語彙・語尾マーカーに基づく関西方言の地域推定

藤田フェリペ　高田秀志
立命館大学情報理工研究科
is0596kh@ed.ritsumei.ac.jp　htakada@is.ritsumei.ac.jp

概要

関西方言は一般に「関西弁」として一括りにされがちであるが，実際には話者の出身地域に応じて，語彙・語尾・言い回しといった言語特徴が隣接地域間で徐々に変化することがある．本研究では，このような地域差を捉えることを目的として，関西圏の短単位話者テキストから言語的特徴量を設計し，4つの地域ラベルに対する地域推定手法を構築する．分類の結果，正解率 0.783（F1 値 0.706）を達成し，終助詞や「ねん」「やん」などの語尾マーカーが識別に大きく寄与することを確認した．

1　はじめに

方言・言語変種を対象とした自然言語処理は，類似言語・変種・方言を横断的に扱う研究領域として発展しており，識別（どの方言か），正規化（標準語寄せ），下流タスクへの適応（翻訳・対話・音声認識）など幅広い課題設定が整理されている [1, 2]．一方で，言語技術は標準変種・多数派の言語実践に最適化されやすく，方言話者が入力を「標準語化」する圧力を受けたり，特定集団に誤認識が生じたりする点が批判的に指摘されている [3]．こうした観点から，方言処理は単なる性能向上だけではなく，どの変種がどの程度不利になるかを測り，原因を説明可能にする評価設計が求められる．

方言間の性能格差を測る取り組みとして，英語の自然言語理解（NLU）における方言格差を体系的に扱う枠組みが提案されており [4]，さらに会話理解に基づいて方言頑健性を評価する研究も報告されている [5]．これらは，「平均的に高性能」なモデルであっても，方言入力に対して誤りが偏って現れる可能性を示唆し，方言差を定量化する指標や可視化が不可欠であることを示している．

日本語においても，方言資源の整備と統計的分析は重要である．音声対訳コーパスの構築は，方言音声資源不足の課題に対して，標準語との対応づけを含む実用的な基盤を提供する [6, 7]．また，関西方言に関しては，既存コーパスを短単位へ分割し形態論情報を付与することで，計算処理に適した形で資源化する試みが報告されている [8]．

本研究の目的は，

- 解釈可能な語彙・語尾マーカーと形態素パターンの分布にもとづいて地域推定を行い，
- 重要特徴と誤り傾向を分析し
- 地域平均特徴量から地域間距離を導出して可視化することで，地域差の「根拠」と「近さ／連続性」を同時に示す

ことである．方言は連続体として捉える必要があるという問題意識のもと，モデルを方言へ適応させるだけでなく，方言側をモデルに寄せるという双方向の適応を扱う枠組みも提案されているが [9]，軽量で説明可能な特徴量により，関西圏内部の地域差を定量化する足場を築くことを目指す．

2　関連研究

方言・言語変種処理は，類似言語・方言の識別や正規化から，下流タスクへの適応までを含む広い研究分野であり，研究動向・評価の枠組み・データ収集上の論点が整理されている [1, 10]．しかし，方言を扱う際には，単に平均性能を上げるだけでなく，どの変種が不利になりやすいか，その原因はどこにあるかを説明可能にすることが重要である．言語技術の偏りを「権力」として捉え直す批判的サーベイは，データと評価の前提が社会的要因と結びつく点を強調しており [3]，方言研究における資源整備・評価設計の透明性が求められる．

方言差を「格差」として定量化する方向として，方言間性能差を明示的に扱い，標準語を中心とした変動の評価が方言話者に不利をもたらし得ることを示した [4]．会話理解に基づく方言頑健性評価は，対話文脈や話者交替を含む状況で誤りが増幅される可能性を示し，方言入力の扱いを「単文分類」

から「会話理解」へ拡張する必要性を提起している [5, 11]．我々の研究は分類タスクを採用するが，後段で LLM ／対話へ接続することを見据え，まずは地域差の根拠となる表現を説明可能に抽出・可視化することを重視する．

