Kolmogorov 記述量に基づく類似度距離による方言自動分類の試行

田中 ゆかり
日本大学 文理学部 国文学科

谷 聖一
日本大学 文理学部 情報システム解析学科

Kolmogorov 記述量に基礎を置いた類似度距離が Li らにより提案されている．この手法の日本語方言分類に対する有効性を検証するために行った基礎的な計算機実験の結果を報告する．

An application of the similarity metric based on Kolmogorov complexity to automatically classification of Japanese dialects

Yukari Tanaka
College of Humanities and Sciences
Nihon University

Seiichi Tani
College of Humanities and Sciences
Nihon University

The similarity metric based on Kolmogorov complexity introduced by Li et al. is known to be useful in clustering various subjects. We investigate the validity of this method for clustering of Japanese dialects.

1．まえがき

コンピュータ，ストレージ，ネットワークの発展に伴い，近年ますます大量のデータが電子的に蓄えられている．これらの大量のデジタルデータを活用するため，さまざまな場面でデジタルデータから自動的に特徴を抽出するデータマイニング技術の発達が期待されている．デジタルデータ間の類似性の自動判定も，データマイニングにおける重要な課題の１つである．そのような手法のひとつに Li らにより提案された Kolmogorov 記述量に基礎を置いた類似度距離がある [10, 13]．この類似度距離は次の特徴を備えている．

1. 統計的な性質に依存せず，個々のデータ間の類似度が決まる．
2. ある程度の差を無視すれば，ある条件を満たしたどのような正規化距離と比較しても，この距離が大きくならないことが証明されている．つまり，比較対象の特質に応じて最適な類似度距離は異なるであろうが，ある程度の差を無視すればこの類似度距離は最適な類似度距離と同等の性能を有するといえる．このような意味で，この類似度距離は万能性を持っているといえ，対象を選ばない．

ただし，Kolmogorov 記述量は計算不能であるため，この類似度距離も計算不能である．そのため，計算機実験を行う際はデータを圧縮したサイズで代用する方法が提案されており，対象によってはこの方法が有効であることが実験的に確かめられている．また，圧縮したサイズで代用した場合でも，圧縮方法がある条件を満たしていれば依然として万能性を持つことが知られている[11]．本研究では，この手法の方言の分類に対する有効性を探るため計算機実験を行った．

2．「方言区画」に関する先行研究と本研究の位置付け

日本語の方言がどのように分類されるか，という関心は強く，多くの先行研究が存在している．方言の区画・分類に際しては，区画・分類に用いる指標，分類方法，いずれもさまざまな観点からの分類が可能である．ここでは，主観による区画・分類と，計量的手法による区画・分類のうち，よく知られている区画・分類について概観し，方言区画・分類本研究の中に位置づける．

2.1 主観による方言区画･分類

主観による区画・分類のうち，もっともよく知られ，かつ現在の方言研究においてもよく用いられる区画・分類は東條 [7] によるものである．これは，アクセント・音韻・文法・語彙の諸事象を東條自身が総合的に捉え，判断した方言区画案である．この分類では，日本の方言を，まず「本土方言」と「琉球方言」に大区分し，「本土方言」を「東部／西部／九州方言」に中区分する．細かい区分には言語事象の特徴に加え，旧国境・都道府県境など行政区分も勘案された区分となっている．

日本語
- 本土方言
  - 東部方言：北海道，東北，関東，東海東山道，八丈(島)
  - 西部方言：北陸，近畿，中国，雲伯，四国
  - 九州方言：豊日，肥筑，薩隈
- 琉球方言：奄美，沖縄，先島

方言体系のうち，ある事象に重点化した方言区画も存在する．このうちよく知られた区画・分類に，アクセントの分布を主とした金田一による区画 [5] などがある．金田一の区画は，東條の区画と大きく異なり，日本語の方言を次のように4区分する．近畿方言を同心円の中心（内輪方言）とし，その外周に中輪，外輪，南島方言が位置する区画図を示している．

日本語
- 内輪，中輪，外輪方言
- 南島方言

その他に方言敬語システムの複雑さの観点からの区分 [3] や，方言意識による区画 [8] なども提案されているが，1970 年代以降は，主観による方言区画・分類論が一段落ついたかたちとなった．

