方言分布から見た文化事象の流通と滞留—岐阜県における方言分布の事例—

小野原　彩香・藤本　悠

A Study of Dialects Distribution Reflecting Flows and Stasis of Cultural Phenomena -A Case of Gifu Prefecture-

Ayaka Onohara, Yu Fujimoto

Abstract: It is difficult to comprehend ‘culture’ from a single cultural fact such as decorations of ancient potteries because a certain culture consists of various human activities. Therefore, we attempt to set up ‘a priori cultural territories’ as a ‘seed’ of culture at first and to continuously modify multifaceted culture by overlaying various cultural phenomena as a result. Consequently, our final goal of this project is to find multifaceted culture but here is not to tackle this purpose; instead we first propose new method for analyzing similarity of ‘vocabularies’ to construct fundamental a priori cultural territories by calculating Normalized Compression Distance (NCD) and classifying each territory with Neighbour Net. We tentatively attempt to apply the method to a study of dialects in Gifu prefecture Japan. Additionally we aim to make clear all the process and verify transparency of entire workflows of our methods with using UML activity diagram.

Keywords: 文化、方言、正規化圧縮距離（NCD)、Neighbor Net

1　はじめに

「文化」は、人間活動の時空間的な共通性あるいは相違性として現れる様々な文化現象によって構成される。それゆえに、文化研究は特定の文化現象のみからではなく多角的な視点から行う必要がある。藤本が代表者を務める「CITMAS プロジェクト」では、こうした文化に対する概念をリッケルトの考え(1915)を参考に情報モデル化している(図 1)。本プロジェクトにおいては、これを多様な文化現象を情報化するための「メタ・メタモデル」として位

小野原彩香　〒610-0394 京都府京田辺市多々羅都谷 1-3
同志社大学大学院文化情報学研究科
Phone：0774-65-7610
E-mail: aonoa68@gmail.com

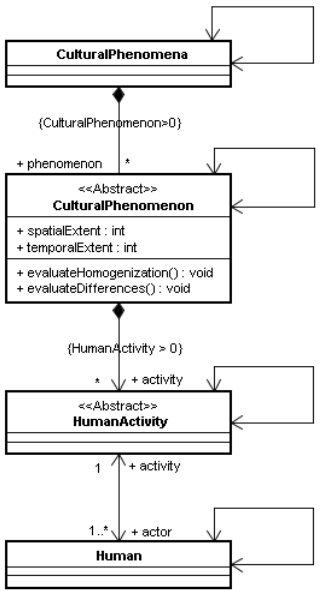


図 1　文化現象のメタ・メタモデル

置づけている。本研究ではこのメタ・メタモデルを元に文化現象としての「方言のメタモデル」を設計・実装し、他の文化現象を総合的に重ね合わせるための分析手法を検討した。

2　研究の背景と先行研究

本研究では、コミュニケーションという人間活動の時空間的共通性と相違性を反映する方言分布に注目し仮説的に文化圏を構築した。

本研究の先行研究である『徳山村のあけぼのを求めて』(1984)では、岐阜県旧徳山村において、関東・東海系アクセントを取り巻くように、関西系アクセントが存在し、一部で極めて特殊な中間型アクセンであることが紹介されている。泉ら(2007)は、これらの研究成果を踏まえ、考古学における遺跡情報と方言分布、集落特性の明確化を試み、方言分布に違いが見られた地域に存在する戸入村平遺跡や塚奥山遺跡などは、集落の規模や土器系統に周辺遺跡とは異なる特徴があることを指摘した。この研究結果から、方言分布と縄文集落には、地理空間的脈絡が反映されていると考えられる。

3　方言モデルの設計とデータの実装

本研究では、最初に文化現象としての方言をメタモデルとして設計した。このメタモデルは前述のメタ・メタモデルを継承する。図 2 において、「Dialects(方言)」クラスは「Speaker(話し手)」クラスと、「Vocabulary(語彙)」クラスと関係を持つ「Message(メッセージ)」クラスから成る。これらのクラスは、「地理情報標準」に準じていて、「応用スキーマ(ISO19109)」と同様の役割を果たす。