方言識別の方法論自体は，多言語で成熟している．たとえばアラビア語方言分類では，深層学習モデルにより高精度な識別が可能であることが報告されており [12]，大規模データと言語固有の表記揺れを前提とした設計が検討されてきた．これは，高性能化の観点だけでなく，どの特徴が識別に寄与しているかを分析する重要性も示唆する．一方，方言を単なる離散ラベルとして扱うのではなく，**連続体**として捉える観点から，モデルを方言へ適応させるだけでなく方言をモデルへ寄せるという双方向適応を通じて言語連続体をモデル化する枠組みが提案されている [9, 13, 14]．この観点では，地域間の「近さ」を距離として扱い，可視化する設計が重要になる．

3　提案手法

本研究では，深層表現（埋め込み）に依存しない軽量な設計として，話者ごとのテキストから **解釈可能な表層特徴** のみを抽出し，関西方言の下位地域（京都・大阪・兵庫・周辺県）を推定する枠組みを提案する．一連の処理フローは図 1 に示す．まず，入力データである各話者の発話テキストに対して MeCab を用いた形態素解析を行い，語彙マーカーおよび形態素パターンの正規化頻度を特徴量として抽出する．これらの語彙マーカーおよび形態素パターンを連結した解釈可能特徴ベクトルを分類器（地域ラベル分類器）への入力とする．学習時にはクラス不均衡を緩和するためオーバーサンプリングを適用しつつ，ランダムフォレストにより地域分類を行う．この方法により，高性能な深層モデルに依存することなく，各特徴量の寄与を明示的に分析可能な分類を実現する．さらに，地域平均特徴ベクトルを集約してユークリッド距離を算出し，距離表および多次元尺度構成法（MDS）により地域間の近さを可視化する．

3.1　マーカー特徴量（語彙・語尾）

方言の地域性は，典型語彙や語尾の偏りとして現れやすい．そこで，認識に使う言語パターンを「マーカー集合」として定義し，テキスト中の出現頻度を正規化して特徴量化する．マーカーは次の 2 群である：地域性が強い語彙や関西弁を代表する語彙（例：**て　はる，ほんま，めっちゃ，ちゃう**），および終助詞群（**わ，で，な，やん**）．終助詞は関西圏全体で頻出である一方，**なやわ**等は地域差が出やすいことがあり，方言分類で頻度にもとづく特徴が有効であるという知見とも整合的である [15]．

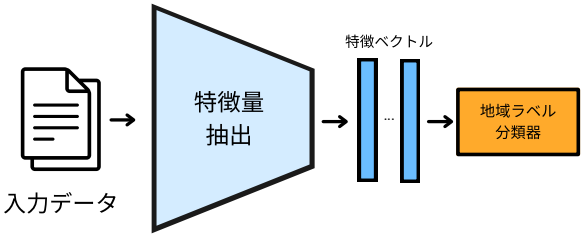

図 1　提案手法の概要

各マーカー m について，話者テキスト x 中の出現回数を $\mathrm{count}(x, m)$，トークン数を $|x|$ とし，正規化頻度を

$$f_m(x) = \frac{\mathrm{count}(x, m)}{|x| + \epsilon}$$

で定義する（ϵ はゼロ除算回避）．

3.2　形態素パターン特徴量

表 1 に，本研究で用いた具体的な形態素パターンを示す．

表 1　形態素パターン（形態素列の連接による検出）

特徴名	意図	表現例
「**(ん ／ の) ＋ や**」	説明・断定の言い回し	「そう**なん や**」
「**(ん ／ の) ＋ ねん**」	説明・強調（ねん系）	「見**てん ねん**」
「**て＋はった**」	過去・尊敬の定型（はる系列）	「行っ**て はった**」
「**て ＋ くれはる**」	待遇表現（尊敬・丁寧さ）	「見**て くれはる**」
「**て＋はる**」	V ＋て＋はる	「し**て はる**」