2.2 計量的手法を用いた方言の区画・分類

1980 年代以降，コンピュータの発達とその文系領域への進出に伴い，計量的手法を用いた方言区画・分類研究が盛んになった．代表的なものとして，[4]，[2]，[1] が上げられる．国立国語研究所がこの間に刊行した語彙を中心とした『日本言語地図 1 ～ 6 』の 2400 地点における 82 項目にあらわれる標準語形データを用いて，方言区画・分類を示した．河西は標準語形の出現率による区画・分類を行い [4]，井上・河西は同じデータを因子分析によって区画・分類した [2]．文法項目を中心とした計量的手法を用いた井上の結果 [1] は，概ね東條などによって総合的観点から示されてきた「琉球／九州／西部／東部」という方言区画が妥当であることを示した．その後，地点間一致率行列を用いてクラスター分析による分類を行なった [6]，活用形の共通語との語形一致の観点から数量化第 3 類によって方言の分類を行なった [9] など，各種の対象・手法から方言の区分・分類について検討されている．

2.3 本研究の位置

本研究は，新しい試みとして Kolmogorov 記述量に基づく類似度距離を用いて，方言の自動分類を行なう．分類対象は，『方言ももたろう』に集録された全国 53 地点の伝統的方言話者による「各地方言訳ももたろう」の談話を文字化した資料である．従来の計量的方言区画・分類研究は，語彙ならば語彙，活用形ならば活用形と事象を限定した区画・分類の試行であった．本研究は，談話文字化資料を用いることから，アクセント・イントネーションのような音調に関する事象は欠落するものの，談話文字化資料として表現できる範囲においてではあるが，各地の方言を包括的な観点から分類する試みであり，新たな知見を従来研究に追加できる可能性をもつ点である．

3. 手法の概説

Kolmogorov 記述量は，文字列の複雑さをその文字列を生成するプログラムの最小サイズで評価するという方法であり，1960 年代に R. Solomonof, A. Kolmogorov, G. Chaitin により独立に提案された [14]．U をある万能チューリング機械とする．形式的には，文字列 x の Kolmogorov 記述量は次のように定義される．

$$K(x) = \min\{|p| : U(p) = x\}$$

補助情報 y を用いた文字列 x の Kolmogorov 記述量 $K(x|y)$ を

$$K(x|y) = \min\{|p| : U(p,y) = x\}$$

と定義する．$K(x|y)$ は補助情報として y を用いて x を生成する最小のプログラムのサイズである．$K(x)$ に比べて $K(x|y)$ が小さいならば，y の中に x に関する情報が多く含まれていると解釈することもできる．この考え方を基礎に，Bennett らは文字列間の類似度を表す information distance という概念を提案した [10]．この概念を発展させ Li らは文字列 x と y の normalized information distance (NID)を次のように定義した [13]．

$$NID(x,y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}$$

[13] では，

- ある程度の誤差を許せば NID は距離の公理を満たすこと
- ある程度の誤差を許せば，ある条件を満たしたどのような正規化距離と比較しても，この距離が大きくならないこと

などが証明されている．つまり，比較対象の特質に応じて最適な類似度距離は異なるであろうが，ある程度の差を無視すればこの距離は最適な類似度距離と同等の性能を有するといえ，このような意味で，万能であるといえる．このようなことから，Li らは NID を similarity metric と呼ぶことができると主張している．

NID は，このような数学的に良い性質を持つが，計算不能な Kolmogorov 記述量を用いて定義されているため，計算不能である．そこで，Li らのグループは，比較対象の文字列をある圧縮方法で圧縮したサイズで代用し計算機実験を行い，いくつかの対象に対して実際に有効であることを示している [13, 11]．x と y の連接を xy と表すことにする．すると，

$$NID(x,y)=\frac{K(xy)-\min\{K(x),K(y)\}}{\max\{K(x),K(y)\}}$$

が成り立つことを示すことができ，$K(x)<K(y)$ のときは，

$$NID(x,y)=\frac{K(xy)-K(x)}{K(y)}$$

となる．文字列 x をある圧縮方法 C で圧縮したサイズを $C(x)$ と表すことにする．$K(x)$ を $C(x)$ で代用して *NID* を計算した値を normalized compression distance (*NCD*) と呼ぶことにする．つまり，

$$NCD(x,y)=\frac{C(xy)-\min\{C(x),C(y)\}}{\max\{C(x),C(y)\}}$$

であり，$C(x)<C(y)$ のときは，

$$NCD(x,y)=\frac{C(xy)-C(x)}{C(y)}$$

となる．Li らのグループでは，この *NCD* を用いて実験を行っている．これらの実験では，経験上良い結果がでるという理由で bizip2 が用いられていた．*NCD* の利点は，比較する文字列およびその連接を圧縮し，簡単な計算を行うだけで求まる点である．

