語形のベクトル化による最適な言語地図の描画

近藤泰弘 [1]　持橋大地 [2]
[1] 青山学院大学　[2] 統計数理研究所・国立国語研究所
yhkondo@cl.aoyama.ac.jp　daichi@ism.ac.jp

概要

本論文では，国立国語研究所でかつて刊行された『方言文法全国地図』[1] の原データを用いて，新しい構想で言語地図を作成した．もとの地図における語形の標識はすべて人手で作成されていたが，本研究では方言の各語形をベクトル化し，そのベクトルのクラスタによって地図のアイコンの色を割り当てた．この際に色の割り当て方についても，カーネル整列法 [2] を用いて，色の類似度と語形ベクトルの類似度が最大になるような配色を行った．さらに元の地図の人手による語形分類データと，自動化された語形分類との一致度を調査することで，提案方法の妥当性を検証した．こうした統計的な手法により，非常に困難だとされてきた方言文法の言語地図を非常に見やすく，研究に使いやすい形にすることができた．

1　はじめに

言語地図とは，言語地理学 [3] [4] の中で発展してきたものであり，ある単語や音韻の，方言や特定の言語による語形・音形をその出現した地点の地図上に色や形でアイコン化して配置したものである．たとえば，「かたつむり（蝸牛）」の言語地図では，方言形「マイマイ」の類を○，「デンデンムシ」の類を△のように区別して表示することで，その分布を可視化し，方言形の変異の歴史の考察を行うことなどが可能である．日本では，その初期から国立国語研究所による諸研究が中心となっており，『日本言語地図』[5] がその代表的業績である．その後の研究の中で『方言文法全国地図』[1] は，文法現象に焦点を当てたものである．

言語地図作りはその手順を明確にしにくい作業であったが，我々の以前の研究（論文 [6] および 論文 [7]）において，単語の音の分布の Bag of Words (BoW) や FastText によって語形をベクトル化し，それをクラスタ化することで語形を分類し，それにアイコンの色をあてはめることによって，言語地図作りを自動的に行うことを提案した．それにより，『日本言語地図』や『新日本方言地図』の地図を再構成し，また，その地図を分類することを試みた．

今回の論文で提案するものは，その手法をさらに洗練させ，ベクトルの生成，次元圧縮，アイコン色の割り当てなども根本的に設計し直したものである．これにより，見やすさ，研究での使いやすさが格段に向上する．

2　提案手法

2.1　ベクトル化の方法

今回は語形のベクトル化としては，それぞれの方言の回答語形（音声記号）を分解して，文字単位のユニグラムとバイグラムを取得し[1)]，それを BoW でベクトル化したものを回答語形の語形ベクトルとした．次元数は，対象地図の語形に出てくる音声記号の種類によって異なるが，約 450 次元程度である．

2.2　クラスタ化の方法

次に，その 450 次元のベクトルを t-SNE で 2 次元[2)]に圧縮してから，K 平均法でクラスタ化した．この際のクラスタ数の選択については，4 節で説明する．各クラスタを，カーネル整列法 [2] によって人間の直感的な一致度に合わせるように Python の `Folium` の標準のアイコン色セットに対応させる．

カーネル整列法 (Kernelized Sorting) [2] は，機械学習のカーネル法の分野で提案された方法であり，別のドメインに属する同数のデータ $\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathcal{X}$ と $\mathbf{y}_1, \cdots, \mathbf{y}_N \in \mathcal{Y}$ の間で，カーネル的な相互情報量である HSIC が最大となるような対応づけ (例えば，自

1)　この際に，バイグラムでは文頭記号 (^) および文末記号 ($) を含めている，よって，たとえば語形 foo は `^foo$` となり，バイグラム `^f, fo, oo, o$` に分解される，

2)　3 次元以上にすることも可能であるが，特定の概念を表す語形だけを扱っていることから，最低限の 2 次元でも広がりを充分に表現できるため，本研究では 2 次元に圧縮した．

