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

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
yhkondo@cl.aoyama.ac.jp
daichi@ism.ac.jp

概要

本研究では, 異なる語に対する「言語地図」を機械学習によって自動分類する手法を提案した. 従来「言語地図」をコンピュータで自動作成する方法としては, 発表者の考案した単語語形音素の単語集合表現（BoW）による方法があったが, 今回の研究では, 新たに, 語形の地域による共起情報を FastText でベクトル化し, 語形情報と地理情報を埋め込んだ言語地図を様々な語について作成した. そして, その言語地図の持つベクトル情報を階層的クラスタに分類することで, 言語地図そのものを分類することを行った. これによって, 基本語・漢語などの分布と, 文法要素の分布とでは, 異なった分布の様相を示すことが明らかになった.

1　はじめに

社会言語学の研究の中で，方言語形を地点ごとに配置する，いわゆる「言語地図」を作ることが行われている．近年は，このような言語地理学の分野でも，GIS の手法を用いて，電子的に表現する方法が一般的になってきた．発表者も以前に，語形をベクトル化し，それをクラスター分析することで，語形をグループ化することを試みた [1]．今回の発表では，ベクトル化の方法を変更し，語形情報だけでなく，相互の位置も含めてベクトル化する方法を考案した．さらに，それによって作られた地図情報をさらにクラスタ分類して，分布状況のパターンを発見することを試みた．従来も，言語地図を処理することで語形の距離から，東日本・西日本・九州などのように方言区画を導き出す研究は多かったが，分布のパターンという考え方で研究されたものはなかった．以下に、その方法を述べていく。

なお，研究対象としたのは，国立国語研究所で企画・開発された『全国方言分布調査』（FPJD) のデータベース [2] である．

2　関連研究

日本語の言語地図をコンピュータで作成する試みとして，メインフレームの時代の荻野綱男の GLAPS[3] ,PC になってからの福嶋（尾崎）秩子の SEAL [4] などがある．大西拓一郎も，Adobe イラストレータや GIS 系のアプリを駆使して言語地図を発表している [5].

語形（音形）のレーベンシュタイン距離を用いて，方言間の差異を計測した研究も多い [1] [6]. 日本語においても，方言分類を論じる場合に使われた．最近では，Heeringa・Inoue による方言区画の研究がある [7].

また，先に述べたように，発表者も，語形（音素）の BoW を用いたベクトル化によって言語地図を自動作成する研究を行っている [8].

3　提案手法

1 節で述べた通り，本研究では，単語の語形の音節（モーラ・仮名単位）を基本データとして用いた．そして，位置情報を反映させるために，各語形の採取地点の緯度・経度を 1 次元に圧縮し，その上に語形を配置した．これを一種のテキストであると考え，FastText によって語形を 100 次元にベクトル化した．これにより，地域で隣接する語形は近いベクトルを持つことになる．この 100 次元ベクトルを PCA で 2 次元に圧縮し，K-Means で分類することで，語形の分類情報を得て，これを用いて言語地図を作成した．さらに，この各語形の持つ 100 次元ベクトルの平均値をとり，さらに，地図内のすべての語形の持つ上記平均値をさらに平均して，ひとつの地図のもつベクトルの特徴量と考えた．この特徴量を階層化クラスタリングして，地図の分類を行った．

3.1　モデルと学習方法

各語形の採取地点の緯度・経度を 1 次元に展開し、それぞれの地点の採取語形、それぞれ 2000 個を順に配列する。

[’ カサブタ’, ’ カサブタ’, ’ カサブタ’, ’ カサビタ’, ’ カサブタ’, ’ カサブタ’, ’ カサブタ’, ’ カサブタ’, ’ カサブタ’, ’ カサビタ’, ’ カサ’, ’ カサブタ’, ’ カサベタ’, ’ カサブタ’, ’ カサビタ’, ’ カサブダ’, ’ ガンベ’, ’ カサブタ’, ’ カタフサ’, ’ カサビタ’, ’ チノカタマリ’, ’ カサブタ’, ’ カサピタ’, ’ カサフタ’, ’ カタプタ’, ’ カサ’, ’ カサブタ’, ’ カサブタ’, ’ カサプタ’, ’ ガンベ’, ’ カサ’, ’ カサ’, ’ カサ’, ’ カサデキタ’,（以下略）]

