Vol.2012-CH-94 No.4
2012/5/26

揖斐川上流域の語彙に関する系統推定

小野原彩香†

本研究では，揖斐川上流域における基礎語彙の調査結果を元に系統推定を行い，徳山村の村落が他の集落と系統上で隔絶されることが確認できた．また，ランダムフォレストを用いて，集落ごとの特徴語彙を抽出した．

Phylogeny Estimation about Words in the Upstream Region of Ibi River

AYAKA ONOHARA†

In this study, using words in the upstream region of Ibi river, we estimated the phylogeny on dialects in those regions. We have confirmed that Tokuyama District is separated from others on the phylogenetic network. Moreover, we determined characteristic words in each colony by use of “Gini coefficients in random Forest”.

1. はじめに

元来語彙は音韻，文法などに較べて，変化しやすいものであるが，その中心的部分においては，比較的変化の速度がおそく，借用要素が入りにくいことは Morris Swadish にはじまり，服部[1]により発展させられた言語年代学(Glottochronology)における基礎語彙の諸研究において言及されている．しかしながら，言語年代学が，言語変化の速度を一定としたことや，比較する二言語が比較言語学的に同系かつ言語間に分岐後の接触が無いことを前提としたことによって，観察される言語現象と言語年代学の理論が合致しないという事態が数多くあった[2]．このため，基礎語彙から語彙の分出年代を求める研究は下火となった．しかしながら，基礎語彙を用いることが無効となった訳では無く，語彙統計学[3]や数理言語学的方法[4]の中で，基礎語彙を用いて，地域間の使用語彙の差や語彙の分岐についての考察が行われている．

本研究では，上記のトピックとして興味深い地域である徳山村及びその周辺地域の基礎語彙を用いて各集落における使用語彙の系統推定を行い，基礎語彙を用いた研究事例として方法論の有効性を示すことを目的とする．

2. データ

2.1 調査地域

徳山村の方言は，成立して以来何回ともなく中央―京都，或いは古くは奈良の言語より改新の波を被っているものと推定される多重層語的な方言であり，かつ内部より独自の変化を起こしているとされる[5]．

本研究では，山田[6][7]で報告されている基礎語彙に関するデータを用いた．山田の調査は，1976 年に岐阜県の揖斐

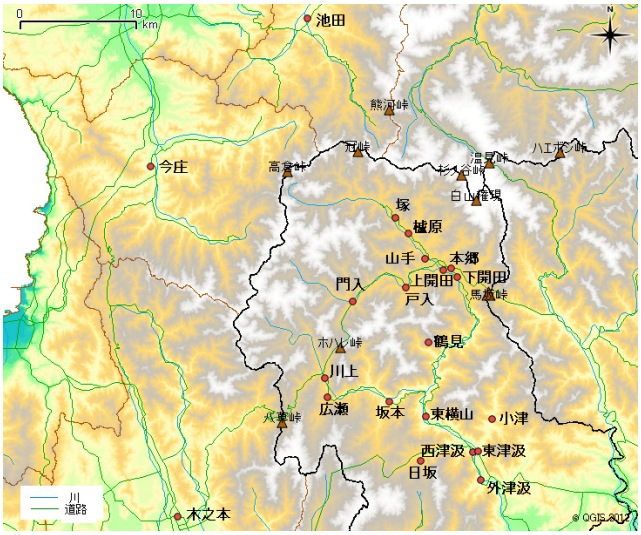

図 1　調査地点と主な峠

川上流で行われた後，1979 年に揖斐川上流と交流があった福井県側と滋賀県側の数地点で行われた．

対象地域は，岐阜県の西部に位置し，北を福井県に，西を滋賀県に接する．対象地域は，図 1 のように周囲を 1,200m 級の高山にかこまれ，その山狭の切りたった谷底を流れる揖斐川沿いに点在する部落を集めてできた四ヶ村（徳山村，藤橋村，久瀬村及び支流坂内川沿いの坂内村）と周辺部の滋賀県木之本町，福井県今庄町，池田町[a]である．

調査当時，同地域に入る道は，川沿いに整備され，徳山村の場合は，根尾村方面より馬坂峠越しの道路も出来ているが，明治時代までは特に徳山村では川沿いの道は危険に満ち，ところどころ寸断され，一般的には利用出来る状態ではなかった．往時わずかに開いていた小道は，一つは塚から冠峠越しに福井県越前地方に通ずるもの，一つは門入からホハレ峠越しに坂内村の川上へ，更に八草峠越しに滋