これらのクラスは既存の資料から収集した情報によって実装し、「Communication」クラスは『岐阜県方言の研究』(奥村, 1976)に記載された岐阜県全域の方言語彙の地域差を示す 10 枚の分布図から実装した。これらは、「唐がらし」、「唐もろこし」、「薬指」、「旋毛」、「竹馬」、「お手玉」、「粳米」、「馬鈴薯」、「真綿」、「里芋」という 10 の調査語彙の分

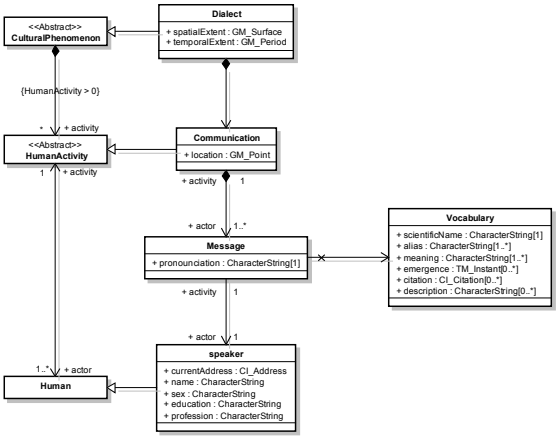


図 2　文化現象としての方言の定義（メタモデル）

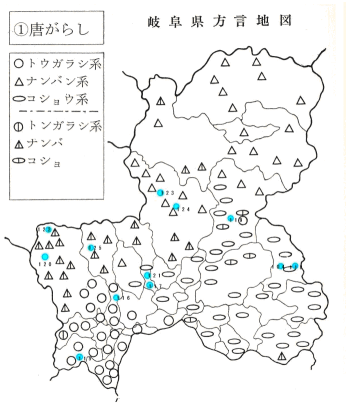
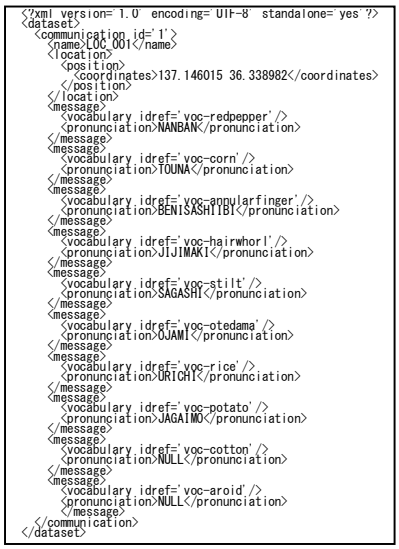


図 3　「唐がらし」の図

布を示している。それぞれの分布図では 125 箇所の調査地点について、各語彙をどのように発音するかを記号でプロットしている(図 3)。本研究ではこれらの図から位置データと語彙の発音データを取得し、最終的に「地理情報標準」における「符号化規則(ISO19118)」に従って XML ドキュメントでインスタンス化した(図 4)。また、 ArcGIS を用いて地形の傾斜角度と調査地点からコスト・アロケーション領域を構築し、これを暫定的に仮説領域を構成する小地域とした。

4　方言の類似性分析と分類

本研究では、方言の類似性を分析するために、汎



```
<?xml version='1.0' encoding='UTF-8' standalone='yes'?>
<dataset>
  <communication id='1'>
    <name>LOC_001</name>
    <location>
      <position>
        <coordinates>137.146015 36.338982</coordinates>
      </position>
    </location>
    <message>
      <vocabulary idref='voc-redpepper'/>
      <pronunciation>NANBAN</pronunciation>
    </message>
    <message>
      <vocabulary idref='voc-corn'/>
      <pronunciation>TOUNA</pronunciation>
    </message>
    <message>
      <vocabulary idref='voc-angularfinger'/>
      <pronunciation>BENISASHITBI</pronunciation>
    </message>
    <message>
      <vocabulary idref='voc-hairwhorl'/>
      <pronunciation>JIJIMAKI</pronunciation>
    </message>
    <message>
      <vocabulary idref='voc-stilt'/>
      <pronunciation>SAGASHI</pronunciation>
    </message>
    <message>
      <vocabulary idref='voc-otedama'/>
      <pronunciation>OJAMI</pronunciation>
    </message>
    <message>
      <vocabulary idref='voc-rice'/>
      <pronunciation>URICHI</pronunciation>
    </message>
    <message>
      <vocabulary idref='voc-potato'/>
      <pronunciation>JAGAIMO</pronunciation>
    </message>
    <message>
      <vocabulary idref='voc-cotton'/>
      <pronunciation>NULL</pronunciation>
    </message>
    <message>
      <vocabulary idref='voc-aroid'/>
      <pronunciation>NULL</pronunciation>
      </message>
  </communication>
</dataset>
```