この語形を順に並んだテキストであると考え，FastText のモデルに学習させた．ハイパーパラメータとしては

```
vector_size = 10, window = 10 , min_n = 3 ,
max_n = 6, sg = 0
```

とした．ここで作成されたモデルに，再度語形を入力して，各語形のベクトルを得て，得られた 100 次元の語形ベクトルを PCA によって 2 次元とし，さらに K-Means でクラスタに分けて語形を分類し，それぞれの語形に色を定めた．その色を用いて，2 次元のベクトルの散布図および、地図上に語形アイコンを色わけして配置した．

Python のライブラリの gensim の FastText および scikit-learn の PCA を用いた．また，PCA によって次元圧縮し，いくつかの次元のデータを作成した後，2 次元にすることを決定し，またその 2 次元データを K-Means で学習させることで，クラスタに分類したが，クラスタ数の最適化については，語形分類が実用的にできるかどうかを主として考え，エルボー法などで結果を裏付けた．こうしてベクトル化された地図の階層クラスタリングには，scipy.cluster.hierarchy のライブラリを用いた．

4　実験と結果

4.1　実験設定

提案手法の評価には，代表的な日本語方言の言語地図用のデータセットである「全国方言分布調査データベース（FPJD)」（国立国語研究所編）[2] を用いた．公開されている「言語地図」データは Excel 形式で，語形（カタカナ），地点番号，緯度，経度，などからなっている．

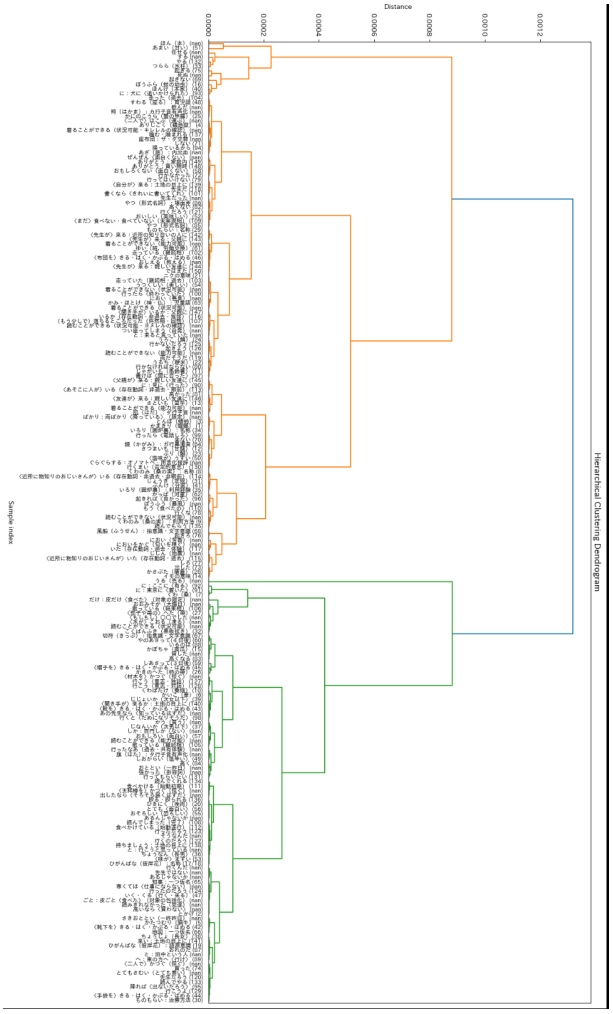

図 1　地図の分類（全体）

「全国方言分布調査」のすべての語形についてではなく，その一部について公開されているが，今回は公開されている部分すべてについての約 200 項目のデータを用いた。それぞれ，約 2000 件の地点数である．処理プログラムの実装には Python 3.8.16 と scikit-learn 1.0. を用いた．地図へのプロットには動的地図を作成できるため，folium 0.14.0 を用いた．