† 同志社大学大学院文化情報学研究科
Graduate School of Culture and Information Science, Doshisha University

a) 各地名は調査当時のもの．

表 1　対象地域

都道府県名	郡名・村名（当時）	集落名
岐阜県	徳山村	塚
岐阜県	徳山村	櫨原
岐阜県	徳山村	山手
岐阜県	徳山村	本郷
岐阜県	徳山村	下開田
岐阜県	徳山村	上開田
岐阜県	徳山村	戸入
岐阜県	徳山村	門入
岐阜県	藤橋村	鶴見
岐阜県	藤橋村	東横山
岐阜県	坂内村	川上
岐阜県	坂内村	広瀬
岐阜県	坂内村	坂本
岐阜県	久瀬村	西津汲
岐阜県	久瀬村	東津汲
岐阜県	久瀬村	小津
岐阜県	久瀬村	日坂
岐阜県	久瀬村	外津汲
福井県	南条郡	今庄町
福井県	今立郡	池田町
滋賀県	伊香郡	木之本町

賀県木之本方面に通ずるもの，及び徳山から馬坂峠越に根尾村に通ずるものの三本のみで，いずれも非常な困難と危険を伴っていた．下流部も徳山ほどではないにしろ集落間の交通はかなり難しかったようである.調査された地点は，徳山村 8 地点，藤橋村 2 地点，坂内村 3 地点，久瀬村 5 地点，木之本町 1 点，今庄町 1 点，池田町 1 点の計 21 地点であり，これは外津汲以北のほぼ全集落にあたる．調査地点の地点名と群名・村名および都道府県名との対応関係については表 1 に示す.

2.2 調査方法と被調査者

調査項目は，服部[1]の Japanese Dialects の 200 項目である．山田による資料の処理方法は，服部[8]に従っている．中核的形態素が東京方言あるいは京都方言と対応する場合には＋で，対応しない場合は－で，又対応関係が疑問の場合は○で，△は東京・京都方言の語に相当する語が二つあり，その一つのみが対応し，かつ代表形を決めることが難しい場合である.山田による集計結果の一例を図 2 で示す.山田で紹介されているのは，200 項目のすべてではなく，21 地点のうち 1 点でも東京・京都に対して－，△，○の関係のあるもの及び両者に対して中核的形態素の点ではすべて＋であっても，形式上興味のあるものの二つである．表 2 に該当項目と語例として東京方言，戸入方言を挙げる．インフォーマントについては，一地点で 1 名のこともあれば 2〜3 名の場合もあった.いずれも土地生え抜きの方々で，最年長者で 87 才，最年少者で 54 才，3 人を除いて 60 才以上である.