語彙（単語）だけでは捉えにくい地域差として，**定型的な言い回し**や**文末周辺の構文**がある．例えば，「〜や」「〜ねん」のような断定・説明の言い方や，「〜てくれはる」のような待遇表現（尊敬・丁寧さ）に関わる構文は，地域・レジスター（改まり／くだけ）によって偏りが生じやすい．しかし，これらは必ずしも単一語として現れるとは限らず，形態素単位に分割したときの**連接（並び）**として現れることが多い．そこで本研究では，MeCab による形態素解析結果（表層形列）から，事前に定義した**形態素列パターン**をヒューリスティックに検出・計数し，語彙マーカー特徴と同じ枠組みで正規化して特徴量化する．

3.3 地域推定と地域距離の定量化

地域推定は多クラス分類として定式化し，説明性を維持しつつ非線形性も扱えるランダムフォレストを用いる．学習は話者単位で行い，クラス不均衡に対して訓練データのみランダムオーバーサンプリングを適用する．さらに，本研究の目的は「当てる」だけでなく「近さ」を可視化することにある．そこで，各地域 r の平均特徴ベクトル $\mu_r = \mathbb{E}_{x\sim r}[f(x)]$ を計算し，地域間距離を

$$d(r_i, r_j) = \|\mu_{r_i} - \mu_{r_j}\|_2$$

で定義する．

4 評価

本節では，抽出した特徴量に基づいて関西方言の地域差を捉え，4 地域間の関係性を定量的に示す．

4.1 前処理および験設定

関西弁コーパス [8] を使用し，まず留学生を含むサブコーパス（RGS）を除外した．RGS は第二言語としての日本語使用や学習環境の影響が混入し得るため，本研究が対象とする「母語話者における地域差」の推定を歪める可能性があるからである．除外後，対象話者は 147 名となった．

ここで，話者ごとのテキスト量には大きな偏りがあり，極端に長い話者テキストが学習・評価を支配しうる．この偏りは，地域差の学習よりも特定話者の語癖・話題への過適合を誘発し，また推定分散の違い（長文は分散が小さく短文は大きい）によりクラス間比較を不安定化させる．そこで，MeCab によりトークン化した上で，トークン数が 5,000 を超える話者テキストは，トークン数が概ね均等になるよう複数チャンクに分割し，分析単位を調整した．分割後のサンプル集合に対する地域ラベルの内訳は，兵庫県 87 個，京都府 37 個，大阪 54 個，周辺県 52 個，合計 230 の処理されたデータを使って分類を行った．

そのうち，70%のデータで学習を行い，残りの 30%のデータで評価した．方言差の評価では平均性能だけでなくクラス間の偏りも重要であるため [4, 5]，クラス別の正解率/適合率/F1 値 と混同行列も報告する．

4.2 分類性能

提案手法に対し，学習時にランダムフォレスト分類器を使用した．4 つのラベル（京都・大阪・兵庫・周辺県）分類の結果，正解率 は 0.7826 であった．クラス別の適合率・再現率・F1 値 は表 2 の通りである．「兵庫」ラベルの F1 値は 0.78 と比較的安定しており，「周辺県」ラベルは適合率・再現率ともに 1.00（F1 値 1.00）で完全に分離された．一方，「京都」ラベルは適合率 1.00 と高いが再現率が 0.55 と低く，京都と判定される条件は強い一方で取りこぼしが発生していることが示唆される．「大阪」ラベルは適合率 0.58，再現率 0.69（F1 値 0.63）であり，兵庫との混同が残る傾向がある．全体のマクロ平均は適合率 0.83，再現率 0.76，F1 値 0.78 であった．また，データ分布を反映した全体性能を示すために各クラスのサポート数を重みとして算出した重み付き平均は，適合率 0.81，再現率 0.78，F1 値 0.78 であった．なお，深層表現を用いた BERT による分類の試行については，付録 A にまとめる．

