Vol.2014-MPS-100 No.6
2014/9/25

基礎語彙に基づく言語系統樹の推定手法に関する調査

呉　靭[1,a)]　松浦　佑哉[2,b)]　松野　浩嗣[3,2,c)]

概要：近年，分子系統学の手法を言語学に応用し，言語系統樹を推定するアプローチが注目を集めている。本発表では，基礎語彙に基づいた分子系統学の手法による言語系統樹の推定方法を調査し報告する．これを踏まえ，我々の研究展開の位置づけおよび方向性などを示す．

1. はじめに

この世界で話されている言語の数は 6,000 を超えると言われている．言語には語族（language family）があり，同じ語族に属する言語は共通の祖語（先祖にあたる言語，proto-language）を共有し，変化してきたとされている．

言語と言語の親近関係をしらべるため，基礎語彙が用いられてきた．言語と言語との基礎語彙を比較し，言語間の近さの度合いを数字ではかり，統計的手法により同系（同じ語族に含まれる）の言語であることを確認する．この方法は，言語の系統分類に効果をもたらした [1], [2]．近年は分子系統学の応用により，基礎語彙を用いた新しい方法論による言語間の関係の推定に関する研究が，再び注目を浴びている [3][4][5].

本研究では，言語間の親近関係をしらべるための言語学における計量的研究の基礎とともに，近年のこれらの研究における分子系統学の応用を調査し，我々の今後の研究展開の方向性などを探ることが狙いである．

本発表では，以下 **2.** では，まず基礎語彙と基礎語彙表について述べ，基礎語彙表を用いた統計的手法による言語間の親近関係の推定に説明する．**3.** では，分子系統学を応用した基礎語彙に基づく言語系統樹生成の新しいアプローチとして，Brown らの世界諸言語の自動分類法（中核となるのが Automated Similarity Judgment Program (ASJP) で，以降 ASJP と呼ぶ）[5] を紹介する．最後の **4.** では，今後の展望について述べ，本稿をまとめる。

[1] 山口短期大学情報メディア学科
〒 747-1232 山口県防府市台道 1346-2
[2] 山口大学理学部
〒 753-8512 山口県山口市吉田 1677-1
[3] 山口大学大学院理工学研究科
〒 753-8512 山口県山口市吉田 1677-1
[a)] wu@yamaguchi-jc.ac.jp
[b)] r050de@yamaguchi-u.ac.jp
[c)] matsuno@sci.yamaguchi-u.ac.jp

2. 統計的手法による言語間の親縁関係の推定

2.1 基礎語彙と基礎語彙表

A，B の 2 つの言語は祖語 P から分裂し，独立に発展した言語とする．言語 A，B は，最初は祖語 P と同じ基礎語彙をもっていたが，だんだんほかのことばにとってかわられた．この言語 A，B の親縁関係の推定には，現在の 2 つの言語の基礎語彙において，CVC（子音-母音-子音）の一致などの客観的な基準にてらし，きわめて高い一致が認められたならば，この 2 つの言語が同じ祖語から分かれた可能性が高いと考えられている [2].

ここでの基礎語彙とは，身近な親族名称や身体部位名，人間としての基本的動作や振る舞いなどを表現する，基礎的な，非文化的語彙で，かならずしも使用頻度の大きい語とは一致しない．例として，身体語「手」「目」「耳」，天体に関する語（「月」「日」「雲」など），色彩語（「黒」「白」など）が挙げられる．「コーヒー」や「サッカー」や「ハンバーガー」のような言葉は基礎語彙ではない．基礎語彙は，文化的な語彙にくらべ，はるかに借用されにくく，借用語の侵入に対するかなりな免疫性をもち，それぞれの言語は，他の言語からほぼ独立に，それぞれの基礎語彙を発展させる傾向の強いことが知られている．つまり，基礎語彙は時間に対する抵抗力が強く，古い時代の語がそれほど変化せずに残る傾向が強い [2].

基礎語彙などの摩滅のしかたに数学的モデルを設定し，共通の祖語をもつ 2 つの言語がいつ分裂したかについて研究を行ったアメリカの言語学者モリス・スワデシュが初めて基礎語彙表の作成に着想した．スワデシュの基礎語彙表には，初めは 215 ないし 200 の基礎語彙を含めたが，最終的には 100 項目に縮約している．基礎語彙の選定は，現存する言語とその古語とを比較し，変化していない語をえらぶという基準にしたがって行われた．これは通時的基準とよばれ，印欧語などの古い時代の文献が豊富な言語に用い

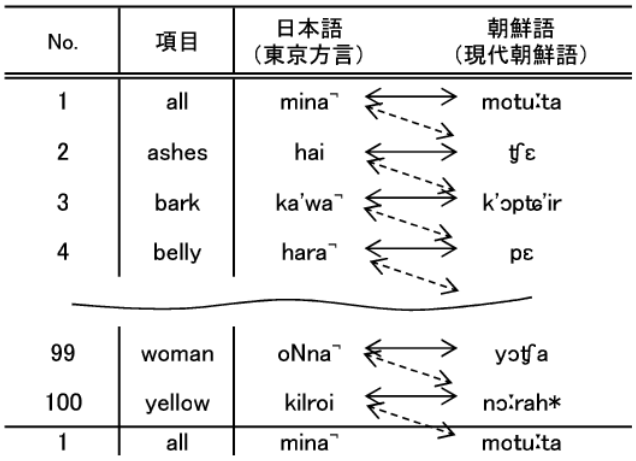

図 1 日本語（東京方言）と朝鮮語（現代朝鮮語）の基礎語彙表

られることが多い．その後，多くの言語学者によって多くの言語の基礎語彙表が作られたが，単語意味項目の選定は通時的基準によるばかりではない [2].

