北星学園大学文学部北星論集第55巻第２号（通巻第67号）（2018年３月）・抜刷

【研究ノート】

言語年代学の基本公式の改良

吉　田　知　行

北星論集(文)　第 55 巻　第 2 号（通巻第67号）　March 2018

研究ノート

言語年代学の基本公式の改良

吉　田　知　行

目次

[要旨]
ひとつの言語の1000 年あたりの単語の残存率を r とする。A, B 両言語が分岐して T 千年たったあとの単語の一致率を c とするとき，言語年代学におけるスワデシュによる基本公式は次で与えられる。

$$T = \frac{\log c}{2 \log r}$$

これによって両言語の分岐年代 T を推定できる。しかし言語年代学の考えについては当初から多くの疑念や反論があった。反対に支持する側からは，基本公式のさまざまな改良が試みられてきた。このノートでは，確率過程の立場から，言語の変化の様子を探り，さらに基本公式の改良案として次の公式を提唱する。

$$T = \log\left(\frac{c-\mu_0}{1-\mu_0}\right) \Big/ 2\log\left(\frac{r-\mu_0}{1-\mu_0}\right)$$

ここで，μ_0 は，偶然による一致率で，一致反復率とよばれる。

1　言語年代学のはじまり

統計やコンピュターを使って言語を研究する学問領域を計量言語学という。スワデシュが 1950 年代初めに提唱した言語年代学 (glottochronology) はそのはしりと言える。『言語学大辞典』第 6 巻術語編によると，スワデシュは，考古学で遺物の年代を算定するための炭素 14 法に着目してこの方法を考案したという。

その前提となったのは，基礎語彙の 1000 年あたりの残存率がどの言語でもほぼ一定という事実である。スワデシュは，借用がおきにくいとされる 215 項目（別に 200 項目や 100 項目）の語彙リストを用意した。13 の言語について千年単位の残存率を計算した結果，千年単位の平均残存率として，$r = 0.8048 \pm 0.176$ が得られた。この数値はリーズ (R/B.Lees) による。

また 100 項目リストによるならば，1000 年あたりの平均残存率は 0.854 となる。

例えば，千年で約 8 割の単語が残るとすれば，1 万年後には，約 1 割の単語しか残らないことになる。それでも，200 語の日本語語彙の中には，1万年前の縄文時代の痕跡をとどめている単語が 20 語程度存在することになる。

千年あたりの平均残存率を r とする。200 語の場合は $r = 0.81$，100 語の場合は $r = 0.854$ とする。ふたつの言語が分岐して t(千) 年経ったとすると，

キーワード：言語年代学の基本公式，スワデシュ，確率過程，日本語の起源

基礎語彙の一致率は

$$c = r^t \times r^t = r^{2t}$$

となる。両辺の対数を取って t を求めると

$$t = \frac{\log c}{2 \log r}$$

となる。これがスワデシュが与えた言語年代学の基本公式である。

2 言語年代学への批判

言語年代学が適用できるためには，以下の前提条件がなり立っている必要がある（安本(1995)）。

仮定 I（同系性）2 つの言語 A,B は，同系である。
仮定 II（時間的恒常性）ある言語の 1000 年あたりの基礎語彙の残存率 r は，いつの時代でもほぼ一定である。
仮定 III（恒常性）1000 年あたりの残存率は，A, B 両言語でほぼ同じである。
仮定 IV（独立性）A, B 両言語は，分裂後基礎語彙において没交渉であった。