然言語処理の場合は異言語間の対訳関係) を与える写像，すなわち順列 $\pi = (1, \cdots, N) \mapsto (1, \cdots, N)$ を求める方法である．与えられたカーネルの下で，$\mathbf{x}_1, \cdots, \mathbf{x}_N$ および $\mathbf{y}_1, \cdots, \mathbf{y}_N$ で計算したグラム行列を $\mathbf{K}, \mathbf{L}$ とおくと，中心化行列を $H_{ij} = \delta_{ij} - 1/N$ として，問題は中心化した $\bar{K} = \mathbf{HKH}$ と $\bar{L} = \mathbf{HLH}$ について

$$\mathrm{tr}\left(\bar{\mathbf{K}}\pi^T\bar{\mathbf{L}}\pi\right) \tag{1}$$

を最大とする順列 π を探すことに帰着する．直感的には，これは (中心化した) 二つのグラム行列の間の類似度がもっとも高くなるような対応づけ π を求めていることに相当する[3]．グラム行列の計算には，本研究では t-SNE で求めた 2 次元の座標および，`Folium` が内部に持つ標準的なアイコン色の RGB 値の間のガウスカーネルを用いた．

式 (1) の最適化には様々な最適化法によって局所解を求めることもできるが，今回は色の数が最大 10 程度と少ないため，全探索を行っても数秒から最大でも数分で全体最適解を求められる．こうして得られた最適な色割り当て π を用い，語形を色マーカーとして地図上に配置した．[4]

3 言語地図の描画

3.1 語形ベクトルの散布図

図 1（先生が来られた）と 図 2（行かなければならない）に，クラスタ数 9 の語形ベクトルのサンプルの散布図を示した．このように，特に地図 2 に対応するもの（図 1）の方は非常にクラスタの形がはっきりしており，後で述べるように，人手による地図との一致率の高さを裏付けている．

また，クラスタのドット色については，どちらの散布図でも，近い部分には近い色相の色が配置されているのがわかる．これはカーネル整列法によるもので，地図のアイコン色にもそのまま使われており，後述のように地図の見やすさに繋がっている．

地図 206（図 2) ではクラスタの明瞭度はやや落ちるが，十分に実用的なクラスタとなっている．

3.2 言語地図

そして地図（図 3 と図 4) は，ここに示す通りである．このように，それぞれの語形ごとの地域区分が

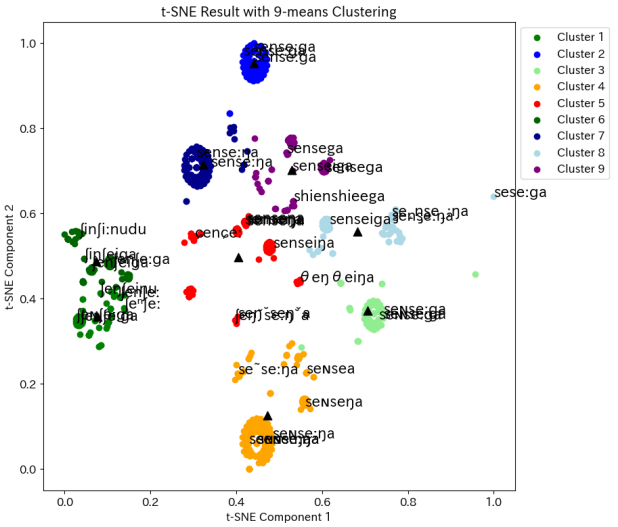

図 1 地図 2 の語形ベクトルの散布図.

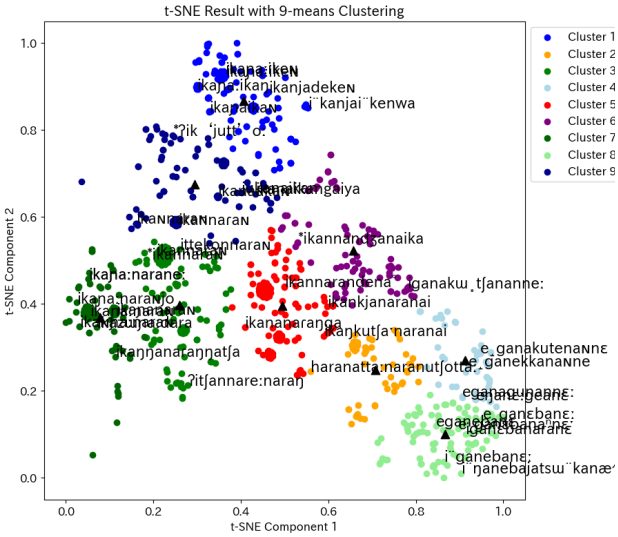

図 2 地図 206 の語形ベクトルの散布図.