基礎語彙表の単語の表記は国際音声字母 (International Phonetic Alphabet=IPA) という音声記号が使われることが多い．各言語の正書法では文字と音声との間にずれがあり，かつ言語ごとに違っているのに対し，IPA を使えば，言語を問わず一貫した表記ができる．IPA は音と記号との間に 1 対 1 の対応関係が成り立つようにしている [6].

図 1 に日本語（東京方言）と朝鮮語（現代朝鮮語）の 100 項目基礎語彙表のイメージを示している．意味を表す「項目」の列は英語表記である．「No. 1」という 1 つ目の項目の意味「all」に対し，日本語が [mina¬] で，朝鮮語は [motu:ta] となっている．

2.2 基礎語彙表の単語の一致をしらべる方法

共通の祖語から分裂し，独立に発展した 2 つの言語は分裂してから時間が経てば経つほど，その間の相関は失われていくことが考えられる．2 つの言語が同じ祖語から分かれた，つまり同系であることを確認するために，基礎語彙が使われる．2 つの言語の基礎語彙が統計学的にきわめて高い一致が認められたならば，同系であるとみなしてよいと考えられている．

では，2 つの言語の基礎語彙表にある個々の単語が，どのような場合に「一致」しているかについて，判定する客観的な基準が必要であると思われる．

まず，ポイヤの方法 [2] がある．これは，語頭音，つまり「頭文字」だけをとりあげて考察するものである．例として，図 1 の「No. 1」の「all」の項目では，日本語と朝鮮語の両方がともに/m/で，一致しているが，「No. 2」の「ashes」の項目では，/h/と/ʧɛ/になっており，一致していない．これらの語頭の音の一致と不一致の数を数え，2 つの言語の基礎語彙表の単語の一致する度合いをはかるものである．

次に，ベンダーの CVC（子音-母音-子音）[2] によって比較する方法がある．CVC による比較の基準の一部は次のようになる．(1) 単語が短く，CVC の 3 つをそえなていない場合は，度外視する，(2) CVC がそなわっている長い単語の場合は，最初の CVC をしらべる，(3) まん中の母音が一致し，一方の子音が一致し，他方の子音が「一定の範囲」で違っている場合も，一致とみなす．ほかにも CVC の基準を緩めた CV（子音-母音）の基準がある．

基礎語彙表の単語の一致をしらべる基準として，上に述べたポイヤの方法やベンダーの方法により，2 つの言語の基礎語彙表の単語の一致度をしらべ，統計的検定を行う．それらの方法はそれぞれ語頭音検定法と CVC 検定法とよばれている．後の研究において，同じ基礎語彙データにこの 2 つの方法をそれぞれ適用したところ，一方は有意の一致を示すのに対し，他方は有意の一致を示さないという結果が表れた．

そこで，2 つの言語間の意味が対応していない単語についても比較し，その偶然による単語の一致度もしらべる，シフト法 [2] とよばれる方法がオズワルトによって考案あれた．意味が対応している単語を比較したときの一致数が，偶然による一致数よりずっと大きければ，その一致は偶然によるものではないという考えからである．

シフト法は語頭音検定法や CVC 検定法などにも適用できる．図 1 に語頭音検定法に適用した語頭音シフト法 [2] について，ひとつずつずらす様子を示している．語頭音シフト法を使い検定を行うには，まず 2 つの言語の意味が対応している単語の一致数を数える．これは粗点（x_0 と表記する）という．次に，意味項目をひとつずつずらして比較を行い，単語の一致数を数える．これを繰り返して行ったうえで，一致数の平均値 m，さらに標準偏差 s を求め，式 (1) によって偏差値 z を求める．

$$z = \frac{x_0 - m}{s} \tag{1}$$

偏差値 z が 1.65 以上であれば 5% 水準で有意，2.33 以上であれば 1% 水準で有意である．この z の値が有意であるとき，上で述べた一致は偶然によるものではないといえ，2 つの言語は同系で，親縁関係をもつと推定されることになる．

3. 言語系統樹生成の新しいアプローチ

ASJP では，245 言語について，100 項目のスワデシュ基礎語彙表を作成している．（個々の言語が 100 項目のすべてについて単語が全部埋められているわけではない）．コンピュータ処理をしやすくするため，音声記号の正書法として，アルファベットの大文字と小文字，数字，および記号を使った独自のものを用いている．母音の音声記号は 7 つのみにまとめられ，複数の母音が 1 つの音声記号に対応