図 4　ISO 標準に準拠した XML データベースファイル(調査地点 1)

用類似度算出技術の一つである Normalized Compression Distance(NCD)を算出し、分類する方法を検討した。NCD とは、以下に示す式によって表された正規化圧縮距離である。

$$NCD = \frac{C_{XY} - \min\{C_X, C_Y\}}{\max\{C_X, C_Y\}}$$

ここで、C はある圧縮アルゴリズム、X および Y は任意のファイルである。 X, Y それぞれを圧縮したファイルのサイズを C_X,C_Y とし、X に Y を接続して圧縮したファイルのサイズを C_{XY} とする。

本研究においては、X および Y は、XML ドキュメントでインスタンス化された各々のファイルである。この分析方法の利点の一つは、一般的な分析手法と比較して、事前のデータ抽出と整形などが簡略化できることにある。また、本研究のように、地理情報標準に準拠して実装されたインスタンスの分析手法としても注目できる。本研究においては、藤本が開発した NCD 算出ソフトウェア「百花壱式 Ver1.0」を用いて NCD 行列を作成した。NCD は、圧縮アルゴリズムに依存するが、同ソフトウェアの現バージョンは GZIP 圧縮のみをサポートしている。地理情報標準に準じてインスタンス化された XML

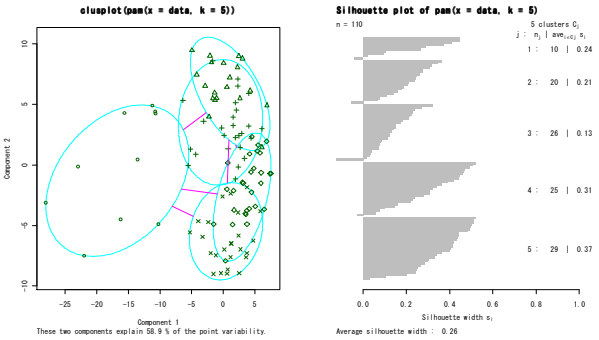


図 5　PAM5 クラスタの主成分 plot と Silhouette plot

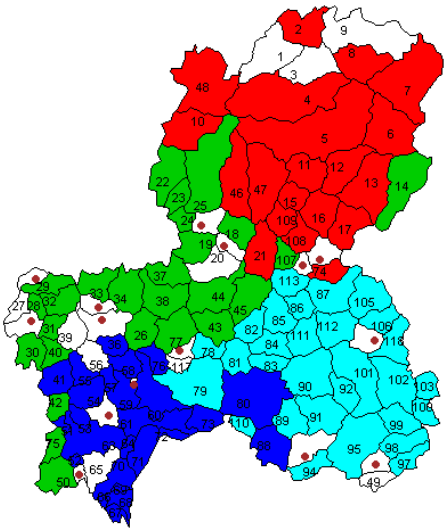


図 6　PAM での 5 クラスタの結果を地理空間上に投影したもの

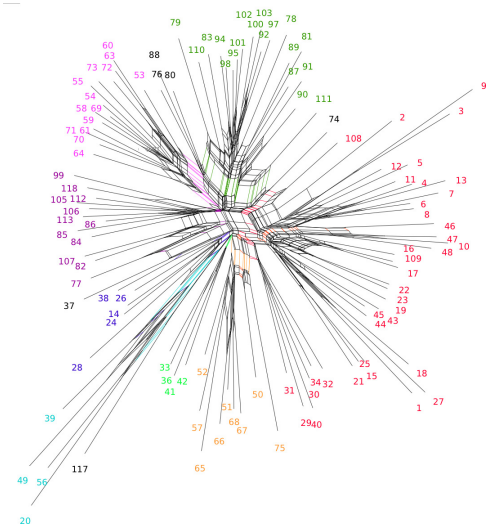


図 7　NCD における Neighbor Net ネットワーク

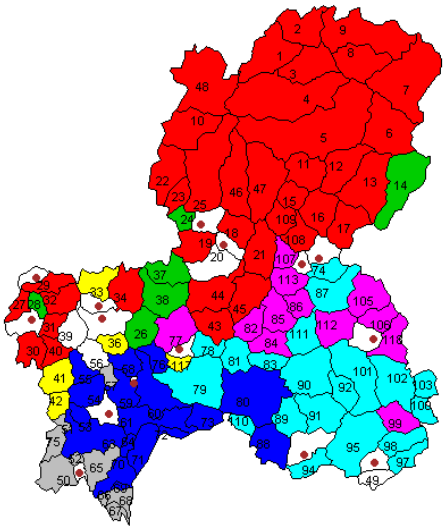