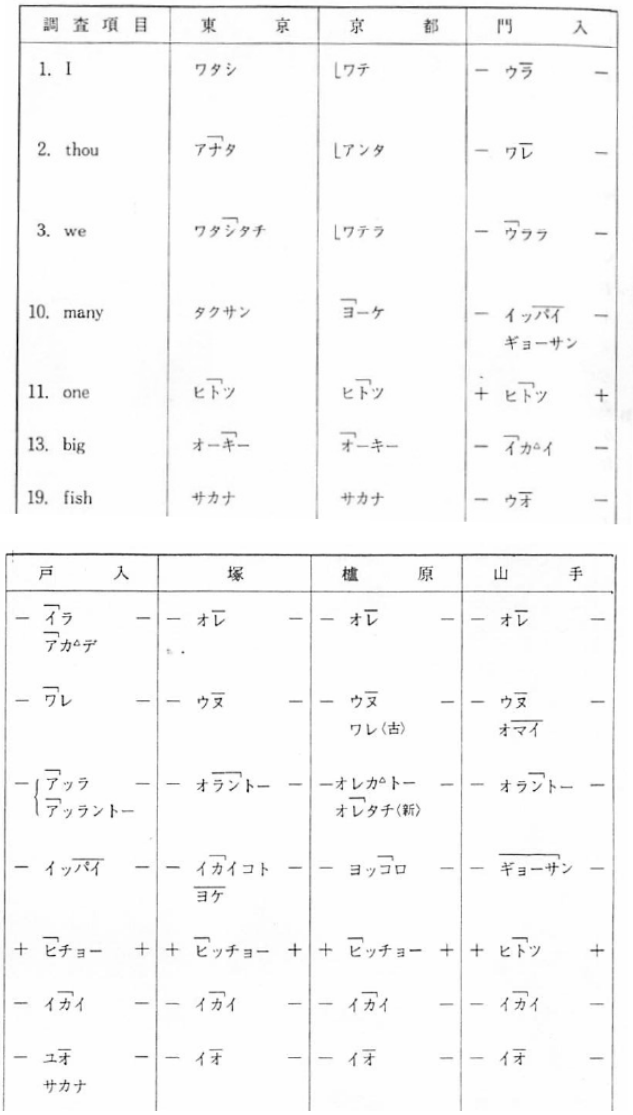

図 2　山田[6]の集計結果一例

表 2　分析対象語彙

No.	name	東京方言	戸入方言
1	I	ワタシ	イラ，アガデ
2	thou	アナタ	ワレ
3	we	ワタシタチ	アッラ，アッラントー
10	many	タクサン	イッパイ
11	one	ヒトツ	ヒチョー
13	big	オーキー	イカイ
19	fish	サカナ	ユオ，サカナ
32	grease	アブラミ	シロミ
35	tail	シッポ	オッポ
38	head	アタマ	カシラ
44	tongue	シタ	ベロ
49	belly	オナカ	ハラ
55	eat	タベル	クウ
56	bite	カミツク	クイツク

60	sleep	ネムル	ネル
65	walk	アルク	アリク
67	lie	ネテル	ネコダットル
68	sit	スワッテル	スワル，ジョーラカク，ツクバル
70	give	アゲル	ヤル
72	sun	オヒサマ	オヒーサン
92	right	ヨル	ヨル
96	new	アタラシイ	アタラシイ，サラ
97	good	イー	ヨイ
99	dry	カワイタ	イヤイタ
101	ye	アンタタチ	ワッラ，ワッラントー，イカデントー
102	he	アノヒト	アノモン
103	they	アノヒトタチ	アノモントー
113	few	スクナイ	チート
114	sky	ソラ	テン
116	fog	キリ	キリ
128	arm	ウデ	ウジェ
135	milk	オチチ	チッチ
143	mother	オカーサン	オッカ
144	father	オトーサン	オトッサン
145	husband	シュジン	オトッサン
146	wife	カナイ	イエノハー
150	freeze	コール	コール
158	thick	アツイ	アツイ
167	smooth	スベッコイ	ツルツル，スベコイ
169	correct	イー	ヨイ
176	throw	ナゲル	ナゲル
177	hit	ナグル	ナグル
187	smell	カグ	カグ
188	puke	ハク	アゲル，ハク
191	fear	コワガル	オソガガル
196	ripe	ジュクシタ	イロッダ
199	rope	ツナ	ホソビキ

3. 分析の手順

3.1 NeighborNet による系統推定

NeighborNet (NN)は Saitou & Nei[9]による Neighbor Joining (NJ)法を，ネットワークを許容するように，すなわち複数の樹形の可能性を表現できるよう改良したものである[10]．NJ および NN では，系統樹の枝の長さの総和が最小の樹形を選び，樹形を限定し比較を行う．本研究では，ネットワークの描画に SplitsTree4[11]を使用した．

本稿では，集落間の使用語彙に関しての系統樹を求めた．

3.2 対応分析

クロス集計票に適用する場合と，0, 1 の 2 値変数のデータ（多値の場合は 0, 1dummy 変数に展開）に適用する場合があり，定式化は同じだが意味合いが異なる．データ M=(m_{ij})が与えられたとき，ケースC_iと変数V_jとに数量化値x_i,y_jを与え，データ値m_{ij}（2 値変数の場合は 0,1)に対する評価関数 f(x,y)=$\Sigma\, m_{ij}x_iy_j$を，$\Sigma\,(p_ix_i)^2$=$\Sigma\left(q_jy_j\right)^2$=1 のもとで最大化する．ここで$p_i$=$\sum_j m_{ij}$，$q_j$=$\sum_i m_{ij}$つまりデータ$m_{ij}$の行和と列和である．つまり，x=($x_i$)と y=($y_j$)の重み付き相関を最大にするといっても良い．2 値変数の場合，よく似た反応パターンに，近接した数量化値が与えられることになるので，数量化の結果 x, y の昇順（あるいは降順）に，ケース，変数を並べ替えると，0, 1 のパターンが似た順に配列される．

なお，分析及び結果の提示には R Commander のプラグイン FactoMineR 中の Correspondence analysis(CA)を用いている．

本稿では，集落とその集落に特徴的な語彙の関係性を明らかにするために対応分析を行った．

3.3 クラスター分析

本稿では，対応分析によって別れた集落をクラスタリングしており，FactoMineR 中の CA の結果を受けて，クラスタリングを行った結果を示している．このアルゴリズムは階層的クラスタリングを用いており，距離にはユークリッド距離を，クラスタリングの方法にはウォード法を用いている[12].