しかしこれらの仮定は島嶼から厳しい批判にさらされた。

言語年代学には当初からさまざまな問題が指摘されている。『言語学大辞典』の関連項目をまとめておく。

1. 誤差の大きさ。言語年代学の適用範囲は，分裂してから 3000 年以上経過していると，分岐年代の推定値はほとんど信用できないとされる。分岐年代が 1000 年前という推定でさえ，誤差は最大 500 年の可能性がある。
2. 仮定 I:同系の証明の困難。すなわち。仮定 I の比較言語学による証明が困難である。これは単語同士でも同じである。
3. 仮定 IV 独立性への疑い。そのため分岐年代が新しく出る。言語年代学によれば，日本語京都方言と沖縄の方言は分岐して約 1000 年と推定されるが，この数値は明らかに過小で，実際の分岐年代はもっと古い。分裂後のふたつの言語は，完全とは言わないまでもある程度交流が続いていたと考えるのが自然である。
4. 仮定 III:r の恒常性への疑問。千年あたりの単語の残存率は，ほぼ 0.81 であるとされてきた。しかしこの値はほとんどが印欧語族の言語から得られたもので，その数値にはすでにバイアスがかかっている。アイルランド語では 0.90 以上，エスキモー語では 0.10 未満と実際の言語では大きく違いがあり，一定しない。ただし，千年あたりの残存率 r が 0.1 であっても，100 年あたりの残存率は 0.8 となる。この程度なら，親子の意思の疎通に問題ないであろう。
5. 基礎語彙は安定していない。文化人類学の専門家によると，文化的に中立な語彙などないし，基礎語彙も借用・禁忌・比喩などによって置き換わる。

このように，言語年代学には多くの問題点があり，言語学者からの評判は芳しいものでなかった。あまり指摘されていないようだが，次の問題点もある。

6. 確率過程のとの矛盾。スワデシュの公式のままでは，$t \to \infty$ のとき，一致数 $c = r^{2t}$ は 0 に収束する。しかし，確率過程の理論によれば，この極限値は，偶然による一致率（言語や一致の基準によって異なるが 0,1 程度）のはずである。
7. 仮定 II:時間的恒常性への疑問。例えば，英語の場合，古英語が中英語に変わるとき単語から文法・音韻までが一斉に変化している。
   一部の言語集団だけが生き残った場合このような現象が起こると考えられる。生物集団の自然淘汰やボトルネック効果に似ている。

3 基本公式の改良と千単位年の導入

さまざまな批判があるが，ほとんどは解決可能である。これについては安本（1995）に詳しい。底には，言語の同系性の判定方法，基礎語彙の残存率のほぼ 0.81 であること，言語は変化しにくく，特に

基礎語彙は借用語の侵入などに対する免疫性を持つことを十分な根拠を上げて説明している。これについてはこれ以上述べない。ただ基本公式の改良については触れておきたい。

割と早い時期に日本に言語年代学を紹介したのは，著名な言語学者の服部四郎であった。服部は単なる紹介だけでなく，スワデシュの公式の改良を発表している。

$$c = r^{1.4t}, \quad t = \frac{\log c}{1.4 \log r}$$

この公式は，ふたつの言語が分裂後もある程度の交渉を保っていた場合に使える。つまり基本仮定 IV はなくても良い。例えば，日本語と琉球語の分岐年代は，1000 年だったのが，2/1.4 倍されて，今から 1400 年前となる。それほどの不自然さは感じられない。

ふたつの言語 A, B で，1000 年あたりの基礎語彙の残存率 r_A, r_B が等しくない場合は，樺島の公式がある

$$t = \frac{\log c}{\log rA + \log r_B}$$

この公式を使うなら，基本仮定 III はなくても良い。

ドブソンら 4 人の数学者が，言語年代学への批判における数学的基礎の誤りを整理し，鋭く反批判した。彼らは次の様に述べている。

> ある特定の語彙統計学のモデルが，ある点で指示できないことが示されても，そのモデルを完全にすててしまうよりは，それを改良し，修正するのがより当をえているであろう。

安本は，「言語年代学に対するまずは妥当な見解であるように思われる」としている。

なお追加しておきたことがある。言語年代学でふたつの言語の同系性の仮定は不要と思う。それは，スワデシュによる言語年代学の基本公式を 2 言語の基礎語彙による「距離」(情報理論のハミング距離) を「時間的距離」に換算する公式と考えることである。必要なら，服部や樺島の公式のようにさらに修正を加えればよい。この場合，基礎語彙の 8 割が残るのに要する年数として**千単位年**の概念を提唱したい。これだと 5 千単位年後の残存率は 0.5^5 となり，3 分の 1 の単語が残ることになる。1 万単位年後だと 1 割が残ることになる。スワデシュの公式によれば，千単位年と現実の 1000 年は，ほぼ一致する。