図 8　Neighbor Net ネットワークによる分類を地理空間上に投影

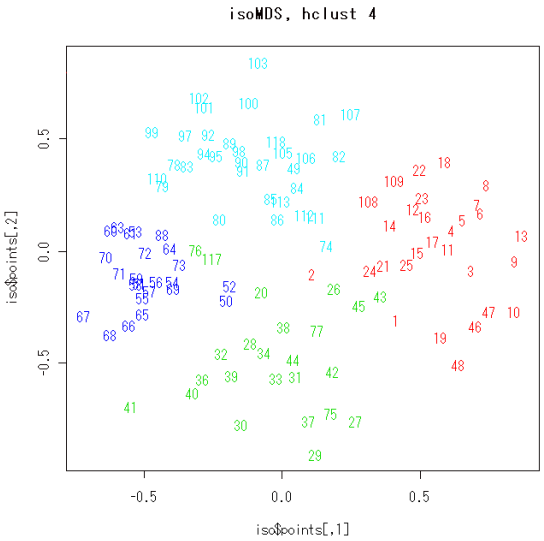


図 9　isoMDS による 2 次元配置の散布図

文書から NCD を求めるという方法は、今回、藤本が新たに考え、試みた手法である。なお、欠損値が多い調査地点 35,62,93,96,104,114,115,116,119,120,121,122,123,124,125 を予め除外した。

結果として算出した NCD 行列は、 非階層的クラスタリング Partitioning Around Medoids(PAM)、階層的クラスタリング Ward、Average 、系統学的 Neighbor Joining(NJ)、Neighbor Net、およびグラフクラスタリング DPClus の各手法を用いて分類した。まず、PAM を用いた分類を通して、分類基数および地理空間的な傾向を観察した。この分析の結果、PAM では、5 クラスタとし、外れ値の 1 クラスタを除いた 4 分類とするのが適切であると分かった(図 5)。なお、図 6 のような地図と調査地点の描画には、R のパッケージ maptools などを用いていて、欠損値として除外した調査地点については小円・茶色のラベルで示している。

各分類の結果、NCD 行列からの分類法としては、NJ のアルゴリズムに基づく Neighbor Net が、先行研究の成果を最も忠実に再現できていることが解った(図 7、図 8)。

図 8 では先行研究にみられる調査地点 28 の孤立や、63 と 51 の東西アクセントの対立が現れている。

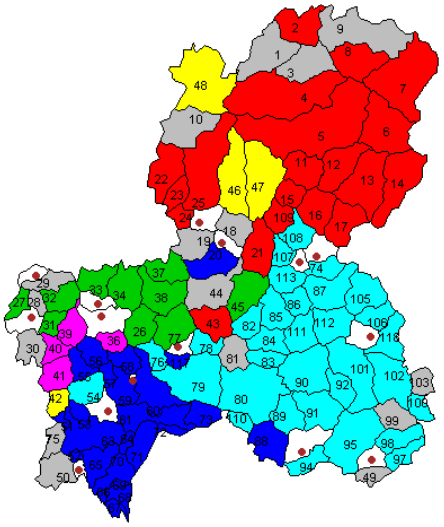


図 10　Jaccard 係数による DPClus の結果を地理空間上に投影

5　分析結果の検証

ここまでの結果によれば、XML に基づく NCD 距離行列は、発音の類似性を適切に反映している可能性があるため、他の一般的な手法との比較検証を行う。本研究では先行研究と検証用データセットの両側面から検証した。

