(y) &:= \binom{a_\lambda}{y}\binom{n-a_\lambda}{b_\lambda - y} \Big/ \binom{n}{b_\lambda} \quad \text{(超幾何分布の密度関数)} \end{aligned}$$

を計算するのはいかにも大変です．したがって，これまでは偏差値 $z = (x_0 - m)/s$ に対する正規上側確率 $Q_n(z)$ を $P(x_0)$ として採用していました．正規分布の上側確率 $Q_n(z)$ は，優れた数値計算公式があって計算に困ることはないのですが，正確な確率の値に比べて，ときには何桁も小さくでることがあります (特に x_0 が m に比べて大きい場合)．もっとも，偶然の確率が百万分の一でも１万分の１でも結論が変わりはありません．そのほかエッジワース展開を使う手もあるのですが，計算公式は相当複雑になります．

シフト法の持つこの最後の難点は，安本美典氏が提案した二項検定法で解決します．これはポリアの検定法とオズワルトのシフト法のハイブリッド方式といえます．その考えは次のようなものです．n の基礎語彙を使うことにします．$N = [n] = \{1, 2, \cdots, n\}$ は語彙リストの項目番号の集合とします．対称群を使ったシフト法では，片方の言語 LB の単語だけを置換 π でかき混ぜました．π に対応する背景点は

$$x(\pi) = \sharp\{i \in N \mid f(i) = g(\pi(i))\} \quad (\pi \in S_n)$$

で与えられます．ここで $f, g : N \longrightarrow Q$ は，前と同じく項目番号に語頭音を対応させる写像です．

LB だけでなく LA もかき混ぜましょう．そうすると，背景点は

$$y(\sigma, \tau) = \sharp\{i \in N \mid f(\sigma(i)) = g(\tau(i))\} \quad (\sigma, \tau \in S_n)$$

となります．特に背景点の個数は $n!^2$ です．容易に分かるように，$y(\sigma, \tau) = x(\sigma^{-1}\tau)$ ですから，背景点の分布は前の LB だけをかき混ぜる方式と変わりません．

今までは語彙のかき混ぜに順列 (置換) を使っていましたが，重複順列を使ってみましょう．つまり，対称群 S_n の代わりに，重複順列全体の集

合 E_n を使うのです．ちなみに E_n は，N から N への写像全体の集合であり，**対称半群**といいます．E_n 上の合成は結合法則を満たし，恒等写像が単位元となっています．しかし逆元は一般に存在しないので，群にはなりません．対称半群を使った場合，背景点 (偶然による一致数) は

$$y(\sigma,\tau) = \sharp\{i \in N \mid f(\sigma(i)) = g(\tau(i))\} \quad (\sigma,\tau \in E_n)$$

となります．背景点の個数は n^{2n} です．この場合，そう難しくなく，背景点の分布が二項分布 $B(n,p)$，ここで $p = (1/n^2)\sum a_\lambda b_\lambda$，であることが分かります．したがって，この方式では，上側確率

$$P(x_0) = \sum_{r=x_0}^{n} \binom{n}{r} p^r (1-p)^{n-r}$$

が例えば 0.05 未満のとき，粗点 (語頭音の一致数) x_0 の大きさは偶然でない (有意水準 5 パーセント) と判断します．二項分布の上側確率なら，二項分布を正規分布で近似する公式がいくつもあるし，直接計算機にやらせても大した時間はかかりません．

なぜ二項分布になるかは，**ポリヤの壺の問題**から容易に理解できます．ふたつの壺にそれぞれ n 個の玉が入っており，玉にはそれぞれ発音記号が書いてあります．発音記号 λ が書いてある玉は，それぞれの壺に a_λ 個と b_λ 個入っています．ふたつの壺からひとつずつ玉を取り出して，一致するかどうかを見ます．ふたつの壺から一つずつ玉を取り出し，同じ発音記号が書いてあったらそれを記録します．今度は取り出した玉をもとの**壺にもどし**，それからまたふたつの壺から一つずつ玉を取り出し同じことをします．この操作を n 回繰り返します．1 対ずつ玉を取り出してもどす n 回の操作が重複順列 σ,τ に当たります．それぞれの壺からひとつずつ取った玉の発音記号が偶然一致する確率 p は

$$p = \frac{1}{n^2}\sum_{\lambda\in Q} a_\lambda b_\lambda.$$

で与えられます．p は，シフト法の一致の確率 m/n と一致しています．また，玉の対を取り出す操作は毎回独立なので，一致数の分布は二項分布 $B(n,p)$ で与えられます．

安本氏によるこの検定方式は，ポリアの方法と似ていますが，偶然の一致率 p はシフト法のものです．ポリアのモデル II の方式では，$\sum_\lambda p_\lambda^2$

で，語頭音が λ である確率 p_λ として，両方の言語で共通の値を使っていました．ポリアの方法も，単語の総数が少ない場合は有効です．

ポリアの壺で考えると，ポリアと安本の検定法はともに復元抽出でした．非復元抽出 (取り出した玉をもどさない) だと，対称群を使ったシフト法になります．この場合，背景点の分布が (独立とは限らない) いくつかの**超幾何分布**の和であることも納得できます．