3.4 ランダムフォレスト

ランダムフォレスト（RF : random Forest）法は，決して精度の高くない分類方法を複数組み合わせて精度を向上させ，分析に用いる方法である．

本稿では，系統樹によって分けられた地域を分類コードとして用いた．すなわち，集落を分けている要因になっている語彙が何であるかをランダムフォレストによって求めた．

3.5 多次元尺度構成法

多次元尺度構成法(MDS : Multi Dimensional Scaling)は，多次元データを低次元に縮約して示す統計的手法である．個体間の親近性の類似したものを近く，そうでないものを遠くに配置し，データの構造を考察する．

本稿では，random Forest で算出された集落間のデータの相違度を MDS で描画した．

4. 結果と考察

4.1 語彙から見る集落の類似性

図 3 は集落ごとの使用語彙のデータを用いて系統樹を描いた結果である．そして，図 4 が当該地域の地理的な図である．図 3 の系統樹の結果を図 4 と対応させて考えてみると，地理的に近接している地域は，系統上でも近接しており，地理的に離れている地域は，系統上でも離れた場所に位置することが分かる．また，系統上でまとまっている集

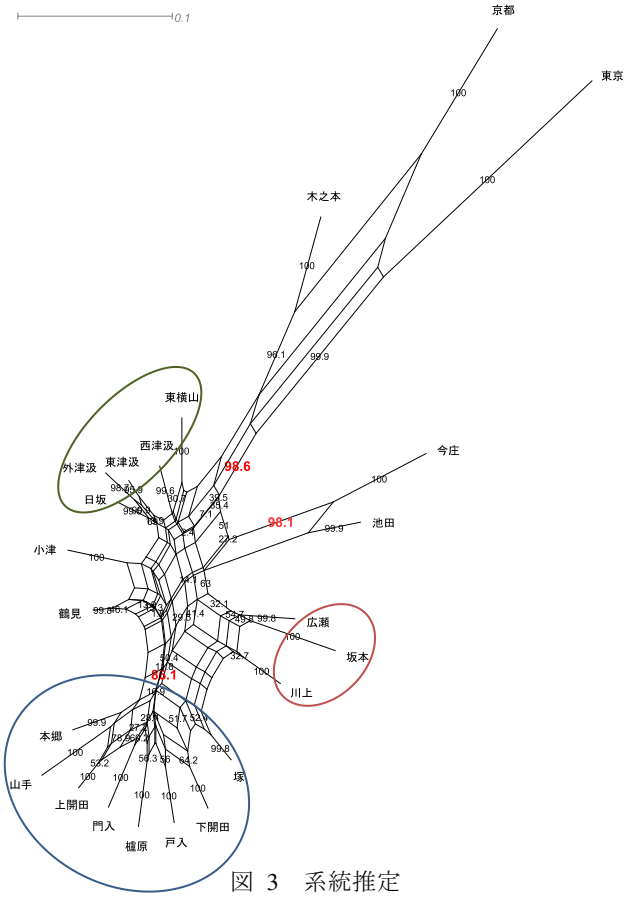

図 3　系統推定

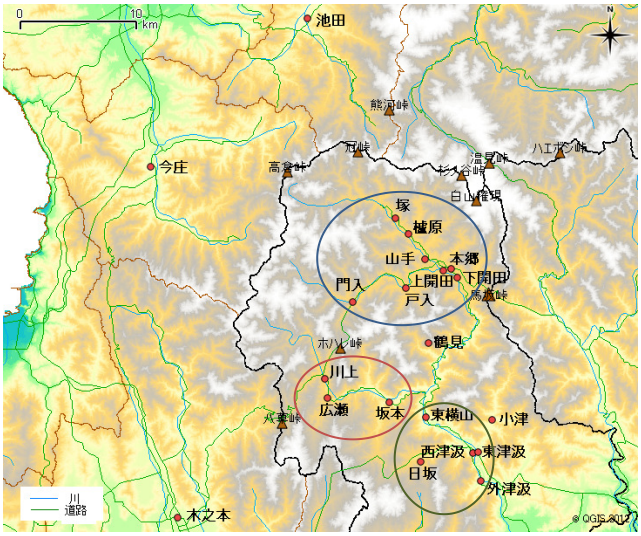

図 4　系統推定のまとまりを記入した地図

落単位の一部は，そのまま村の単位に相当することが分かる．系統上でまとまっている，山手，上開田，門入，櫨原，戸入，下開田，塚は徳山村である．また，川上，坂本，広瀬は坂内村である．

系統上の枝上にある数字は，系統樹の統計的な信頼性を調べるためにブートストラップ法を用いて導いた信頼度(%)の結果である．ここでは，1000 回の繰り返しによってブートストラップを行っている．図 3 の結果から，徳山村