検証に用いたデータセットは、列を調査語彙、行を調査地点番号とし、要素として各調査地点での発音を一文字の記号に置き代えて入れた。一つの調査地点に二つの発音がある場合は、横に二つ記号が並んでいれば左側を、縦に二つ記号が並んでいれば上を採用した。また、列を調査語彙のすべての発音、行を調査地点番号にした、2 値データ(各調査地点で語彙が存在する場合は 1、存在しない場合を 0 とする)も準備した。なお、NCD の距離行列の場合と同様に欠損値の多い調査地点は予め除外し、合計 110 地点にて分析を行った。

最初に、この検証用データセットでの適切なクラスタ数を選択するため、多次元尺度構成法(MDS)を行い、データの概要把握を試みた。この分析の結果、全調査地点を 4 つにグループ分けすることが適当であると分かった(図 9)。その一方で、先行研究を通して明らかとなった地域差を見ようとするならば、どのクラスタ分析においても、7 から 8 のクラスタに分ける必要があることも明らかとなった。

MDS の結果を受け、クラスタ数を 4 から適宜増やして分析を行った。今回は、距離に Gower と Jaccard（２値データの場合)、クラスタリングでは、NCD の場合と同様の手法を試した。

方言学における先行研究では、揖斐川が東方アクセントと近畿アクセントの境界として古くから知られている。北の境界は大垣市(調査地点 63)と垂井町(調査地点 51)の間にも走っており、垂井町およびその付近には特種のアクセントが行われている(服部,1930;柴田,1950)。また、徳山村の戸入地区(調査地点 28)は、東方アクセントであるとされている(徳山村の歴史を語る会,1984)。

分析の結果、 全ての分析の結果では、Jaccard の距離行列から構成したグラフに対する、DPClus (距離<0.85、density=0.95、最小クラスタにおける構成地点数=4)を用いた分類が、これらの先行研究結果を最も良く反映していた。この分析の結果である図 10 では、徳山村の戸入集落(調査地点 28)の孤立を始めとした揖斐川流域における東方アクセントと近畿アクセントの対立が現れており、従来の研究成果との間に齟齬は無い。この図において、欠損値として除外した調査地点については小円・茶色で示し、特定のクラスタに分類されなかった調査地点については、灰色で示している。したがって、灰色の調査地点内で一つのクラスタを形成することを示すものではない。

NCD による分析結果と検証データによる分析結果を比較すると、各グループの中核的な構成地点に差は無く、NCD による分析の有効性は示された。今回は語彙分布の類似性データを用いたのであるが、アクセント分布による先行研究結果との間にも齟齬が無いことにも注目すべきである。

6　処理プロセスの明確化

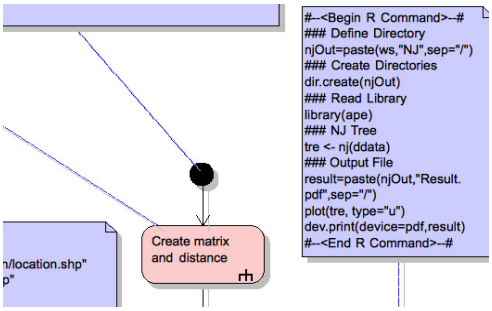


図 11 UML アクティビティ図と R スクリプトによる処理の明確化

本研究では、UML アクティビティ図を用いて、処理プロセスを記述する試みも行った。近年、文化情報学分野などでは高度な分析手法を文化現象研究に応用されているが、一方で分析プロセスが不透明化し、方法論の再利用を著しく阻害している。これらの問題に取り組むためには、より客観的な方法によって、分析プロセスの明確化に努める必要がある。