一方，各圧縮方法 C に対して $C(x)$ が $K(x)$ をどのくらい良く近似しているかを評価する方法は知られていない．そこで，Cilibrasi と Vitányi [11] は，以下のように圧縮方法 C が normal であるという概念を提案し，*NCD* を計算するのに用いる圧縮方法 C が normal であるなら，*NCD* もある程度の万能性を持つことを示している．（*NID* が万能であるというときの万能性の定義を *NCD* は満たすことができないので，定義を弱めている．詳しくは，[11] を参照のこと．）

次を満たすとき圧縮方法 C は normal である：

1. 任意の文字列 x に対して $C(xx)=C(x)$ であり，空文字列 λ に対して，$C(\lambda)=0$ である．
2. 任意の文字列 x, y に対して $C(xy)\geqq C(x)$．
3. 任意の文字列 x, y に対して $C(xy)=C(yx)$．
4. 任意の文字列 x, y, z に対して

$C(xy)+C(z)\leqq C(xz)+C(yz)$．

ただし，等号・不等号の両辺に現れる文字列の最大長を n とすると $O(\log n)$ の誤差を許す．

4．計算機実験

著者らのグループでは，先行研究が扱った対象に対し追実験を行い，これらに対する *NCD* の有効性を再確認した上で，日本語方言の分類への有効性を確認するための基礎実験を行った．本報告における実験の分類対象は，CD-ROM『方言ももたろう（杉藤監修）』[15] に収録された全国 53 地点の伝統的方言話者による「各地方言訳ももたろう」の談話を平仮名と句点を用いて文字化した資料である．

4.1 文字化方法

3 名の研究協力者が文字化を担当した．3 名が独立に個人内のルールで統一し文字化をしたものと，個別の文字化を終えた後に 3 名が共同で文字化したものの 4 種類の資料を作成した．本報告で用いた資料は，3 名が共同した作成したものである．

4.2 圧縮方法と文字コード

本報告では，圧縮方法は bizp2 を用いた．

図 1 は，6 つの文字コードに対して文字化資料の圧縮後のサイズを比較したものである．

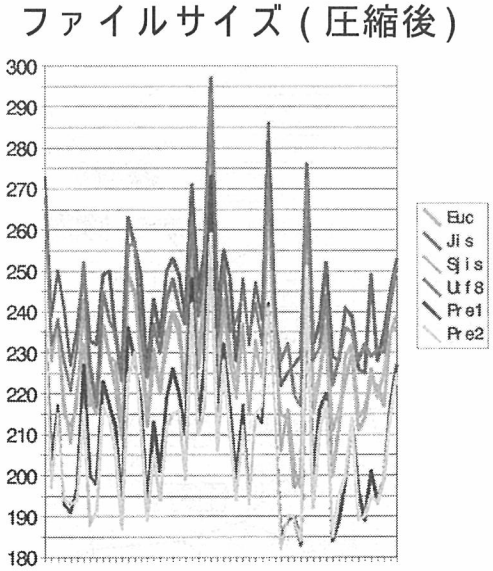

図 1: 文字化資料圧縮後サイズのによる文字コードの比較（横軸：地点）

- Shift_JIS
- ISO-2022-JP

- EUC-JP
- UTF-8
- 「ぁ」の文字コードとの差分を 7bit 表記（このコードを Pre1 と呼ぶ）
- 53 地点の文字化資料中の頻出頻度を 7bit 表記（このコードを Pre2 と呼ぶ）