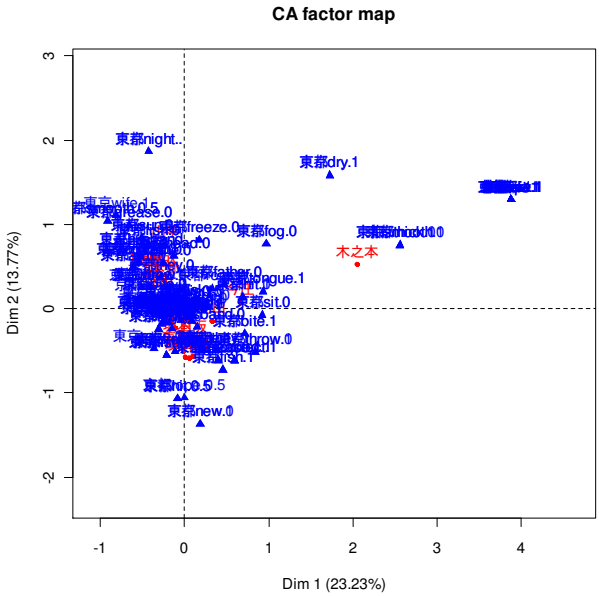

図 5　対応分析（東京，京都以外）

とそれ以外は，86.1%の信頼度で系統が分かれ，今庄，池田とそれ以外は，98.1%，東京，京都，木之本は 98.6%の信頼度で系統が分岐することが分かる．

図 5 は東京，京都以外の集落と集落に特徴のある語彙について対応分析を用いて示した図である． x 軸の正の方向には，I，we，sit，few，ripe が特徴語として現れ， x 軸の負の方向には，smooth，wife，grease．また， y 軸の正の方向には，I，we，sit，few，ripe，dry が，y 軸の負の方向には，new，hit，rope，fish，good，correct，grease が現れる．

また，図 6 は対応分析を行った際に算出される集落の得点を用いて主成分分析を行った結果を示した図である．図 3 の系統推定の結果と図 6 の集落のまとまりは，ほぼ一致することが見て取れる．図 7 は，対応分析を行った際に算出される集落の得点を用いてクラスター分析を行った結果を示している．この結果についても，図 3 での系統推定の結果とほぼ一致する．

4.2 集落に特有の語彙

図 8 は，図 3 で求めた系統推定の結果をもとに，地域を典型=(東京，京都，木之本)，徳山=(門入，戸入，塚，櫨原，山手，上開田，本郷，下開田)，坂内=(川上，広瀬，坂本) 藤橋・久瀬=(東横山，西津汲，東津汲，日坂，外津汲)，小津・鶴見=(小津，鶴見) ，周辺=(今庄，池田)とコード分けし，このコード分けに依存する語を random Forest によって求めた結果である．各回の決定木の数は 500 としてそれを 100 回繰り返し，Gini 係数の分布を求めた．

また，random Forest を行った際に，東京，京都間のみに違いが現れる語を省いた他，△，〇で表現されるデータが一集落でも含まれる語については，その語の列そのものを削除して分析を行った．