4.2　結果

地図の持つ平均化したベクトルを階層化クラスタリングで分類した結果全体を図 1 に示した．その最上部と最下部を拡大したものが，図 2，図 3 である．

100 次元の語形ベクトルから平均値をとってそれを分類するとは，概略，2 次元の散布図を似た形で分類することと等しい．例えば，クラスタ分類図の上部（図 2）にある語である，図 4 ・図 5 の「甘い」の散布図では，重要な語形が独立してマッピングされていて，中央にその他の語形が位置する．このよ

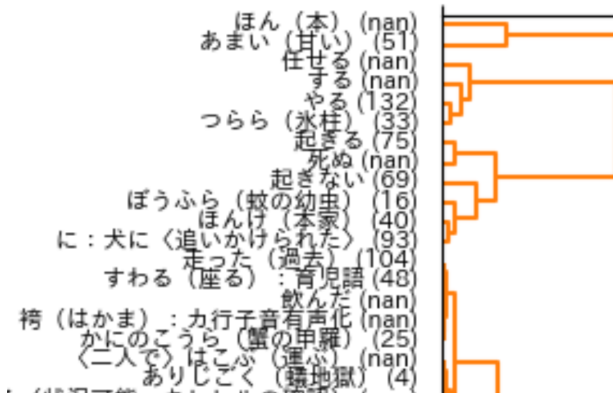

図 2　地図の分類（上部）

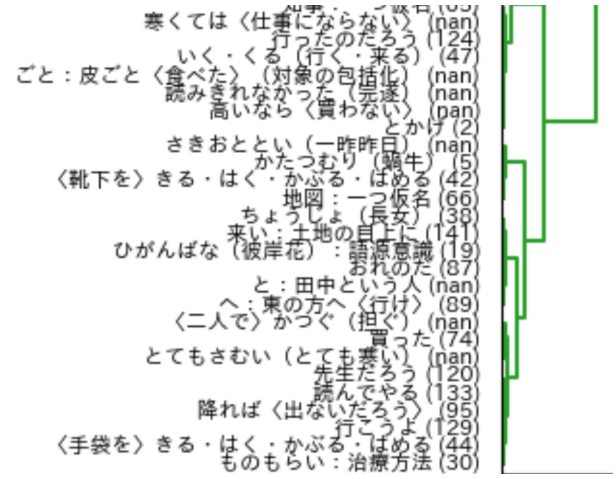

図 3　地図の分類 (下部)

うな散布図が地図になると，主要語形が西日本四国や九州に地域ごとにブロックとして配置される．これに対して，クラスタ分類図の下部（図 3）にある語である図 6 ・図 7 の「行こうよ」では，全体に語形がばらばらに散布図に存在し，この場合は，地図の上でも，各語形がまだらに分布することになる．

また言語の内容面からは，大きく見て，ブロック状の分布の地図には，「本」（図 8）「本家」などの漢語・「あまい」「つらら」などの基本語が多く存在している．それに対して，まだら状の分布には，「行こうよ」「先生だろう」（図 9）などの文法要素を含む語形が多い．漢語は渡来が新しいことによる固定化と思われる．文法要素の語が比較的ばらばらに分布することの原因はつかみにくいが，文法要素が様々な付加的な接尾辞により構成されているため，地域差が大きいことが原因ではないかと考えられる．これについては，さらにこの方法を洗練させて，より詳しい分析が可能になるように改良を図りたい．