非常に明瞭に観察できる．ここに示したのは固定された状態だが，folium で実際に作成された地図はインタラクティブに縮小・拡大・移動が可能であるので，より使いやすい．

この地図の視認性がよい原因は，アイコンの色付けの工夫にある．語形が近いものは近い色が付与されるようになっているので，一見して分布の特徴を判別しやすいのである．

付録の，図 7 と，図 8 は，元の『方言文法全国地図』からの分布の引用であるが，分布の傾向性を捉えることがやや難しいのがわかる．

4 クラスタ数の最適化について

K 平均法でクラスタ化を行う時に，クラスタ数をいくつにするかは重要な問題である．一般的には，

3) 一般に，行列 $\mathbf{X}, \mathbf{Y}$ について $\mathrm{tr}\,(\mathbf{X}^T\mathbf{Y})$ は，ベクトルの場合と同様に $\mathbf{X}$ と $\mathbf{Y}$ の「内積」とみなすことができる．

4) `Folium` の標準のアイコンはやや大きく，言語地図が見づらくなるため，任意の HTML の div 要素をアイコンとして使用できる `DivIcon` の機能を使って描画している．

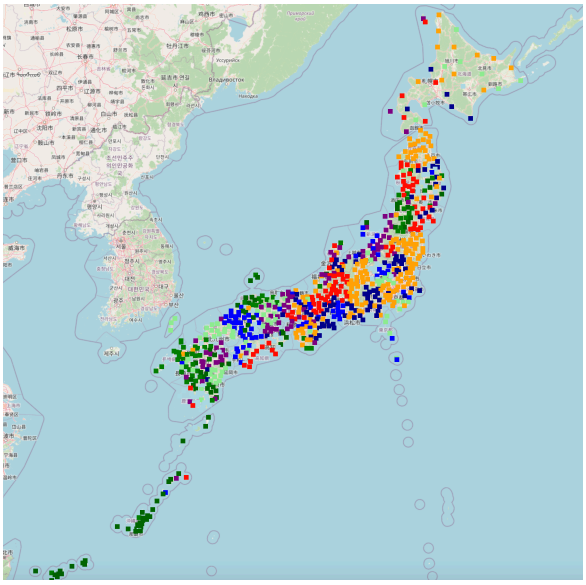

図 3　提案手法による地図 2 のインタラクティブ言語地図. Python の `Folium` で描画されているため, 任意に拡大や縮小も行うことができる.

エルボー法や，シルエットスコアによるクラスタリングの性能評価が行われることが多いが，よく指摘されるようにこれらの方法は，対象とする研究の内容とマッチしないことがある．

そこで今回は，クラスタリングの性能評価として，分析内容と直接に関係する方法を用いた．

4.1　一致率とエントロピーの計算

まず，計算方法であるが，次のように，クラスタ内で最も頻出した見出しコードの出現数を，クラスタ内の総見出し語データ数で割ったものを，クラスタと見出しコードの一致率とした．i は，そのクラスタ内の見出し語コードのインデクスである．

$$\text{match\_rate} = \frac{\max(\text{count}_i)}{\sum_{i=1}^{n} \text{count}_i} \tag{2}$$

また，クラスタ内での見出しコードのばらつきをエントロピーで表した．ここで p_i は，下の式のように，各クラスタ内の見出しコード別の出現割合である．なお，以下では，スペースの関係上，表では一致率のみ掲載した箇所がある．

$$\text{entropy} = -\sum_{i=1}^{n} p_i \log_2(p_i); \quad p_i = \frac{\text{count}_i}{\sum_{j=1}^{n} \text{count}_j} \tag{3}$$

4.2　計算対象の地図

元データには，合計 350 図の地図があるが，元方言地図の語形分類（見出し語の選定）において，大きく分けて次の 2 種類の語形の選定方法がある．

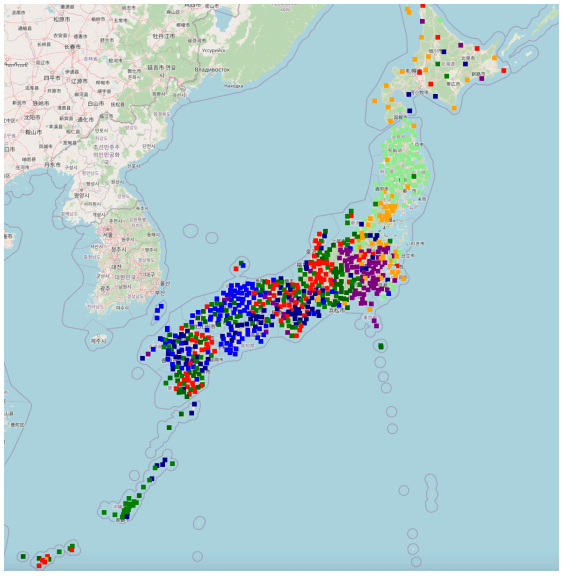

図 4　提案手法による地図 206 の言語地図.