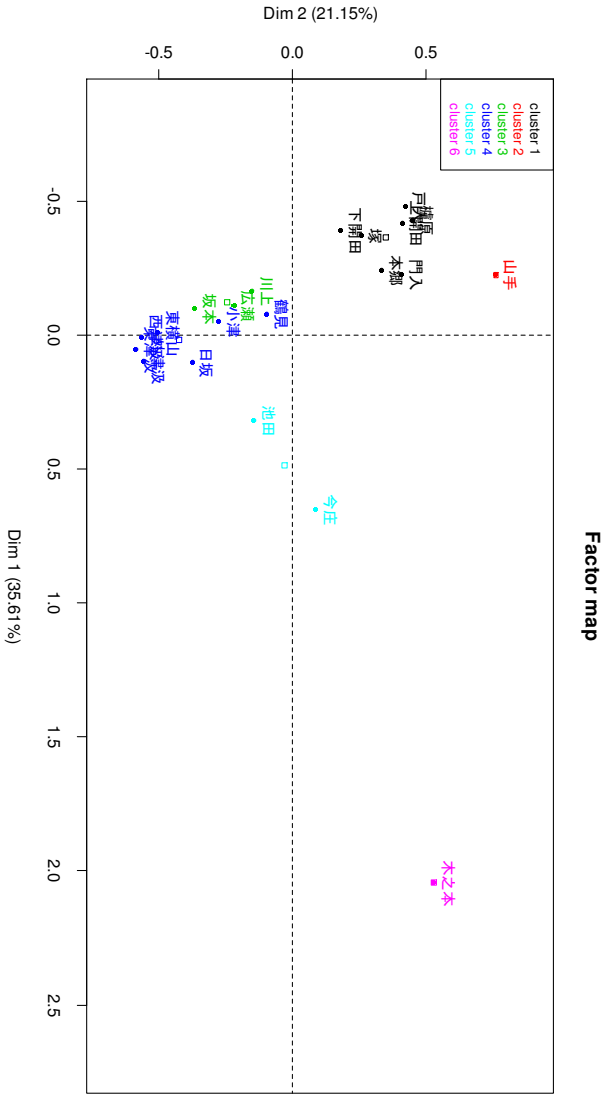

図 6　主成分分析（対応分析後）

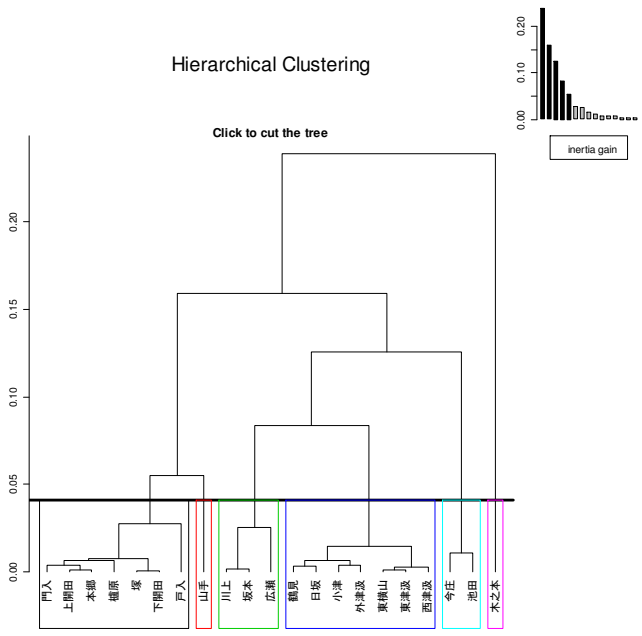

図 7　対応分析の得点を用いたクラスター分析

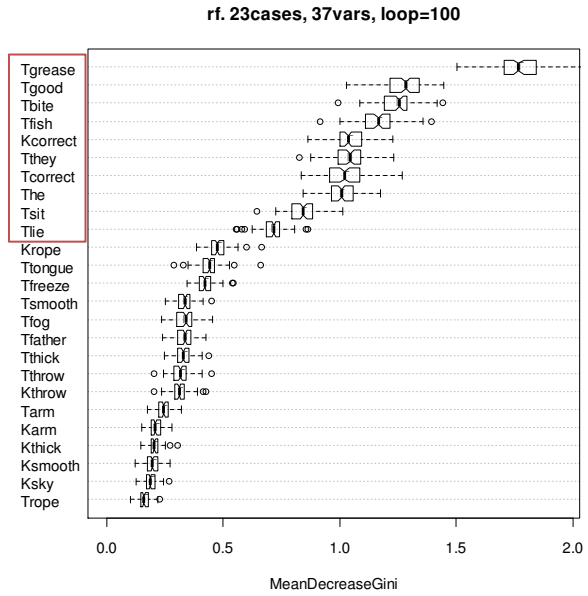

図 8　random Forest

表 3　対応分析による特徴語と RF による特徴語

I, we, sit, few, ripe, dry (y 軸正)	
smooth, wife, grease (x 軸負)	I, we, sit, few, ripe (x 軸正)
new, hit, rope, fish, good, correct, grease (y 軸負)	