対称群を使ったシフト法も，ポリアと安本の二項検定法も，統計学的な根拠を持っています．例えば基礎 100 語で比較する場合，その基礎 100 語を母集団と考えるならシフト法の方がふさわしく，二項検定法はその近似と考えられます．しかし基礎 100 語を，ふたつの言語の持つ膨大な単語の中からランダムに取った標本と考えるなら，二項検定法の方がふさわしいと考えられます．いずれにしても，考えている言語における単語の総数が大きいなら，n 個の単語の非復元抽出も復元抽出も大した違いはありません．それなら，二項検定法を使うという手法は考え方も明白だし，計算が単純で実用的に優れています．これなら，正規近似による一致数の過大評価もありません．シフト法が精緻でやや過敏なのに比べると，二項検定法は実用に優れたタフな方法だと思います．

9　言語群同士の比較

3 つの言語 LA, LB, LC についてのシフト法もあります．それぞれの言語で，n の単語からなる語彙リストを用意しておきます．次の記号を用意します：

・$a_\lambda, b_\lambda, c_\lambda$：それぞれの言語で音 λ で始まる単語の個数．
・x_{AB}, x_{BC}, x_{AC}：ふたつの言語で同じ語頭音を持つ単語の数．
・x_{ABC} ：3 言語すべてで同じ語頭音を持つ単語の数．
・$m_{AB}, m_{BC}, m_{AC}\, s^2_{AB}, s^2_{BC}, s^2_{AC}$：2 言語間の一致率平均と分散．

(方式 1) 偶然による $x^{(2)}_{ABC} = x_{AB} + x_{BC} + x_{AC}$ の平均値 m' と分散 s'^2 は次で与えられます：

$$m' = m_{AB} + m_{BC} + m_{AC}, \quad s'^2 = s^2_{AB} + s^2_{BC} + s^2_{AC}.$$

したがってこの方式で $x^{(2)}_{ABC}$ の大きさが偶然で得られるかどうかの検定が出来ます．独立でもない確率変数の和について，分散がこのようにきれいに表されるのはおもしろいことです．

(方式 2) x_{ABC} についての平均 m'' と分散 s''^2 は，次で与えられます：

$$\begin{aligned} m'' &= \frac{1}{n^2}\sum_{\lambda\in Q} a_\lambda b_\lambda c_\lambda \\ s''^2 &= \frac{2n-1}{(n-1)^2}m''^2 + \frac{n^2-2n}{(n-1)^2}m'' \\ &\qquad - \frac{1}{n^2(n-1)^2}\sum_{\lambda\in Q} a_\lambda b_\lambda c_\lambda (a_\lambda + b_\lambda + c_\lambda) \end{aligned}$$

これらの公式で上古日本語・現代アイヌ語・中期朝鮮語を比べてみると，偶然でこれほど大きな粗点 ($x^{(2)}_{ABC} = 151$ と $x_{ABC} = 23$) は得られないことが分かります (有意水準 1 万分の 1 以下)．したがって日本語・アイヌ語・朝鮮語の 3 言語の基礎語彙には偶然以上の語頭子音の一致があるとの結論が得られます.

	J × A	J × K	A × K	JAK(1)	JAK(2)
x_0	41	53	57	151	23
m	36.57	36.155	37.555	109.88	8.2553
s	5.0987	5.1647	5.1922	8.9160	1.9850
z	0.8688	3.2615	3.7450	4.6119	4.9396
P(正規)	0.1925	0.0^3554	0.0^4902	0.0^5252	0.0^6511
P(二項)	0.2450	0.0^2238	0.0^3658	0.0^4227	0.0^4123

表 5: 日本語・朝鮮語・アイヌ語の比較

10　日本語の謎と不思議

日本語に関してはたくさんの不思議があります.

(1) 数詞の語頭子音の倍数構成 (1:hi と 2：hu，3:mi と 6: mu，4:yo と 8: ya，5:itu と 10: to)．確率計算から，この対応が偶然とは思えません．しかし，何千とある言語の数詞を調べても，倍数構成の痕跡を持つ言語はハンガリー語などきわめてわずかです．このような数詞の倍数構成は，日本列島の内で生じたのでしょうか，それとも外で生じたのでしょうか．そもそも日本語の数詞はどこでどのようにして生まれたのでしょうか.

(2) 高句麗語の数詞.『三国史記』という高句麗・百済・新羅三国の正史を表した 12 世紀高麗の書物の中に，高句麗の地名がたくさんあります．広辞苑の編者として有名な新村出は，その中に数多くの日本語と思われる単語を見つけました (1913)．例えば,「三紛縣（一云密波兮)」「五谷郡（一云于次呑忽)」「七重縣（一云難隠別)」「十谷縣（一云徳頓忽)」とあります．つまり数詞「三」「五」「七」「十」の一云 (読み) が「ミツ」「ウチャ」「ナン」「トク」だというのです．こうなると両言語での類似は明らかでしょう．よくこのようなことを発見したと驚きます．数字以外の文字についても，日本語との対応がいくつか発見されています．これらの語が本当に高句麗語 (朝鮮語系) の単語なのか疑問が残りますが，それならいったいどのようにして日本語の単語があのような場所にまで入り込んだのでしょうか.