Pre1 と Pre2 では圧縮後サイズはほぼ同じであったが，ほとんどの地点でわずかに Pre2 の圧縮後サイズの方が小さい．

bzip2 が Pre2 で記録されたデータに対して normal である 4 つの条件をどの程度満たすか検証した．3 番目の条件は，bizp2 がブロックソートを用いていることから明らかに成り立つ．実際に，2 番目と 4 番目が成り立つことを確認した．1 番目に条件については，平均して

$$C(xx) : C(x) = 280.7857 : 231.9821$$

となり，満たしているとはいえなかった．

4.3 系統樹の構成と描画状図の解釈

4.2 節で示した 6 つの文字コードそれぞれに対して，53 地点の文字化資料の全ての 2 点間の *NCD* を算出した．つまり，6 種類の 53×53 距離行列を作成した．得られた 6 種類の距離行列に対して，3 つの古典的な手法（NJ 法，UPGMA 法，Quartet 法）で系統樹を構成した．この系統樹の構成までは自動で行っている．図 2 は，NJ 法で構成した系統樹を，C++ 用クラスライブラリ LEDA のグラフ描画用クラスを用いて，ランダムに描画した後，spring-embedder という再配置メソッドを数回手動で施したものである．

比較するデータを葉に対応させ，内点は次数が 3 となる木を構成する．このような木から葉を 4 つ選び，選んだ 4 つの葉とこれらの葉同士を結ぶ経路からなる部分グラフを考えると，図 3 のように u と v の組，w と x の組のように経路が交差しない 2 組に分けることができる．葉の 4 つ組に対して，経路が交差しない葉間の距離の和（図 3 の場合は，$NCD(u, v) + NCD(w, x)$）をこの 4 つ組のコストとし，全ての葉の 4 つ組のコストの総和を木の Quartet 木コストと呼ぶことにする．つまり，距離が小さい葉の組をなるべく交差しないように木を構成すれば Quartet 木コストは小さくなる．しかし，距離行列が与えられたとき，Quartet 木コストを最小にする問題は NP 困難であるので，効率の良い解法は期待できない．そこで，Quartet 法ではヒュー

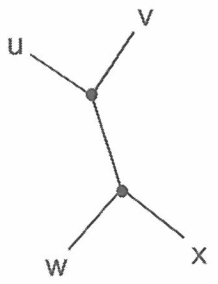

図 3：葉 u, v, w, x から定まる部分木

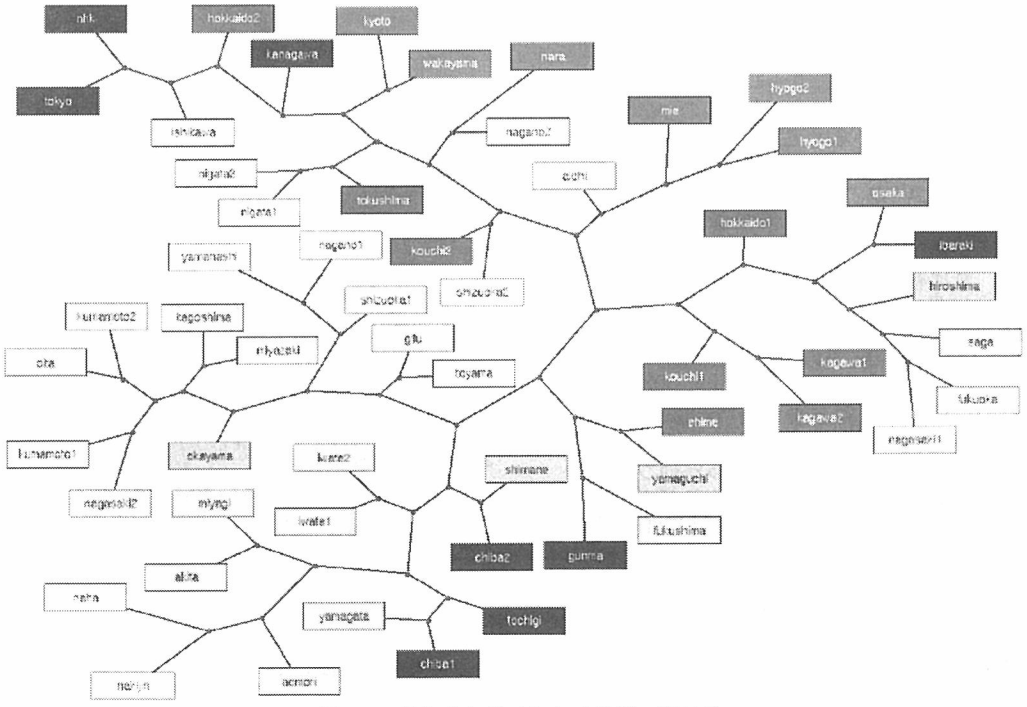

図 2: NJ 法により構成した系統樹の描画例

リスティクスを用いて Quartet 木コストがなるべく小さな木を構成しようとする．ある程度 Quartet 木コストが小さな木は，多くの葉の 4 つ組において距離の近い葉の経路が交差しないよう良く分別されているはずであり，その有効性は多くの対象において確認されている．しかし，方言ももたろうの文字化資料に対しては，良い結果は得られなかった．