日本語と琉球語の分岐年数はほぼ千単位年だが，服部の公式によれば，それを 2/1.4 倍することによって 1400 年となる。

例. 服部によれば，日本語京都方言と朝鮮語京城方言は，93 項目中 10 から 18 語が同源にさかのぼるという。一致数 $c = 10/93 \sim 18/93$ である。$r = 0.8$ として分岐年数を計算すると

$$t = \frac{\log(18/93)}{2 \log 0.8} = 3680 \sim \frac{\log(10/93)}{2 \log 0.8} = 4997$$

すなわち 4 千から 5 千単位年前に分岐したことになる。$r = 0.81$ とすれば，3.897 ～ 5.291 となり，5 パーセントほど古くなる。基礎 100 語の場合の残存率 $r = 0.854$ を使うなら，5203 ～ 7065 年前に分岐したことになる。

また服部の公式によれば

$$t = \frac{\log(18/93)}{1.4 \log 0.81} = 5567 \sim \frac{\log(10/93)}{1.4 \log 0.81} = 7559.$$

$r = 0.854$ として服部の公式を使えば，$t = 7432 \sim 10903$ となる。

これらの数値，特に基礎 100 語用の残存率と服部の公式を使うと 7 千年を超える古い年代が出る。両言語がたとえ同系であったとしても，これほど古い年代が出ると，従来の比較言語学の方法では，同系かどうかは証明できないし，音韻対応の法則を見出すのも不可能と言わざるを得ない。

4　言語と音韻対応表の数学的解釈

すでに述べたように，言語年代学には多くの問題点が指摘されてきた。しかしその多くはすでに解決していると筆者は考えている。しかし残っている課題もある。その一つが，本稿の主題である確率過程の理論との矛盾の解消である。それには基本公式の改良が必要になる。やや数学的な議論をしなければ

ならない。議論を簡単にするために，単語は語頭の音（あるいは語頭の子音や語頭文字）だけを考える。つまり語頭音の時間的変化だけを考える。したがって，ふたつの単語の類似は語頭音の一致として定義する。相当の抽象化であり，切り捨てだが，これでうまくゆけばは話しが簡単になる；

以下では，吉田の論文 (2017) をもとに，数学的な用語と記号を準備する。$N = \{1, 2, \cdots, n\}$ を基礎語彙の項目番号とする。普通は $n = 100$ とか $n = 200$ を取る。$\mathcal{L}$ を音の集合とする。似た音はまとめておく。$\mathcal{L}$ に属する音は，$\lambda, \mu, \ldots$ のようにギリシア小文字で表す。そうすると，ある言語 A には n 個の単語からなる基礎語彙があり，i 番目の単語には語頭音 $f_A(i)$ が付随している。写像 $f_A : N \to$ のことを**語頭音写像**という。以下，$f_A(i)$ や f_A は，単に $f(i)$ とか f と書く。

ここでは一致の判定を語頭音の一致だけで判定するので，言語 A の基礎語彙の語頭音以外の情報は捨てて考える。そうなると言語とは，単なる写像 $[f : N \to \mathcal{L}]$ のである。ただし，ほかの言語 $f' : N' \to \mathcal{L}'$ であっても，全単射 $\sigma : N \to N'$ があって，$f = f' \circ \sigma$ のとき，ふたつの写像 f と f' は同値であると言い $f \cong f'$ と書く。これは単に項目番号の呼び名を取り替えているだけである。

結局，言語とは f の同値類 $[f]$ に過ぎない。このような写像の同型類 $[f]$ を 1 元データセットともいう

言語 $[f : N \to \mathcal{L}]$ の**音分布表**とは，$\mathcal{L}$ で番号づけられた行列

$$X[f] = (a_\lambda)_{\lambda \in \mathcal{L}}, \quad a_\lambda = |f^{-1}(\lambda)|$$