(1) 2 図（表 1)「先生が来られた」のように，「先生が来られた」に相当する文全体ではなく，その中の主格助詞「の／が」の部分など，単語の形だけに注目して見出し語（たとえば，[ga]）を決めてその語形の分布で地図を作るタイプ

(2) 206 図 (表 2)「行かなければならない」）のように，その中の「行かなければ」の部分など，長めの句を見出し語（たとえば，([ika, iga, itteko, eka) njaa -] ）とするタイプ

このように，元方言地図には，内容に応じて 2 種類の方針が混在しているが，今回の論文でも，それぞれの方針に従ってベクトル化した．しかし，結果としては，この 2 種類は，かなり違った状況を示すので，以下にその実際を示しておく．

4.3　一致率の実際

最初の表 1 は，2 図（「先生が来られた」）である．ここでは，格助詞の差異に注目して，見出し語がまとめられている．これは，日本語の歴史の中で，尊敬語かどうかで主語の格助詞が異なるという研究史があるため，それが方言に反映されているかどうかの地図ということになる．クラスタ番号の 1,2 は，一致率が非常に高く，自動判定と，元地図の人間による判定の見出しコードがよく一致していることがわかる．語形が格助詞だけということで, 判定が容易であることが効いていると思われる．

次の表 2 は，地図の 206 図（「行かなければならない」）である．こちらは最もよいところで，0.63 であり，2 図よりも一致率が低い．これは，先に述べた

表 1　地図 2 (8 クラスタ) の一致率

クラスタ	データ数	最頻見出しコード	最頻見出し名	一致率
0	67	1	ga	0.358
1	99	1	ga	0.980
2	77	1	ga	0.987
3	170	3	ŋa	0.776
4	93	3	ŋa	0.677
5	82	1	ga	0.682
6	112	3	ŋa	0.705
7	107	1	ga	0.728
平均				0.737

ように，対象となる部分が長いため，表現が複雑で，見出し語の分類が難しいことによる．従って，この一致率の低さは，機械学習側に問題がある場合もあるが，人間による分類に問題がある場合もある．

表 2　地図 206（8 クラスタ）の一致率

クラスタ	データ数	最頻見出しコード	最頻見出し名	一致率
0	144	85	(ika, iga, itteko, eka) njaa −	0.354
1	74	1	(ika, iga) nakereba −	0.176
2	97	132	(ika, iga, iŋa, itteko, ega, eŋa, juka, Nŋa) N −	0.278
3	102	132	(ika, iga, iŋa, itteko, ega, eŋa, juka, Nŋa) N −	0.147
4	114	132	(ika, iga, iŋa, itteko, ega, eŋa, juka, Nŋa) N −	0.211
5	124	85	(ika, iga, itteko, eka) njaa −	0.258
6	74	51	(ika, iga, iŋa, itteko, eka, ega, eteko) neba −	0.635
7	77	50	(ika, iga, iŋa, uŋa, ega, eŋa) n ɛ ba −	0.273
平均				0.291

4.4　一致率からみたクラスタ数最適化

図 5 に示したグラフは，第 2 図のクラスタ数と語形の一致率の平均，およびエントロピーを示したものである．このようにクラスタ数が 8 から 9 あたりまでは一致率が上がり，またエントロピーが下がるが，それ以上は横ばい傾向にある．したがって，このグラフを見る限り，適切なクラスタ数としては 8 あるいは 9 と考える．