(3) 日本列島に到来した言語．日本列島には１万年以上前から様々な文化が流れ込んできました．例えば，最近の考古学の研究で 3500 年前の縄文時代にはすでに稲作がある程度の行われていたことが分かってきました．それなら，そのような縄文稲作をもたらした人々の言葉は何だったのでしょうか.

安本美典氏の『日本語の誕生』や『言語の科学』に，基礎 200 語の語頭音を比較では，日本語 (上古日本語，現代東京方言，首里方言) がいくつかの言語 (朝鮮語，中国語，インドネシア語，カンボジア語など) と統計学的に偶然とは言えない関係を持つことが示されています．このような関係は統計学とコンピュータの助けを借りて初めて検出できるものであって，従来の比較言語学の方法で検出するのは難しいでしょう．安本氏は，そのほかにもビルマ系のいくつかの言語の中に日本語の身体語とよく似た単語を持つものがあることを発見しています．これらの言語が本当に日本語に何らかの影響を与えたというなら，どんな人々がどの時代にどのルートで日本列島にもたらしたのでしょうか．彼らの痕跡は日本の文化や遺伝子に残っているのでしょうか．いずれにせよ，これらばらばらの言語との関連は，日本語の起源探求に，系統樹のモデルより流入説のモデルの方を強く支持しているように思います.

日本語の「旦那」や「僧」「袈裟」はサンスクリット語起源です．日本語の「将棋」，中国語の「象棋」，ヨーロッパの「チェス」は，インドの「チャトランガ」というゲームが元になっているそうですが，発音が何となく似ています．文化的な単語の中には，思いもかけない世界的な広がりでの借用関係があるのかもしれません.

11　ブートストラップ法

最近の統計学の見地からすれと，オズワルトのシフト検定法 (1970) は，典型的な**リサンプリング法**です．リサンプリング法とは，シフト法同様，すでに得られたデータからのサンプリングによって推定値のバラツキを評価する一種のシミュレーション法です．その考えは 1930 年代に始まる (フィッシャーの並べ替え検定) のですが，大量の計算を必要とするため，実用は計算機の発達してきた 1960 年代中頃からです．代表的なブートストラップ法 (Efron 1979) は，今では標準的な統計手法として広く使われています．

ブートストラップ法のもっとも簡単な例として，日本列島の住人の平均年齢を知りたいとします．全部の人の年齢を調べるのは大変ですから，ランダムに n 人を選んで彼らの年齢のデータ $x_1, x_2, \cdots, x_n$ を集めます．母集団の平均年齢 μ の推定値は，$\overline{x} = (x_1 + \cdots + x_n)/n$ ですが，問題はその精度です．そこで $x_1, x_2, \cdots, x_n$ からランダムに (重複を許して) n 個取ります：$x_1^*, \cdots, x_n^*$ (リサンプリングという)．これを母集団からの新たな標本と見なして平均値 $\overline{x}^* = (x_1^* + \cdots + x_n^*)/n$ を求めます．この操作を多数回くり返すことによって，$\overline{x}^*$ のばらつきが分かり，さらに $\overline{x}$ の信頼度も分かるという仕掛けです．シフト法のように，$\overline{x}^*$ のヒストグラムを書くこともできます．母集団の年齢分布について何の仮定もなしに使える便利な方法 (ノンパラメトリック法) です．

$x_1, \cdots, x_n$ から重複を許して n 個の $x_1^*, \cdots, x_n^*$ 選ぶことは，$1, \cdots, n$ の重複順列 $\rho = \rho(1)\,\rho(2)\,\cdots\rho(n)$ を (同じことだが $[n] = \{1, \cdots, n\}$ から $[n]$ への写像) 選ぶことに他なりません．ふつうのブートストラップ法は重複順列をランダムに取ります．しかし重複順列をランダムに取る代わりに，シフト法のように，すべての重複順列を取り対応する平均値の分布を取ったらどうでしょう．これはリサンプリングの回数を無限大にした場合のブートストラップ法に相当します．シフト法のモメント公式の証明と同様にして (あるいは多項定理により)，次の定理が証明できます．

定理． $1, 2, \cdots, n$ から m 個取った重複順列 $\rho = \rho(1)\cdots\rho(m)$ に対し，

$$\widehat{x}_\rho = \frac{1}{n}(x_{\rho(1)} + \cdots + x_{\rho(m)})$$

とおく．このとき t を不定元とする形式的ベキ級数として，

$$\sum_{k=0}^{\infty}\left(\frac{1}{n^m}\sum_{\rho}\binom{\widehat{x}_\rho}{k}\right)t^k = \left(\frac{1}{n}\sum_{i=1}^{n}(1+t)^{x_i/n}\right)^m.$$

ここで ρ は $1, 2, \cdots, n$ 個から m 個取る重複順列全体を動く.