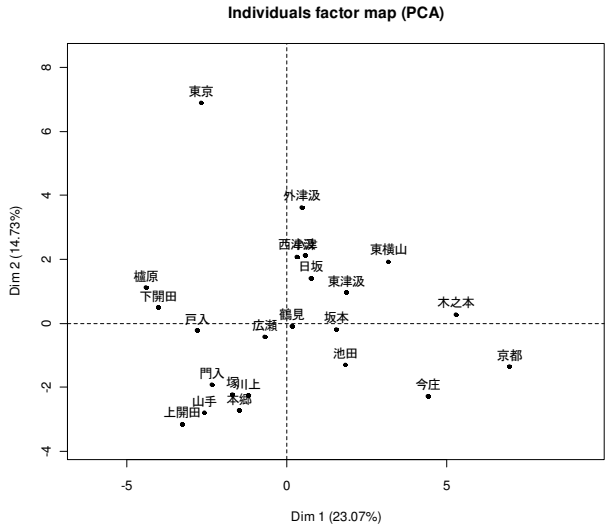

図 9　主成分分析（主成分得点）

特徴語として上位に挙がったのは，grease, good, bite, fish, correct, they, he, sit, lie である．この結果を対応分析によって抽出した特徴語と対照したのが表 3 である．対応分析による xy 軸正負の特徴語を示したのちに，random Forest によって抽出された語と重複するものを赤字で示した．この結果から，random Forest の結果は，対応分析の y 軸負の語を多く抽出していることが分かる．

図 9 図 10 は，random Forest で用いたデータを主成分分析した結果である．さらに図 11 は，主成分分析の集落に関する主成分得点を元にクラスター分析を行った結果である．

図 9 図 10 の結果から，random Forest によって抽出された系統推定を元にした分類に依存する語 grease, good, bite, fish, correct, they, he, sit, lie は，図 10 の丸で囲んだ位置に固まることが分かる．そして，図 9 の結果より，それらの語は，日坂，東津汲，西津汲，東横山，坂本，木之本といった集落において特徴的な様相を示すことが分かる．

図 12 は random Forest で算出された集落間のデータの相違度を MDS で描画した結果である．この配置もまた，図 3 で示した系統推定の配置とよく似た分布を示していることが分かる．