なお，完全には分類できず，一部例外的な地図も存在するが，それについては，今後の改良が必要である．

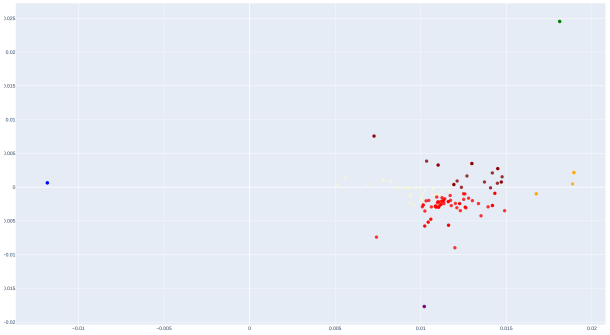

図 4　「甘い」の散布図

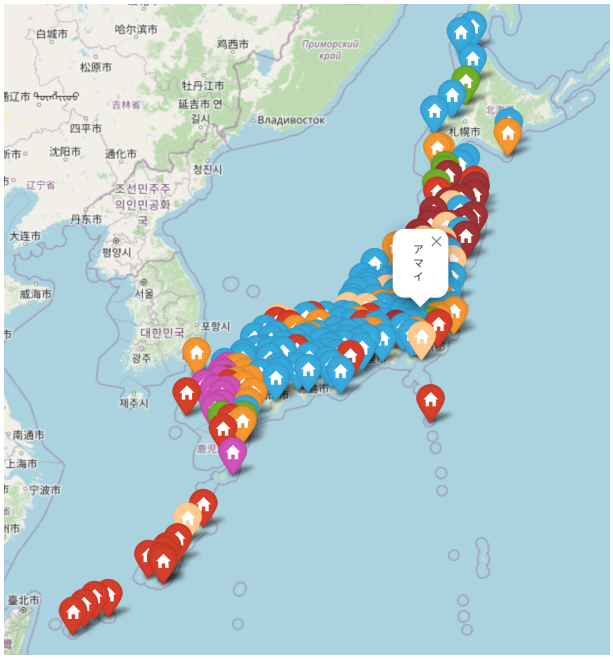

図 5　「甘い」（ブロック状分布）

5　結論

本研究の方法によると，言語地図自体を地域情報を含めてベクトル化して扱うことで，その分布の「模様」を分類することが可能であることを示した．これにより，分布パターンと語の素性との関係という新しい研究分野が開拓可能である．

今後，ベクトル化の方法や，地図の分類方法をさらに探求し，より精度の高い分類ができる方法を試みたい．特に南北型の分布や，周圏分布といった特徴までも分類可能になるように方法を改善していきたい．