系. $x_1, \cdots, x_n$ の平均，分散，歪度，尖度をそれぞれ，$\mu, \sigma^2, \gamma_1, \gamma_2$ とすれば，$\widehat{x}_\rho$ (ρ は n 個から m 個取る重複順列全体を動く) の平均 M，分散 S^2，歪度 G_1，尖度 G_2 は次で与えられる：

$$M = \mu, \quad S^2 = \frac{1}{m}\sigma^2, G_1 = \frac{1}{\sqrt{m}}\gamma_1, G_2 = \frac{1}{m}\gamma_2.$$

特に $m = n$ の場合の $S = \sigma n^{-1/2}$ は，よく知られた公式です．他にも統計量についてもこのような厳密なモメント公式があるかもしれません.

12 系統樹

ふたつの現代の言語 LA と LB が共通の言語から分岐したのが今から t 年前だとし，$\phi(LA, LB) = t$ とおきます．$\phi(LA, LB)$ の値が小さければ LA と LB は近い言語であり，単語も似ているでしょう．このとき次の性質が成り立ちます.

(S1) $\phi(LA, LB) \geq 0$ で，$\phi(LA, LB) = 0 \iff LA = LB$.
(S2) $\phi(LA, LB) = \phi(LB, LA)$.
(S3) $\phi(LA, LC) \leq \max\{\phi(LA, LC), \phi(LC, LB)\}$.

このような性質を満たす関数 ϕ を一般に，**超距離関数**といいます．したがって，現代ある言語の集合は**超距離空間**となります.

p-進解析. 超距離とは勇ましい名前ですが，数学的に重要な例 (非アルキメデス的付値) があります．整数 n が素数 p で割り切れる回数を $\mathrm{ord}_p(n)$ とし，ふたつ整数 x, y の **p-進距離** $\phi_p(x, y)$ を次のように定義します.

$$\phi_p(x, y) := p^{-\mathrm{ord}_p(x-y)}$$

で定義します．このとき ϕ_p は超距離関数になり整数の集合 $\mathbb{Z}$ は超距離空間になります．この距離空間では $\lim_{n\to\infty} p^n = 0$ です！あとは有理数からの実数の構成と同様，コーシー列が収束するように完備化して，関数の極限や微分積分を定義すれば p-進解析学のできあがりです.

言語の分岐の状況は**樹形図** (デンドログラム) によって表すことが出来ます．図はヨーロッパの言語の数詞についてのデータから作った樹形図です．樹形図の枝 (横枝) の長さは，正確ではないにせよ，分岐してから

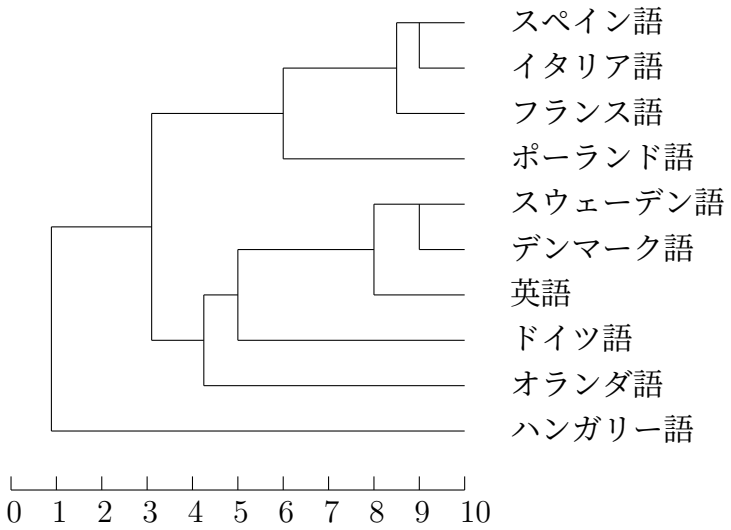

図 2: 数詞によるヨーロッパの言語のクラスター分析

の時間を反映しています．逆に超距離空間は，樹形図から得られることが分かります．

さて，言語の集合にはもう一つの距離関数があります．今基礎 n 語で，言語 LA と LB を比較することにしましょう．

$$d(LA, LB) = n - (\text{基礎 } n \text{ 語での LA と LB の語頭音の一致数}).$$

とおけば，関数 d は**距離関数** (ハミング距離，市街化距離，マンハッタン距離などという) になり，言語の集合は**距離空間**になります．すなわち次の性質が成り立ちます：

(D1) $d(x, y) \geq 0$ で，$d(x, y) = 0 \iff x = y$.
(D2) $d(x, y) = d(y, x)$.
(D3) $d(x, z) \leq d(x, y) + d(y, z)$ (三角不等式).