のことである。吉田 (2017) では，tab$[f]$ と書いている。

他にも言語 B があって，その語頭音写像を $g = f_B : N \to \mathcal{L}$ する。その音分布表を

$$X[g] = (b_\mu)_{\mu \in \mathcal{L}}, \quad b_\lambda = |g^{-1}(\mu)|$$

とする。さらに，**音韻対応表**を，次の $\mathcal{L} \times \mathcal{L}$ 型の長方形行列で定義する：

$$X[f, g] = (x_{\lambda,\mu})_{\mathcal{L} \times \mathcal{L}}$$

これは，$\mathcal{L} \times \mathcal{M}$ 型分割表にほかならない。

$f, g : N \to \mathcal{L}$ に対し，ふたつの言語 $[f], [g]$ の**一致数**を次で定義する

$$x[f, g] = \sharp\{i \in N \mid f(i) = g(i)\}$$

これは音韻対応表の対角和（トレース）である。

$$x[f, g] = \mathrm{Tr}(X[f, g]))$$

S_N を N 上の対称群とする。このとき，$g \circ \tau$ は，言語 B の基礎単語をランダムに並べ換えたものである。したがって $x[f, g \circ \tau]$ は偶然の一致数となる。すなわち**偶然の一致数**とは，

$$x(\tau) = x[f, g \circ \tau], \quad \tau \in S_N$$

のことである。サイズが大きすぎて（$n = 200$ で 200!，そのようなデータを扱うには特別の方法が必要である。よく使われるのは MCMC(マルコフ連鎖モンテカルロ) 法である。しかしカイ二乗統計量とは違って，今の場合は，一致数という線形性を持つ統計量なので，厳密な評価が可能である（吉田 2017)。

偶然の一致数の平均は次で定義される；

$$m[f, g] = \mathbf{E}_\tau[x[f, g \circ \tau]] = \frac{1}{|S_N|} \sum_{\tau \in S_N} x[f, g \circ \tau],$$

一般に S_N 上の関数 φ について，S_N 上の平均を $\mathbf{E}_\tau[\varphi(\tau)]$ であらわす。

定理 1(平均値公式)。$X[f] = (a_\lambda)_\lambda$，$X[g] = (b_\mu)_\mu$ とする。このとき以下がなり立つ：

(1) $m[f, g] = \dfrac{1}{n} \sum_{\lambda \in \mathcal{L}} a_\lambda b_\lambda$ である。

(2) $(1/n)m[f, g] - 1$ は，$X[f]$ と $X[g]$ の共分散である。

(3) 特に $g = f$ の場合，$(1/n)m[f, f]$ は偶然の一致率に等しい。

注. $(1/n)m[f, f] = \sum_{\lambda \in \mathcal{L}} (a_\lambda/n)^2$

を暗号理論ではの**一致反復率**という（フリードマン 1922)。文字の並べ替えによらず，各言語に特有の値を取るため，暗号解読に使われた。

系2(組合せモーメントの公式)

$$\mathbf{E}_\tau\binom{x(\tau)}{t} = \frac{(n-t)!}{n!}\sum_{\Sigma t_\lambda = t}\prod_\lambda \binom{a_\lambda}{t_\lambda}\binom{b_\lambda}{t_\lambda} t_\lambda!$$

特に，偶然による一致数の分布は両言語の音分布表 $X[f], X[g]$ だけで決まる。

一致数 $x[f,g]$ に関する P-値の正確な確率の計算方法は吉田 (2017) にある。

5　確率過程としての言語年代学

ある言語 $[f]$ の t(千年) 後の言語を $f^{(t)}$ とする。特に千年後の $[f]$ を $[f']$ で表す。以下簡単のため，$n = 200$ で考える。スワデシュのモデルの前提条件 (2 節仮定 I〜IV) のうち数学的な部分を単純化し，数式で表すと次のようになる。等号は「ほぼ等しい」あるいは「漸近的に等しい」を意味する。

スワデシュのモデル.