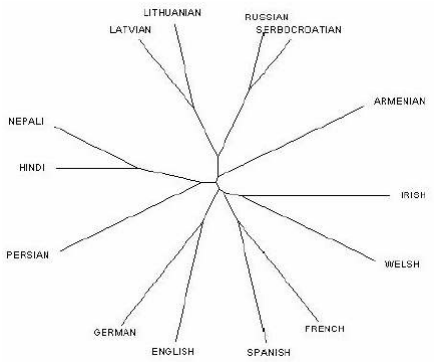

図 2 ASJP によって生成された Indo-European 語族の言語系統樹

INDO-IRANIAN:
IRANIAN: Persian
INDO-ARYAN: Hindi, Nepali
CELTIC: Irish, Welsh
ITALIC: French, Spanish
GERMANIC: English, German
ARMENIAN: Armenian
BALTIC: Latvian, Lithuanian
SLAVONIC: Russian, Serbo-Croatian

図 3 歴史言語学者による分類

するように変換されている．

基礎語彙表の単語の一致をしらべる基準は，基準 1 から基準 8 までの 8 つの条件のいずれかを満たしているときに，2 つの言語がその単語について類似しているとみなす．8 つの条件は，たとえば，基準 1 は次のようになっている．2 つの言語 A, B の基礎語彙の単語のそれぞれに C_1VC_2（C_1 と C_2 はそれぞれ子音，V は母音を指す）があり，$C_1^A = C_1^B$ かつ $C_2^A = C_2^B$ のときに，この意味の単語について言語 A と言語 B は類似しており，一致度合いの数に 1 を増やす．例として，buk vs. bek.

言語間の距離は次のようにして計算されている．(1) 245 言語の個々について，言語間の語彙類似率（The Lexical Similarity Percentage，LSP と表記する) をしらべる．前に述べたように，個々の言語が 100 項目のすべてについて単語が全部埋められているわけではない．言語 A と言語 B が，たとえば，100 項目のうち 95 項目について単語が記載されていて，そのなかの 30 項目について上に述べた基準にしたがって類似していると判定された場合，$LSP = 30/95 \times 100 = 31.6$ となる．(2) 音節類似率 (Phonological Similarity Percentage，PSP と表記する) をしらべる．PSP は **2.2** で述べたオズワルトのシフト法によるもので，意味が対応していない単語について，2 つの言語が偶然による類似と判定される単語の割合である．(3) 差引類似率 (Subtracted Similarity Percentage，SSP と表記する) を計算する．SSP は $SSP = LSP - PSP$ によって計算され，言語間の距離とされている．

245 言語について，任意の 2 つの言語ペアの SSP 値が計算され，245×245 のサイズの SSP 行列が作成された．また，SSP 行列を距離行列として，近隣結合法を用いて，言語系統樹を生成している．

文献 [5] では，Mayan, Mixe-Zoque, Otomanguean, Huitotoan-Ocaina, Tacanan, Chocoan, Muskogean, Indo-European および Austro-Asiatic の一部の言語について，ASJP によって生成された言語系統樹を提示し，伝統的な方法によって作られた言語の分類と大体一致したことを示している．図 2（ 文献 [5] より引用）は Indo-European 語族の 14 言語について ASJP によって生成された言語系統樹のイメージを示している．図 3（ 文献 [5] より引用）は歴史言語学者による分類である．

4. おわりに

言語間の親縁関係や言語の系統分類に関し，語頭音シフト法など基礎語彙に基づく言語学分野の手法とともに，分子系統学の手法を応用した新しいアプローチとしての ASJP について紹介した．

ASJP では，独自の音声記号の正書法を用いており，複数の母音が 1 つの音声記号に対応するように変換されている．また，基礎語彙表の単語の一致をしらべる基準として課している条件の妥当性も検討する余地がある．さらに，系統樹の推定方法として近隣結合法が用いられているが，系統樹の推定方法はほかにも最大節約法、最尤法、ベイズ法など多くあり，言語系統樹の生成に最も適している方法について検討が必要だと考える．今後は，これらの点について検討を行っていきたい．

謝辞

本研究の一部は日本学術振興会科学研究科研費挑戦的萌芽研究 23650129 の助成を受けたものである．ここに記して感謝の意を表す．

参考文献

[1] 安本美典，野崎昭弘：言語の数理，筑摩書房 (1976).
[2] 安本美典：言語の科学，朝倉書店 (1995).
[3] R.D. Gray and Q.D. Atkinson: ”Language-tree divergence times support the Anatolian theory of Indo-European origin”, Nature, Vol. 426, No. 6965, pp. 435-439 (2003).
[4] R.D. Gray, A.J. Drummond, and S.J. Greenhill: Language phylogenies reveal expansion Pulses and Pauses in Pacific settlement, Science, Vol. 323, No. 5913, p. 479-483 (2009).
[5] BROWN, Cecil H., et al. Automated classification of the world’s languages: a description of the method and preliminary results, STUF-Language Typology and Universals Sprachtypologie und Universalienforschung, 61.4, pp. 285-308 (2008).
[6] 風間喜代三ほか：言語学第 2 版，東京大学出版会 (1993).