5．考察

5.1 NJ 法による樹状図の解釈

NJ 法により構成した系統樹の描画例を，従来の方言区画の観点からみると，以下で示す群が割合よくまとまりをみせている．これら従来の方言区画を割合よく反映している方言群は，語形・語末無声化・語中有声化・ズーズー弁など，文字に転写することが可能な特徴をもつ方言群であるようにみえる．文字に転写可能な音声的特徴を多くもつ方言群がまとまりをみせている結果となった．

この結果，地域的には隣接していないものの音声的特徴を共有する方言がまとまりをみせている部分も見られる．たとえば，東北・北関東群に中国地方方言の島根が群化しているが，これは，島根方言のズーズー弁的特徴が文字に転写された結果，東北・北関東群としてまとまりをみせた結果と推測できる．

- 沖縄：那覇・今帰仁
- 東北・北関東：青森／宮城・秋田／山形・千葉／栃木／岩手１・岩手２／千葉２・島根
- 九州中南部：熊本２・大分／熊本１・長崎２／鹿児島・宮崎
- ナヤシ方言：山梨／長野１／静岡１
- 中部：岐阜・富山
- 四国：香川１・香川２／高知
- 九州北部：福岡・長崎１／佐賀
- 関西 A：兵庫１・兵庫２／三重
- 関西 B：京都・和歌山
- 共通語：ＮＨＫ・東京／石川／北海道２／神奈川

一方，NJ 法により構成した樹状図には，従来の方言区画の観点からは説明がしにくい部分も目立つ．もっとも共通語的な変種と想定される NHK と石川が非常に近いところに位置している，大阪と茨城がもっとも近い距離に位置しているなどがそうである．このような部分においては，何によってこのように結果が得られたのか分からない．従来の方言区画で見落とされてきた要素によって新しい分類が提案された，というよりは，用いたデータの何らかの特性に反応した結果，あまり適当ではない結果が導き出されたという解釈になりそうだ．

今回分析に用いたデータは，音声資料を文字化したものであるため，文字に反映されないアクセントやイントネーションのような韻律に関わる部分が分類対象とならない．そのことがうまく説明できない部分を大きくしている可能性も高い．

5.2 NHK からの距離を地図化した結果の解釈

図 4 は，各方言の NHK からの距離を地図上に反映させたものである．東京・神奈川・長野２・石川・京都・奈良・和歌山・北海道２が NHK からの距離の近いものということになる．現在の共通語基盤方言である東京・神奈川や，北海道共通語である北海道２については，妥当な結果といえそうであるが，その他の方言は，他の方言より NHK に近い距離をとる理由がはっきりしない．

東日本において，太平洋側が比較的 NHK に近く，日本海側が遠いという結果は，近年のグロットグラム調査結果などから明らかになってきている交通網の整備時期による東北方言の共通語化の程度差を反映したものともみえる．

5.3 計量的手法を用いた方言分類・区画との関係

今回の音声資料を文字化したデータに基づく試行は，計量的手法を用いた方言分類・区画のいずれとも一致するような分類とはならなかった．部分的な重なりはうかがえるものの，全体として一致する結果とならなかったことは，一つには何らかの語形の一致という統一的な視点に基づくデータでなかったことも関係していそうである．