次のグラフ 図 6 では，先の図 5 とは異なり，一致

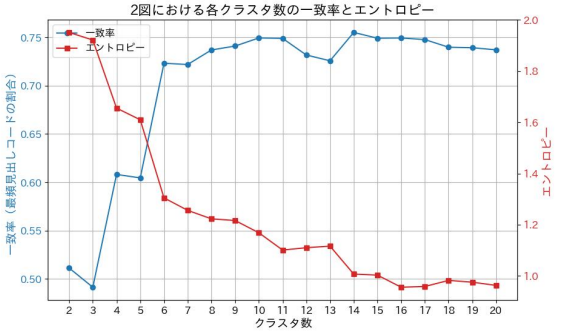

図 5　地図 2 のクラスタ数と一致率・エントロピー

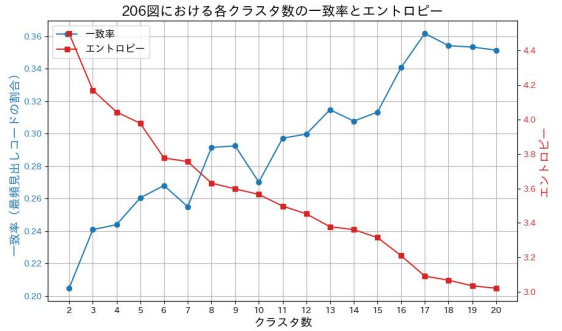

図 6　地図 206 のクラスタ数と一致率・エントロピー

率もエントロピーも大きな屈曲はなく，クラスタ数によって微増・微減である．これは形が複雑であるため，手動でのまとめ方と，自動化されたまとめ方とに食い違いが出ているためであるが，先の地図描画で見たように，自動の方が「間違っている」とも言いにくい部分もある．

以上，語形の両様のパターンから，単純な語形の場合を優先して，クラスタ数は 8 から 9 を適正な数値と考えたい．

5　まとめ

従来の言語地図作成，とくに方言文法の地図においては，その中で歴史的な変遷を明らかにするようなアイコン化が重要であるという立場（論文 [8] 等）や，客観的に語形の分布を表現することが重要であるという立場（論文 [9] 等）があり，意見が分かれていた．それは，ここまで見てきたように，方言文法という枠組みで調査を行った場合，非常に多様な変種を持った表現形が得られるため，それを統一的にアイコン化し，地図化することが困難であったことに起因する部分がある．

今回の新提案では，方言文法地図というその描画が困難だとされている種類の言語地図に対する処理方法を提案し，従来にない研究手段を提供できた．

謝辞

本研究には，国立国語研究所ウェブサイト「方言研究の部屋」（https://www2.ninjal.ac.jp/hogen/dp/dp_index.html）内の，「方言文法全国地図」全データ（1〜6 集）および PDF 版『方言文法全国地図』（1〜6 集）を用いることができた．管理者の大西拓一郎氏に感謝申し上げる．また，原データを作成された，小林隆氏を始めとする皆さまにも感謝申し上げたい．

参考文献

[1] 国立国語研究所. 方言文法全国地図 (1)〜(6). 国立国語研究所, 1989–2006.
[2] Novi Quadrianto, Le Song, and Alex Smola. Kernelized Sorting. In **Advances in Neural Information Processing Systems 21 (NIPS 2008)**, 2008.
[3] 柴田武. 言語地理学の方法. 筑摩書房, 1969.
[4] 大西拓一郎. ことばの地理学: 方言はなぜそこにあるのか. 大修館書店, 2016.
[5] 国立国語研究所. 日本言語地図 (1)〜(6). 国立国語研究所, 1966–1974.
[6] 近藤泰弘. 単語音素のベクトル化による言語地図作成. 言語処理学会第 29 回年次大会 B10-2, 2023.
[7] 近藤泰弘, 持橋大地. 語形の分布状況のベクトル化による言語地図の分類方法. 言語処理学会第 30 回年次大会 D5-1, 2024.
[8] 柴田武. 書評　国立国語研究所編『方言文法全国地図 1』. 『国語学』, No. 162, 1990.
[9] 小林隆. 方言地図の方法について―柴田武氏「書評　国立国語研究所編『方言文法全国地図 1』を読んでー. 『国語学』, No. 163, 1990.