超距離空間は距離空間ですが，逆は一般には成り立ちません．また過去の言語も含めた集合は，$d(LA, LB)$ を LA と LB が分岐してから経過した時間の和とすれば，距離空間になりますが，超距離空間の公理を満たしません．このような距離空間の満たす公理系を見つけるのは良い練習問題です．

今言語の系統樹といいましたが，過去の分岐の様子や時期は，なかなか分かりません．実際には，現代の言語の様子から，言語の過去の分岐の様子

や時期を推定することになります．そこで，現代の言語 $X = \{LA_1, \cdots, LA_m\}$ について，一致数から得られる距離関数 $d(LA_i, LA_j)$ は分かっているとします．このふつうの距離空間 (X, d) から，系統樹 (すなわち超距離空間) ϕ を再構成することを統計学では**クラスター分析**といいます．

$X = \{x_1, \cdots, x_m\}$ を距離空間，d をその距離関数としたとき，求める超距離空間 (X, ϕ) として，

$$\sum_{x,y \in X} (d(x,y) - \phi(x,y))^2 \longrightarrow \mathrm{Min}$$

を満たすものを採用するのが自然でしょう．数学的には，超距離関数の**モジュライ空間**

$$\{(\phi(x_i, x_j))_{i,j} \in \mathbb{R}^{m^2} \mid \text{(S1), (S2), (S3)}\} \subset \mathbb{R}^{m^2}$$

内の点で,「距離行列」 $(d(x_i, x_j))_{i,j} \in \mathbb{R}^{m^2}$ にもっとも近いものを探す極値問題とも解釈されます．

上の条件を満たす超距離関数 ϕ を具体的に求めるアルゴリズムは，多変量解析学で**群平均法**として知られています．群平均法は，生物の系統樹を作るのによく使われています．具体的には次のように超距離関数を構成します．

群平均法のアルゴリズム． $X = \{x_1, \cdots, x_m\}$ で，$d(x, y)$ を X 上の距離関数とする．
(ステップ 1) 初期クラスターを $G_1 = \{x_1\}, \cdots, G_m = \{x_m\}$ とし，クラスター間の距離を $d(G_i, G_j) = d(x_i, x_j)$ とする．
(ステップ 2) クラスター数が 1 なら終わり．そうでなければ，$d(G_i, G_j)\,(i \neq j)$ の中で最小のものを探す．番号を付け替えてそれを $d(G_1, G_2)$ とする．
(ステップ 3) クラスター G_1 と G_2 を融合して新しいクラスター G_{12} を作る．$i \geq 3$ に対し

$$d(G_{12}, G_i) = \frac{1}{|G_1| \cdot |G_2|} \sum_{x \in G_{12}} \sum_{y \in G_i} d(x, y)$$

とすれば，新しいクラスター集合 $\{G_{12}, G_3, \cdots\}$ は距離空間になる．そこで，(ステップ 2) にもどる．

超距離空間の代わりに別の距離空間を使うこともできます．例えば距離空間 (X, d) を，できるだけ無理 (ストレス) がかからないように，ユークリッド平面 $\mathbb{R}^2$(またはより高次元の空間) に埋め込む問題です．この手

の問題は，多変量解析学という統計学の分野で，多次元尺度構成法など，多くの方法が開発されています．先ほどの樹形図は，前にあげたヨーロッパの１０言語のクラスター分析の結果です．

13　さらなる発展と課題

比較言語学における数理科学的方法をいくつか紹介してきました．最近の成果と今後の課題を追加しておきます．

(A) **言語年代学と言語の系統．** 言語年代学では，$x_0(t) = x_0(0)0.81^t$ (t は千年，$x_0(t)$ は時刻 t における一致数) のような公式が使われています．0.81 というのは，千年あたりの単語の残存率です．この公式で，$t \to \infty$ の極限を取ると，$x_0(t) \to 0$ となりますが，これはおかしい．この極限値は，偶然の一致数のはずです．したがって，この公式は何らかの修正が必要です．

実際，確率過程の理論を使うと，t 千年後の一致数は

$$x_0(t) - m(t) = (x_0(0) - m(0))r^t$$

となります．ここで $m(t)$ は時刻 t での偶然による一致数です．応用として，どれくらい前に分岐した言語の間の関連が検出できるかという問題を考えてみます．簡単のため，語彙数 $n = 200$，偶然の一致率 $p = 40/200 = 0.20$ とすれば，語頭音の一致数が 50 以上だと二項分布の上側確率 0.04935 以下となり，偶然でない (有意水準 5%) と判断されます．そこで， 千年あたりの残存率 $r = 0.8$，分岐したのが今から t 千年前として計算すれば，

$$\begin{aligned} x_0 - np &= r^{2t}(n - np) \\ \therefore\ x_0 - 40 &= 160 \times (0.64)^t. \end{aligned}$$

これより $x_0 \geq 50$ となるのは，

$$t \leq \log\left(\frac{50-40}{160}\right) \Big/ \log(0.64) = 6.21257$$

すなわち６千年以上前に分岐した言語の関係が検出できるのです．実際には，もっと古い関係まで検出しているのではないかと思わせるデータ (東北アジアとアメリカ先住民の言語の関係) もあります．

この種の問題は集団遺伝学で盛んに研究されてきました．またこの公式は，クラスター分析などで言語や遺伝子・タンパク質・生物の系統樹を

作るときの枝の長さ (分岐年代) への注意を喚起してます．樹形樹の枝の長さは分岐してからの時間に比例するように描くのが自然ですが，系統樹の多くはそれを考慮していません．また一致数と偶然の一致数が近い (したがって同系かどうか判定できない) ふたつの言語 (や遺伝子) を同じ系統樹に入れて，それが実際の系統樹であるかのように述べるのも問題があるでしょう．要は，クラスター分析 (さらに多変量解析) は分類の技法であって，検定ではないと言うことです．そもそも系統樹を作るのに必ずしもクラスター分析がふさわしいかも大きな問題です．超距離関数よりも弱い何らかの条件を満たす距離関数を見いだす極値問題を考える必要があります．これについては数理分類学では 1970 年代から 4 点条件

$$\varphi(a,b)+\varphi(c,d) \le \max(\varphi(a,c)+\varphi(b,d), \varphi(a,d)+\varphi(b,c))$$

を満たす距離関数 φ が知られていたそうです (三中氏による).