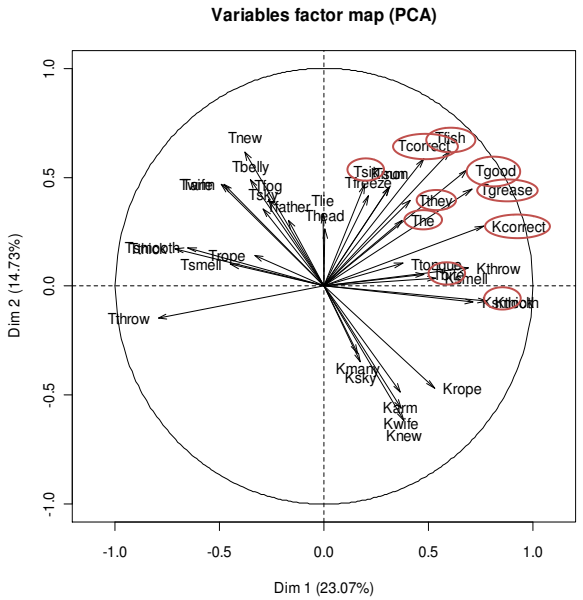

図 10　主成分分析（主成分負荷量）

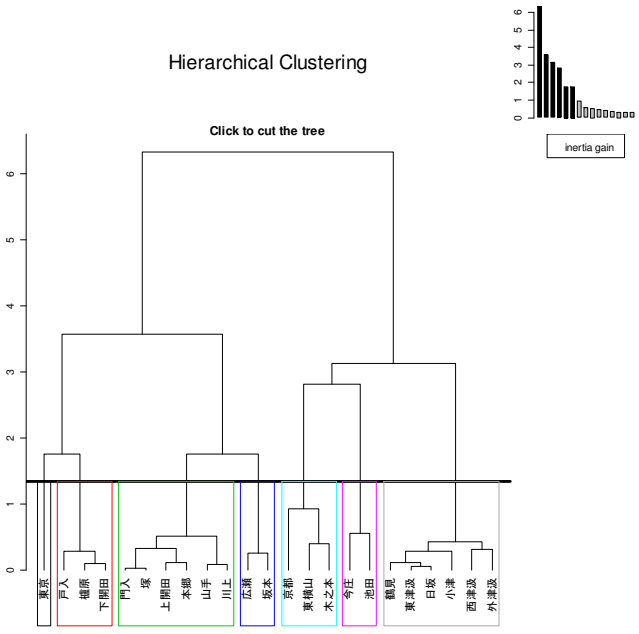

図 11　クラスター分析（主成分分析後）

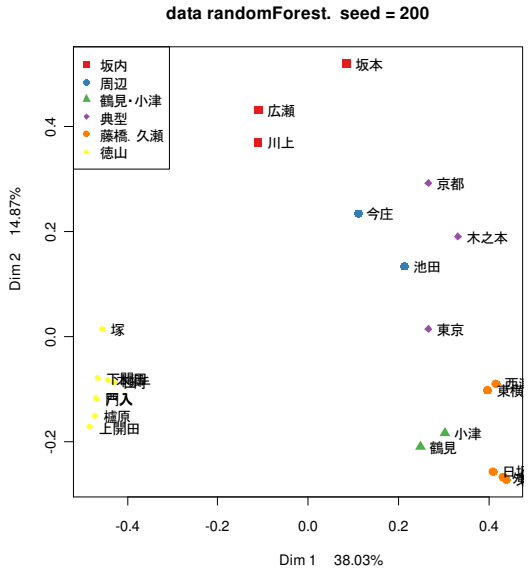

図 12　random Forest で用いたデータの MDS

5. おわりに

以上，基礎語彙を用いて，揖斐川上流域の使用語彙に関する地域差を見てきた．基礎語彙については，言語年代学としての使い方に問題があるとして，敬遠されがちである．しかしながら，地域で話される語彙について，類似性を抽出するという限定した使用においては，今回の分析において，有効な結果が得られたと言える．引き続き，対象地域を拡大し，方法論としての有効性を問う必要があると考えられる．

参考文献

1) Shiro Hattori: Japanese Dialects ,Current Trends ln Linguistics, Vo1, no.11, pp.368-400 (1973)
2) 崎山理: 言語による先史研究，言語人類学を学ぶ人のために，世界思想社, ,pp218-244 (1996)
3) Wurm, S. A.: Papuas Languages of Oceania, Tubingen: Genter Narr Verlag (1982)
4) 安本美典: 縄文語を追う　統計的にみた縄文語の姿，月刊言語 17-2, p.18-26 (1988)
5) 野村正良: 岐阜県揖斐郡徳山村戸入方言の記述的報告及び成立過程に就いての二三の考察——付アクセントの分布と変遷——，名古屋大学文学部研究論集 70, pp.3-17 (1977)
6) 山田達也: 揖斐川上流域における基礎語彙について，名古屋市立大学教養部紀要人文社会研究 22, pp.33-66 (1978)
7) 山田達也: 福井県今庄、池田方言及び滋賀県木之本方言の基礎語彙について—徳山村方言に関連して，名古屋市立大教養部紀要人文社会研究 24, pp.23-34 (1980)
8) 服部四郎:「言語年代学」即ち「語彙統計学」の方法について——日本祖語の年代——，季刊民族學研究 19(1)，日本文化人類学会, pp100-101 (1955)
9) Saitou N. and Nei M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees, Molecular Biology and Evolution, vol.4, no.4, pp.406-425 (1987)
10) David Bryant, Vincent Moulton: NeighborNet: An agglomerative method for the construction of planar phylogenetic networks, Molecular Biology and Evolution, 21, pp.255-265 (2004)
11) D. H. Huson and D. Bryant: Application of Phylogenetic Networks in Evolutionary Studies, Molecular Biology and Evolution, 23(2), pp.254-267 (2006)
12) Package‘FactoMineR’, http://cran.r-project.org/web/packages/FactoMineR/FactoMineR.pdf, pp.1-62 (2012)