仮定 II. どんな言語 $[f]$ についても，$x[f^t, f^{t+1}]/n = r(= 0.81)$.

仮定 III. どんな言語 $[f]$, $[g]$ についても，$x[f,f']/n = x[g,g']/n = r$).

仮定 I, IV は後述する。

まずスワデシュが考えたように，ある言語 $[f]$ の経年変化を考える。一致数 $x[f, f^{(t)}]$ の時間 t に関する連続性により，

$$x[f, f^{(s+t)}]/n = x[f, f^{(s)}]/n \cdot x[f^{(s)}, (f^{(s)})^{(t)}]$$
$$x[f, f^{(t)}]/n = r^t$$

を得る。したがって，無限時間経過後の単語の残存率は

$$\lim_{t\infty} x[f, f^{(t)}] = x[f, \lim_{t\infty} f^{(t)}] = 0$$

となる。ここで，

$$f^{(\infty)} = \lim_{t\infty} f^{(t)}$$

他方，$[f^{(\infty)}]$ の音の分布を $(a_\lambda^{(\infty)})$ とすれば，

$$x[f, \lim_{t\infty} f^{(t)}] = \sum_\lambda a_\lambda a_\lambda^{(\infty)} > 0$$

なお，$[f]$ と $[f']$ の音分布が同じなら，

$$x[f, \lim_{t\infty} f^{(t)}]/n = x[f,f]/n > 0$$

(一致反復率) である。

つまり語頭音の残存率 $x[f, f^{(t)}]/n$ は長い時間の経過後に偶然の一致率に近づく。しかしスワデシュの前提条件のもとでは，0 に近づく。この矛盾はスワデシュの仮定 II, III のどちらかに問題があることを意味する。

6　新しいモーメント公式

前節と同じ記号を使う。

さらに　次の条件を仮定する。

比較言語学の仮定：言語 A において，i 番目の単語 (の語頭音) $f(i)$ が $f'(i)$ に変化する確率は，i には依存せず，音 $(f(i), f'(i)$ だけに依存する。

比較言語学では，ある音は，別の音に一斉に変化すると考えるので，この仮定はそれほど不自然でない。

数学的には，千年後に音 $\lambda \in \mathcal{L}$ が音 $\mu \in \mathcal{L}$ に変わる確率を $p_{\lambda,\mu}$ とすれば，$[f]$ が千年後に $[g]$ に変わる確率は

$$P_{f,g} = \prod_{i\in N} p_{f(i),g(i)}$$

一致数について，

$$x[f,g] = \sum_{i\in N} \delta(f(i), g(i))$$

(δ はクロネッカーのデルタ)。

定理. 千年後の残存率の積率母関数について，

$$F(u) = \prod_\lambda (1 + p_{\lambda,\lambda}(1-u))^{a_\lambda}$$

(証明)

$$\begin{aligned}F(u) &:= \sum_{g:N\to\mathcal{L}} P_{f,g}u^{x(f,g)}\\ &= \sum_{g:N\to\mathcal{L}}\prod_i \left(p_{f(i),g(i)}u^{\delta(f(i),g(i))}\right)\\ &= \prod_i\sum_\mu p_{f(i),\mu}u^{\delta(i),\mu)}\\ &= \sum_i\left(p_{f(i),f(i)}u+\sum_{\mu\neq f(i)}p_{f(i),\mu}\right)\\ &= \sum_i\left(p_{f(i),f(i)}u+1-p_{f(i),f(i)}\right)\end{aligned}$$

となる。結局

$$\begin{aligned}F(u) &= \prod_\lambda\left(p_{\lambda,\lambda}u+1-p_{\lambda,\lambda}\right)\\ &= \prod_\lambda\left(1+p_{\lambda,\lambda}(1-u)\right)^{a_\lambda}\end{aligned}$$