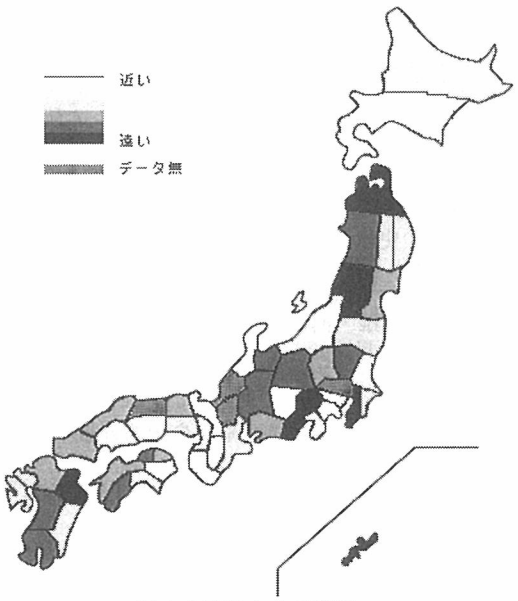

図 4: NHK からの距離

6．今後の課題

本研究は，談話文字化資料を用いていることから，アクセント・イントネーションのような音調に関する事象は欠落するものの，各地の方言を談話レベルから分類する試みであり，包括的な観点から日本語方言の区分・分類の可能性を持っている．例えば，東北における太平洋側と日本海側といった対立が確認できるなど，今日の共通語化の進行程度の違いを読み取ることができ，このような点は，本手法による効果といえるかもしれない．

複数の文字化担当者が同一規準に基づき文字化を行っている．しかし，53 の全ての地域に対して規準のゆらぎがないとはいえない．そこで，今後の課題として，文字化の機械化があげられる．

今回の実験では，Quartet 法で得られた樹状図はあまり良いものではなかったが，Quartet 法の特徴を活かし距離を直接的に反映するような改良が考えられる．本報告では，圧縮方法を bzip2 に限ったが，gzip や PPMZ などの圧縮方法でも実験を行い圧縮法の違いを検討する必要がある．得られた系統樹の優劣を客観的に評価する方法の開発，得られた系統樹の効果的な描画方法の開発なども今後の課題である．

謝辞

堀中幸司氏，関根康人氏，高野浩明氏をはじめとする日本大学文理学部谷研究室卒業生には，音声資料の文字化やソフトウェア開発に協力いただいた．谷研究室の荒堀和広君，福田奏君は現在ソフトウェア開発に携わっている．ここに記して謝意を表したい．

なおこの研究は私立大学学術高度化推進事業 学術フロンティア推進事業 2003-2007 の私学助成を得て行われている．

参考文献

[1] 井上 史雄：方言イメージ多変量解析による方言区画 『現代方言学の課題 第 1 巻社会的研究篇』 明治書院，1983.
[2] 井上 史雄・河西 秀早子：標準語形による方言区画 『計量国語学』(1982) 13-6.
[3] 加藤 正信：方言区画論『岩波講座日本語 11: 方言』岩波書店，1977.
[4] 河西 秀早子：標準語形の全国的分布 『言語生活』(1981) 354.
[5] 金田一 晴彦：私の方言区画 『日本の方言区画』 吉川弘文館，1954.
[6] 沢木 幹栄：方言地図データの活用―GAJ のデータによる地点のクラスター分析 『方言地理学の課題』（馬瀬 良雄監修）明治書院，2002.
[7] 東條 操：国語の方言区画 『日本方言学』吉川弘文館，1954.
[8] 馬瀬 良雄：方言意識と方言区画『日本の方言区画』, 1964 .
[9] 鑓水兼貴：活用形における共通語の分布パターン 『方言文法全国地図』第２・３集データの多変量解析 『計量国語学』(2007) 26-1.
[10] Bennett, C.H., P. Gács, M. Li, P.M.B. Vitányi, and W. Zurek: Information Distance, IEEE Transactions on Information Theory, 44:4 (1998), 1407–1423.
[11] Cilibrasi, R. and P.M.B. Vitányi: Clustering by Compression, IEEE Transactions on Information Theory, 51:4 (2005), 1523- 1545.
[12] Cilibrasi, R. and P.M.B. Vitányi: Clustering by Compression, *IEEE Transactions on Information Theory*, 51:4 (2005), 1523- 1545.
[13] Li, M., X. Chen, X. Li, B. Ma, and P.M.B. Vitányi: The similarity metric, IEEE Transactions on Information Theory, 50:12 (2004), 3250- 3264.
[14] Li, M. and P. Vitányi: An introduction to Kolmogorov complexity and its applications, 2nd edition, Springer-Verlag, 1997.
[15] 杉藤美代子監修・著：『日本語探検シリーズ 方言ももたろう』，富士通ビー・エス・シー.