同様の試みは、ABM(Agent Based Model)に関連する分野においても積極的に行われており、UML クラス図と UML アクティビティ図(あるいは UML シーケンス図)を用いて分析の適用範囲の明確化や処理プロセスを明記することが一般化しつつある。また、OpenABM[3]が主導する業界標準 OD

D プロトコルにおいても、UML クラス図の積極的利用が推奨されている。

本研究における試みでは、こうした試みと同様に処理プロセスを明確化すると同時に、分析方法の明確化を図った。具体的には、統計ソフトウェア「R」のスクリプトを用いて、処理内容をコメントに書き込んだ(図 11)。こうすることで、モデル駆動とまでは言わないまでも、コピー・アンド・ペーストで一連の分析行程が再現できるほか、経験的に、遠隔地での共同研究を大幅に効率化できることが解っている。また、文化現象研究の研究プロセスにおいて、どの程度が自動的な処理として行われているか、あるいは、どの程度恣意性が排除されているかを評価する上でも注目できる。

7　まとめ

本研究では、多角的な文化を見いだすための基礎となる「仮説的文化領域」の構築手法の検討を行った。また、一連の手法の処理プロセスを明確化するための方法として、UML アクティビティ図と R スクリプトを用いた方法の提案を行った。仮説的な文化領域構築として試みた NCD と Neighbor Net による分類は、地理情報標準で規定されているようなオブジェクト指向の分析手法として可能性を持っていることが分かった。また、XML を元に NCD で分析を行う場合には、分析前の準備作業が簡略化できるので、膨大なデータを扱うことになる文化研究には総じて有効であると考えられる。

UML アクティビティ図と R スクリプトを用いた処理プロセスの記述方法に関しては、実験的な試みでとどまっているが、将来的には、地理情報標準に準拠したデータセットを利用した研究や、同標準の規則を利用した研究は、より活発化すると期待でき、UML をはじめとする情報モデルを中心とした研究は、益々、重要性が増すと考えられる。しかし、現実には NCD などの R パッケージに含まれないものも存在することもあり、そうした場合の対処方法を考えることも重要である。

方言研究としては、調査地点は 7 から 8 クラスタに分類でき、そのうち 3 から 4 クラスタは地域的なまとまりを持たないクラスタであることから、地理空間的な視点で見ると、岐阜県下は大きく 4 つのグループに分けることができる。興味深いことに、これら 4 つのグループの境界の内の一つは岐阜市周辺に位置し、この傾向は、藤本(2010)による縄文土器を用いた分析でも指摘されている。

参考文献

[1] リッケルト. 佐竹哲雄訳 (1939): 「文化科學と自然科學」. 岩波書店. (H.Rickert, 1915. Kulturwissenschaft und Naturwissenschaft, 3rd ed.)

[2] Open ABM: http://www.openabm.org/.

[3] 徳山村の歴史を語る会 (1984): 「徳山村のあけぼのを求めて―岐阜県揖斐郡徳山村遺跡分布調査中間報告―」.

[4] 泉拓良, 藤本悠, 三嶋誠 (2007): 旧徳山村所在の縄文集落の GIS 研究,日本文化財科学会第 24 回大会講演論文集.

[5] 奥村三雄 (1976): 「岐阜県方言の研究」. 大衆書房.

[6] D.H. Huson and D. Bryant, 2006. Application of Phylogenetic Networks in Evolutionary Studies, Molecular Biology and Evolution, 23(2):254-267. software available from www.splitstree.o.

[7] 服部四郎 (1930): 近畿アクセントと東方アクセントとの境界線「音聲の研究」. 第三輯. 文學社.

[8] 柴田武 (1950): 揖斐川上流のアクセント「文字と言葉」. 刀江書院.

[9] 藤本悠 (2010): 文化情報学および理念型モデル化分析法に関する基礎的研究 －文化情報学における諸概念と方法について－. 2009 年度同志社大学大学院文化情報学研究科博士後期課程博士論文 2010 年 3 月提出. (http://www.cis.doshisha.ac.jp/kundaikan/students/doctoral/fujimoto/data/pdf/D_Main_ver.2.0.0.pdf).