系 1. 千年後の残存数 $x[f,f']$ の平均について

$$m'[f]=\mathbf{E}_{f'}[x[f,f']=\sum_\lambda a_\lambda p_{\lambda,\lambda}$$

これは，積率母関数 $f(u)$ の $u=1$ での微分係数の計算から得られる。

系 2. 分散について

$$\mathbf{V}_{f'}[x[f,f']]=m'[f](m'[f]+1)$$

例. スワデシュの場合，$p_{\lambda,\lambda}=r=0.81$ であった。したがって

$$F(u)=(1+r(u-1))^n=\sum_{k=0}^n\binom{n}{k}(1-r)^{n-k}r^ku^k$$

どのような遷移確率行列 P を使うにせよ，積率母関数の公式は，結局スワデシュの条件 $(p_{\lambda,\lambda}=r0.81)$ に帰着される。ただ，基本公式だけは改良する必要がある。

7　改良基本公式

前節の記号を使う。さらに，ふたつの言語 $[f],[g]$ の t(千年) 後の一致数 (偶然による一致数) を $x_0(t)$ $(m(t))$ とする。

定理. 両言語の独立性を仮定する。スワデシュの条件 $p_{\lambda,\lambda}=r$ のもとで，

$$\frac{x_0(t)/n-\mu_0)}{1-\mu_0)}=\left(\frac{r-\mu_0}{1-\mu_0}\right)^{2t}$$

特に

$$\lim_{t\to\infty}x_0(t)=m$$

(確率収束)

なお，ここでの仮定のもとで，言語の音分布は平均的に変化しない。すなわち $m(t)=m(0)=m$ で置き換えられる。

これで，はじめにあげたスワデシュの基本公式のおかしな点は解決した。

実例については，前論文 吉田 (2017) 参照。服部の公式に習って基本公式の 2 を 1.4 にするとより納得のいく分岐年代が得られる。

8　日本語の起源は解決可能か？

急速な進展を見せている分子人類学 (DNA 人類学) と，日本考古学の成果を合わせると，日本人の起源が次第に見えてくる。日本人の東アジアへの到達時期，日本列島への渡来時期と拡散の様子を見ると，日本語の起源には強い制約がかかる。ただし，日本列島へのヒトの流入ルートや時期については研究者によってかなりの違いがある。

斉藤成也「日本列島人の歴史」(岩波ジュニア新書)2015
篠田謙一「DNA で語る日本人起源論」

まず 7 万年ほど前に，インドネシアのトバ火山が巨大噴火を起こし，ヒトの人口が数千から 1 万人程度に急減少した。ヒトの言語の起源は 7 万年以上前には遡れない。ヒトはサウジアラビア南岸からイラン南部に到達した。その後三方向に分かれ，東アジアに到達したのは 5 万年ほど前と言われている。

日本列島に到達したのは，3 万 8 千年以前 (朝鮮半島から対馬ルート)，2 万 6 千年前 (サハリンから北海道の陸上ルート)，3 万 5 千年前 (沖縄ルート) の三方向からである。

4 万年前には東アジアには Y-染色体ハプログループの D 系統の人たちが広く分布しており，日本列島に渡ってきたのは彼らである。そうなると，彼らの使っていた言語もある程度近い言語であったと考えられる。

しかし，同じ D 系統の末裔である本土日本人と北海道のアイヌ人の言語はまったく別の言語である。4 万年もたつと単語レベルでは，同系の判定が不可能なほど異なっている。

結局日本語の起源を比較言語学だけで研究することはあまり有益と思えない。これは改良した基本公式からも分かる。

なお，定理に挙げた公式のかたについている $2t$ を服部四郎のアイデアにしたがって 1.4 とすると，日本語と朝鮮語の分岐年代はほぼ 6146 年前となる。

参考文献

・吉田知行『言語間の距離とシフト法』数理科学(1984/12)
・吉田知行『分割表の一致率検定と Fisher の正確確率法』北星論集，北星学園大学経済学部(2017)
・斉籐成也『日本列島人の歴史』岩波ジュニア新書(2015)
・安本美典『言語の科学―日本語の起源をたずねる』朝倉書店(1995)