日本語の起源を説明するモデルとして分岐説の他に流入説があると書きました．分岐説の数学モデルとしては系統樹 (枝に長さの付いた根つき木) がありますが，流入説の数学モデルはまだないと思います．候補としては，ネットワーク上の流れなどが挙げられます.

(B) **インドヨーロッパ語族の起源．** 最近，大規模なインドヨーロッパ語族の統計的研究が発表されました.

Russell D.Gray and Quentin D.Atkinson, Language-tree divergence times support the Anatolian theory of Indo-European origin, NATURE 426 (2003), 435–439.

インドヨーロッパ語族に属する 87 の言語の膨大な語彙データを使って，言語の系統樹を作っています．インドヨーロッパ語族の発祥の地として 7800 年前から 9800 年前のアナトリア (トルコ) が有力としています．語彙表は，スワデシュによる英語のリストに対応するものを使っています.

(C) **音韻対応の法則．** 比較言語学全体がそうなのですが，音韻対応の法則には，数理科学的な雰囲気がします．Q を発音記号の集合，λ, μ はそのような発音記号を表す変数とします．ふたつの言語 LA と LB の基礎 n 語の語彙リストにおいて，LA では λ で始まり，LB では μ で始まる同じ意味を持つ単語の対の個数を $x_{\lambda,\mu}$ とすれば，$Q \times Q$ 型の行列 $X = (x_{\lambda,\mu})$ が得られます．以前のように，写像 $f : N \longrightarrow Q$ (または $g : N \longrightarrow Q$) を，番号 $i \in N = \{1, \cdot, n\}$ に言語 LA (または LB) の i 番目の単語の語頭音を対応させる写像とすれば,

$$x(\lambda,\mu) = |f^{-1}(\lambda) \cap g^{-1}(\mu)|$$

です．そうなると言語 LA における音 λ が言語 LB の音 μ に対応するという音韻対応の法則を統計的に確かめるには，分割表 X の独立性の検定をすれば良いことになります．ただ，語頭音についての音韻対応の法則を見いだすのに，基礎 200 語くらいでは数が足りません．語頭音以外も含めた音韻対応の法則については何らかの数学的方法を工夫する必要があります．そのほか，祖語の再構成を自動的にやってくれる計算機プログラムを望むのは無理でしょうか．

(D) **暗号．**ある言語で，語頭音が λ である確率を p_λ としたとき，

$$p = \sum_\lambda (p_\lambda)^2$$

は，偶然ふたつの単語の語頭音が一致する確率です (ポリアの本にあるモデル II の式)．一方この確率は，暗号解読では，**一致反復率** (または反復生起率) とよばれる重要な量で，太平洋戦争中に日本軍の紫暗号解読に携わったフリードマンが暗号の周期を推測するのに使いました．シフト法の他の公式も，暗号理論で何か意味があるのかもしれません．

(E) **分子進化遺伝学．**ＤＮＡの比較によって，過去の人類の拡散や系統を調べる研究 (分子人類学) が現在盛んです．人類文化の中核をなす言語の系統の研究とは，互いに補い合う研究分野ですし，両方の分野の目標と方法には似たところもあります．ＤＮＡは 4 種類のアルファベットＡＣＧＴが並んだ単語であり，タンパク質はアミノ酸 (20 数種類ある) の並んだ単語と考えるなら，人間は単語が集まった辞書でしょうか．

比較言語学のアイデア (シフト法や音韻対応の法則) を分子進化遺伝学に応用したり，逆に遺伝学の方法を比較言語学に応用することが考えられます．実際シフト法は，塩基配列やアミノ酸配列の比較から生物の系統を調べるのによく使われるブートストラップ法や最尤法によく似ています (違ったところもある)．また，ふたつの生物で遺伝子の違いを数えて分岐年代を計算することはよくやることですが，そこの議論は言語年代学に使えそうです．特に言語の系統樹を作るのに，系統生物学での研究は役に立つと思います．そのほか数理生態学の拡散方程式を使って，言語の拡散や移動の数学理論を作ろうという試みもあるそうです．

言語と生物では違うところもあります．例えば，遺伝子は４種類の塩基が並んだものですが，遺伝子は語彙表に，塩基は語彙表の単語に対応させた方が良いこともあります．さらに生物の進化の様子は系統樹で描くことが出来ますが，言語の場合は，言語の分岐だけでなく，単語の借

用・転用などがあるため，系統樹よりも複雑なネットワーク (グラフ) の方が言語の進化を表現するのに良いでしょう．その場合，数学的な取り扱いは単純な系統樹に比べてはるかに難しくなりそうです.