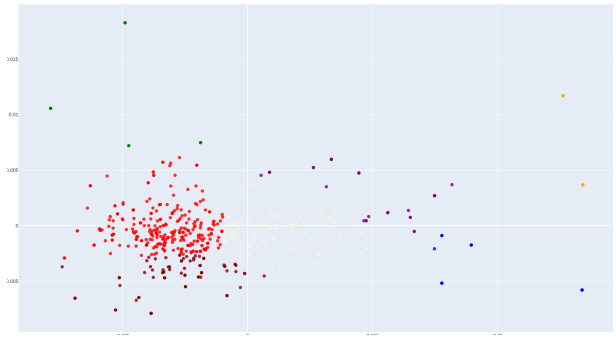

図 6　「行こうよ」の散布図

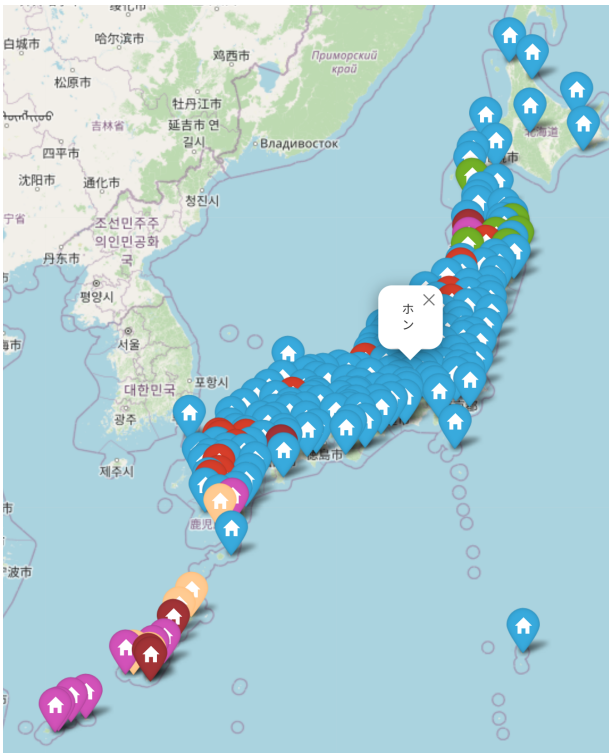

図 8　「本」（ブロック状分布）

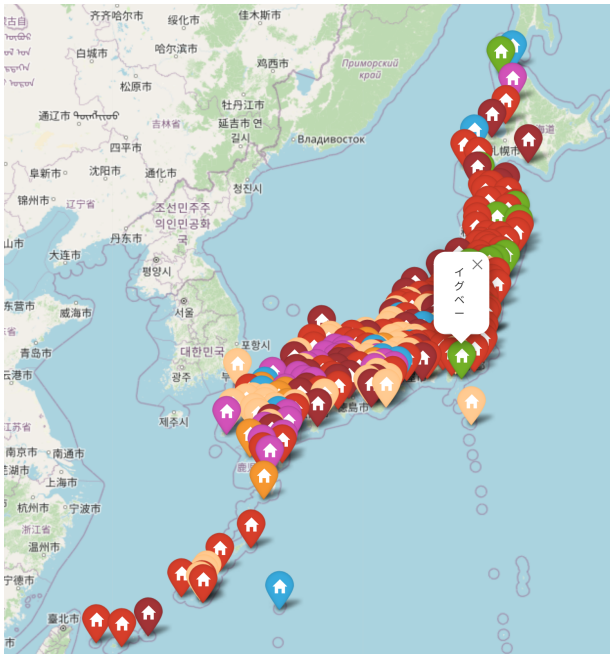

図 7　「行こうよ」（まだら状分布）

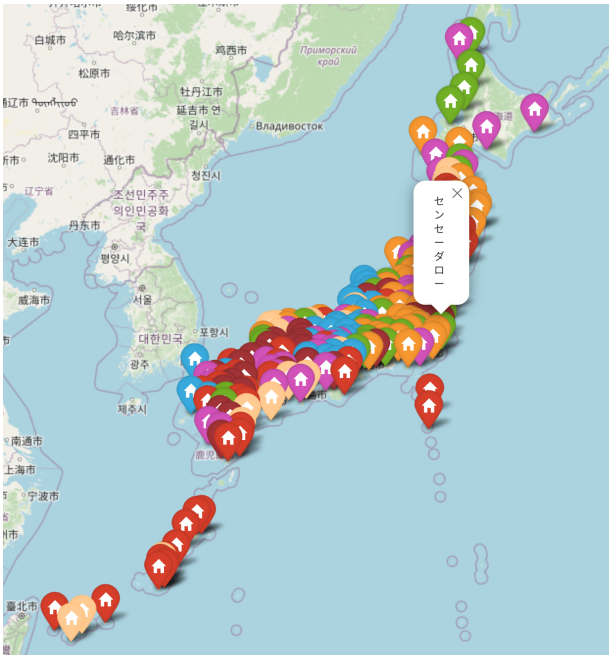

図 9　「先生だろう」（まだら状分布）

謝辞

本研究の評価データセットに用いた「『全国方言分布調査』データベース」(FPJD) を開発した，国立国語研究所，サイトの保守にあたっている 熊谷康雄氏，大西拓一郎氏，また調査関係者の皆様に感謝申し上げます.

参考文献

[1] Wolfgang Vierreck. The computer developed linguistic atlas of england, volumes 1 (1991) and 2 (1997). *International Computer Archive of Modern English: ICAME journal*, Vol. 1997, No. 21, pp. 79–90, 2015.

[2] 国立国語研究所. 『全国方言文法調査 (fpjd)』, 2018. https://www2.ninjal.ac.jp/hogen/dp/fpjd/fpjd_index.html.

[3] 荻野綱男. コンピュータ言語地理学. 言語研究, Vol. 1978, No. 74, pp. 83–96, 1978.

[4] 福嶋秩子. 言語地理学のへや, 2022. https://www.unii.ac.jp/chitsuko/inet/lg7.html.

[5] 大西拓一郎. 言語地図作成の電算化. 日本語学, Vol. 21, No. 11, 2002.

[6] 鑓水兼貴. 共通語化過程の計量的分析. 東京外国語大学博士論文, 2009.

[7] Heeringa Wilbert, INOUE Fumio. Exploring the japanese dialect dialectometorically – division and continuity –. *Studies in Geolinguistics*, No. 3, pp. 1–44, 2023.

[8] 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 回年次大会発表論文集, pp. 2381–2385, 2023.