A　参考資料・『方言文法全国地図』からの引用

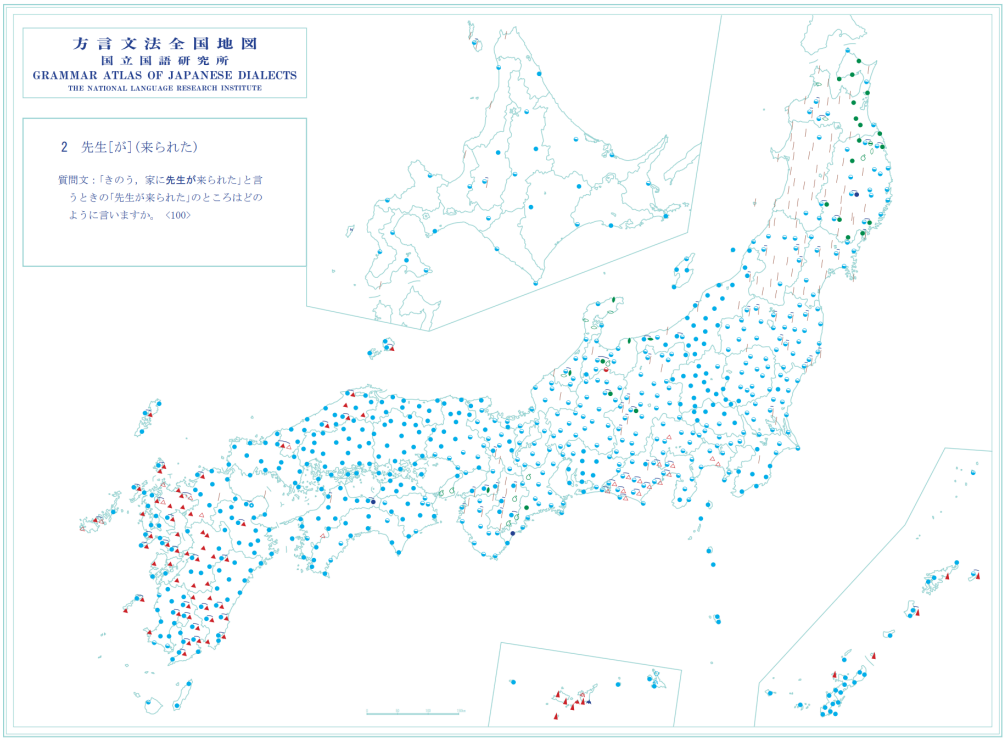

図 7　『方言文法全国地図』における地図 2 の言語地図

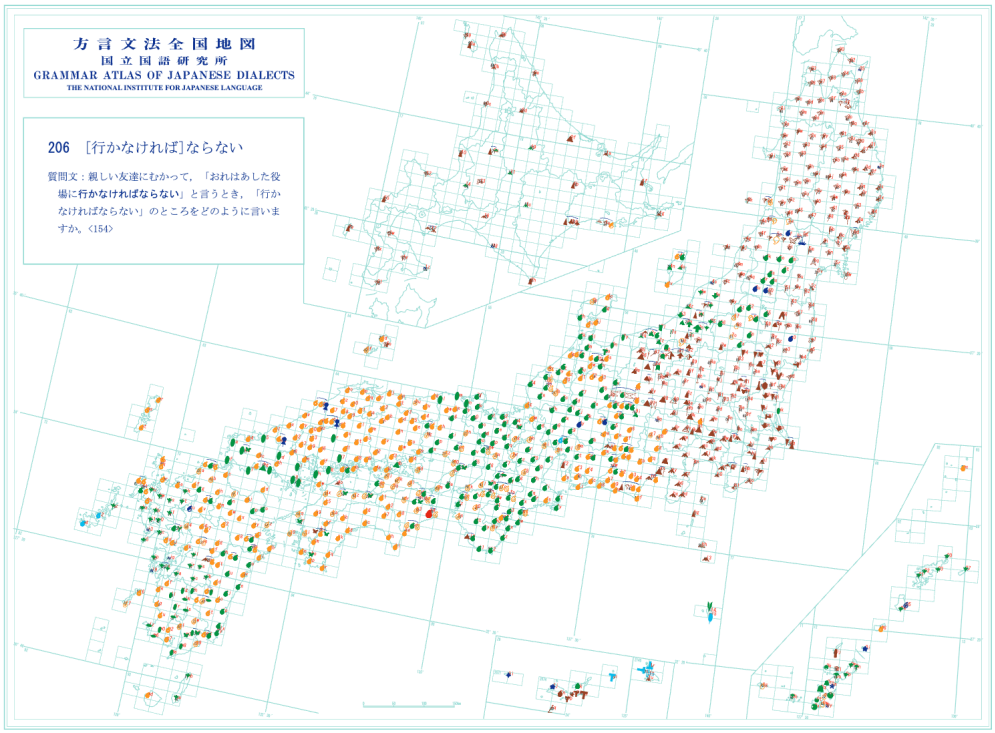

図 8　『方言文法全国地図』における地図 206 の言語地図