生物の系統学では，クラスター分析などで得られた系統樹の信頼性の評価が重要な課題になっているそうです．この手の検定の問題は難しいのですが，最近ではブートストラップ確率で客観的な数値で評価することが可能になりました．しかしこの方法はまだ発展途上にあるようで，ブートストラップ確率を信頼しすぎは良くないという専門家もいます．現在では，心理学，考古学，歴史学，人類学，文献学，言語学など人文科学の分野でも自然科学的方法が使われ，昔とは研究方法が一新されつつあると感じます．推論の客観性のためには喜ばしいことでしょうが，反面結果の解釈の間違いや方法の誤用が目に付きます.

14　数学の核

抽象化したレベルでは，現実も応用も無視した自由な考察が可能になります．例えば，一致数 x_0 は $f : N \longrightarrow Q$ と $g : N \longrightarrow Q$ の**等化**

$$E(f, g) = \{i \in N \mid f(i) = g(i)\}$$

のサイズである，それなら等化の代わりにその双対概念であるところの**余等化**について，シフト法のような公式はないのか．有限距離空間 (距離正則グラフなど) に関する何らかの量の平均値の公式はないのか.

またブートストラップ法のモメント公式もいくつかのことを考えさせます．公式の分母をはらうと完全に形式的ベキ級数の等式になります．それならこの公式は，有限集合族の間の１対１対応を表しているだろうと考えてやってみると，それは局所有限トポスとよばれるタイプのカテゴリーにまで一般化出来て，表現論やホモロジー代数の言葉では加法的誘導写像と乗法的誘導写像の交換法則 (要するに多項定理) を表していることが分かります．さらに，また重複順列についての和が出て来ますが，そのような和の連続版は**経路積分**だろう．それならモメント公式の連続版は何か，証明に出てきた多項定理の連続版は何か，等々たくさんの問題が生じます.

生物の分類や系統さらに生態のような伝統のある分野でも，一昔前とは様子を一変して，最近では様々な数学が使われていることに驚きます．特に離散数学が役立っていると聞くのはこの方面の数学者にとって大き

な励みです．例えば，枝の長さを無視してそのトポロジー (つながり状態だけを考えた系統樹の形のこと) だけを考えるなら，系統樹は，ラベルと根の付いた木 (labelled rooted tree) という組合せ論の研究対象です．組合せ論だけでなく，データ構造とアルゴリズムのような計算機指向の離散数学でも盛んに研究されています．

系統樹に対する数学からのもう一つの見方を紹介しましょう．数学の標準的な研究方針では，個々の系統樹よりも (仮想的なものも含め) 系統樹全体を考えるという立場を取ります．個々の系統樹の研究は，系統樹全体の中でのその系統樹の位置を調べるという形で行います．この方針に沿えば，系統樹に対する次のような「カテゴリー論的解釈」も何か意味のあることかもしれません．

時刻 $t \leq 0$ に地球上にいたすべての生物種の集合を $B(t)$ とします．$s \leq t \leq 0$ のとき，種 $x \in B(t)$ には，祖先 $x' \in B(s)$ がいるはずです．したがって写像

$$B(s \leq t) : B(t) \longrightarrow B(s)$$

があります．当然ながら $B(t \leq t)$ は恒等写像であり，合成に関して

$$B(r \leq s) \circ B(s \leq t) = B(r \leq t), \quad (r \leq s \leq t \leq 0)$$

が成り立ちます．数学の用語を使うと，$B(t)$ $(t \leq 0)$ が区間上定義された集合値の**ファンクター** (関手) になることを意味します．

$$B : [T_0, 0] \longrightarrow \mathbf{Set}$$

T_0 は最初の生命の誕生日です．逆にこのようなファンクターは何らかの系統樹を表しています．

生物全体を距離空間と考えることも可能です．距離関数としては，例えば何か共通の遺伝子についてのハミング距離 (非一致数) が考えられます．したがって，系統樹を，$[T_0, 0]$ から，距離空間と連続写像のなすカテゴリー (圏) へのファンクター $t \longmapsto B(t)$ と見ることも出来ます．そうなると，生物の系統学の目的は，現在の切り口 $B(0)$ の距離空間から，このファンクターを復元することと解釈できます．ファンクターは**カテゴリー論** (圏論) という数学の基礎的分野における重要な概念です．系統樹全体もカテゴリーになります．しかもトポスとよばれるきわめて重要な種類のカテゴリーです．カテゴリー論によれば，トポスとは，直観主義論理のもとでの集合と写像のなすカテゴリーです．直観主義論理とは，二重否定がもとに戻らない論理であって，したがって背理法が使えません．

この論理での数学の証明は，計算機のプログラムに翻訳できることが知られています．カテゴリー論は情報の分野ではひとつの言語としての地位を獲得しています．生物や言語や統計，さらに確率過程の分野にもカテゴリーが使えるとうれしいのですが，そのような研究はまだないようです．

私たちはすでに比較言語学からはるかに遠い数学の深奥を見下ろすところまで来てしまいました．そろそろ筆 (いや指を) をおくことにします．

15　あとがき

数学は科学の研究に役立つと書きましたが，自分の研究に必要な数学がどこにあるか，あったとしてもどんな文献から必要なところだけつまみ食いすれば良いのか，なかなか分からないと思います．科学に必要となる数学のほとんどは，すでに数学者がどこかで見つけているはずです．例えば，ここで紹介した群論は 100 年以上前から発展してきた分野です．やはり数学の専門家の出番だと思います．最近は数理科学として，数学と他分野をつなぎの役を務める分野が大きく発展してきました．

序文でも述べましたが，新しい分野への数学の応用だとすると，その分野と数学にブレークスルーを引き起こすことがあります．私の専門分野 (群論，組合せ論，カテゴリー論) でも，新しい応用分野が広がることを夢に見ています．最後に，大学生には，文系や生物系であっても，数学 (線形代数や微積分，統計，出来れば微分方程式) の勉強してほしいものです．ひょっとすると将来，新しい研究の始まりになるかもしれません．

この記事は，高校生向けの講座や，大学での一般教育の講義と集中講義の内容を整理したものです．安本『言語の科学』をもとにしました．またやや古い本ですが，服部『日本語の系統』も参考にしました．まとめるに当たり，安本美典氏の論文をいくつか読み返してみたのですが，あまりの先駆けぶりに改めて驚かざるをえません．1972 年の雑誌『数理科学』には，日本語と周辺アジアの言語の数詞や基礎語彙の因子分析による分析がのっていました．この当時，比較言語学に統計学やコンピュータを使う研究はまだ黎明期にあったと思います．穿孔カードにバチンバチンと穴を開けてプログラムやデータを入力し，カードデックをかついで計算センターに運んでいた時代です．今ならパソコンで多変量解析のソフトが使え，多くの言語の数詞や基礎語彙のリストの入手など簡単で

す.「早すぎた発見，忘れられし論文」などとは考えたくもありませんが，日本語の起源あつかった本にはあまりにもめちゃくちゃな内容のものが目立ちます．素人の私にとって，安本氏の議論はきわめて説得力があります．今後の研究の方針としては，日本語に流入した言語と，遺伝子・民族・文化の対応を付け，その流入経路を決めることと思うのですが．

参考書を思いつくままに挙げておきます．日本語朝鮮語などの語彙リストは安本『日本語の誕生』のものを使いました．残念なことに，比較言語学は (歴史言語学も)，日本での研究は低調なのか，現在手に入る専門書はごくわずかです．斉藤『ＤＮＡから見た日本人』によれば，人類学分野でも，言語の系統に関心を持っている人はいるようです．遺伝学関係では，三中『生物系統学』が系統樹に関する数学的議論を含んでいます．系統樹のより数学的な研究がどのようなものか知りたい人には，Semple–Steel の本はいかが．

参考書

(1) 言語学
・服部四郎『日本語の系統』岩波文庫 (1959 の文庫版).
・安本美典『日本語の誕生』大修館書店 (1978).
・吉田知行「言語間の距離とシフト法」，数理科学 **258**(1984) 37–42.
・安本美典『言語の科学』朝倉書店 (1995).
・吉田和彦『比較言語学の視点―テキストの読解と分析』大修館書店 (2005)

(2) 人文系科学の数理的研究
・村上征勝『真贋の科学―計量文献学入門』朝倉書店 (1994).
・村上征勝『文化を計る―文化計量学序説』朝倉書店 (2002).

(3) 数学の応用
・ラインズ『物理と数学の不思議な関係』早川文庫 (2004).
・C.Semple–M.Steel,“Phylogenetics”, Oxford (2003)
(4) 遺伝学・生物学
・木村資生『集団遺伝学概論』培風館 (1960).
・木村資生『分子進化の中立説』紀伊國屋書店 (1986).
・根井正利『分子進化遺伝学』培風館 (1990).
・三中信宏『生物系統学』東京大学出版会 (1997).
・岸野洋久，浅井潔『生物配列の統計』岩波書店 (1993).
・斉藤成也『ＤＮＡから見た日本人』ちくま新書 (2005).
・重定南奈子『進入と伝播の数理生態学』東京大学出版会 (1992).
・崎谷満『ＤＮＡが解き明かす日本人の系譜』勉誠出版 (2005).
(5) サイト

・http://yamatai.cside.com/katudou/kiroku193.htm (安本氏の考える日本語の成立の概略)
・http://www.zompist.com/numbers.shtm (世界の 5000 以上の言語の数詞)
・http://www.ntu.edu.au/education/langs/ielex/IE-DATA (インドヨーロッパ系言語の基礎語彙)
・http://cse.niaes.affrc.go.jp/minaka/index.html (三中氏のサイト)

改訂履歴

2005.5 「数学の応用事例—比較言語学への応用」としてサイエンストピックスに掲載.
2005.10 「言語比較の数学的基礎」(縦書き) として『季刊邪馬台国』92 号に掲載. いくつかの誤植の修正と統計学的間違いを修正.
2006.12 文献表の追加と削除. タイトルを「言語比較の数学的基礎」に変更.