表 2 ランダムフォレストの分類結果

クラス	適合率	再現率	F1 値	サポート
兵庫	0.75	0.81	0.78	26
京都	1.00	0.55	0.71	11
大阪	0.58	0.69	0.63	16
周辺県	1.00	1.00	1.00	16
マクロ平均	0.83	0.76	0.78	69
重み付き平均	0.81	0.78	0.78	69

また，表 3 の混同行列を見ると，周辺県は 16/16 が正しく分類されており，他クラスへの誤分類も生じていないため，本特徴量集合では周辺県を他の 3 地域から明確に分離できていることが分かる．一方で，兵庫は大阪に 5 件，京都は大阪に 3 件・兵庫に 2 件，また大阪は兵庫に 5 件誤分類されており，**京都・大阪・兵庫の三者間で相互に混同が残る**．特に大阪と兵庫の取り違えが対称的に見られることから，両地域で共有される語彙・文末表現・形態素パターンが多く，現状の特徴量では境界が連続的になっている可能性が示唆される．

4.3 重要特徴と地域差の解釈

表 4 に，上位の特徴量重要度を示す．重要度が高いのは，文末表現（終助詞）である「な」「で」「や

表 3　ランダムフォレスト分類器の混同行列（行：正解，列：予測）

	兵庫	京都	大阪	周辺県
兵庫	21	0	5	0
京都	2	6	3	0
大阪	5	0	11	0
周辺県	0	0	0	16

ん」「わ」であり，これらが地域推定に強く寄与している．続いて，大阪語彙「めっちゃ」「ちゃう」「ほんま」といった代表的語彙が寄与した．さらに，形態素パターンとして「て＋くれはる」，「（ん／の）＋や」，「（ん／の）＋ねん」，「ねん（トークン）」，「て＋はった」，および京都敬語系（例：V＋て＋はる）に対応するパターンが上位に現れた．一方で，京都語彙「おおきに」は重要度が非常に小さく，また「どす」「しはる」などは本分割では寄与が観測されなかったため表に記載しなかった．これらは，当該表現がほとんど出現しなかった，あるいは地域識別に十分な頻度差を形成しなかった可能性を示す．地域平均特徴ベクトルの定義および地域別の値については，付録 B に詳述する．

表 4　ランダムフォレストの特徴量重要度

特徴量	重要度
文末表現「な」	0.114926
文末表現「で」	0.095356
文末表現「やん」	0.091625
文末表現「わ」	0.089521
大阪語彙「めっちゃ」	0.087030
形態素パターン「て＋くれはる」	0.079679
形態素パターン「（ん／の）＋や」	0.079636
形態素パターン「ねん（トークン）」	0.078638
大阪語彙「ちゃう」	0.070844
大阪語彙「ほんま」	0.070205
形態素パターン「（ん／の）＋ねん」	0.061412
京都語彙「はる」	0.037878
形態素パターン「て＋はった」	0.021569
京都敬語系パターン（例：V＋て＋はる）	0.021528
京都語彙「おおきに」	0.000152

4.4　地域間距離と可視化

抽出した特徴量に基づく地域平均ベクトル間のユークリッド距離を算出し，その距離行列を表 5 に示す．また，同距離行列に対して MDS を適用し，4 ラベル（京都・大阪・兵庫・周辺県）の関係を 2 次元に配置した結果を図 2 に示す．

表 5 より，京都・大阪・兵庫の 3 地域間距離は兵庫—京都 0.00208，兵庫—大阪 0.00351，京都—大阪 0.00379 と小さく，**三府県が互いに近接**していることが分かる．一方，周辺県は兵庫から 0.01079，京都から 0.01208，大阪から 0.00942 と大きく，**三府県から相対的に離れた分布**であることが確認できる．ラベル間の距離を 2 次元に配置した図 2 を見ると，京都・大阪・兵庫が近い位置に集まり，周辺県が離れて配置されるという全体傾向が視覚的に確認でき，**距離表の関係を直感的に再現**している．特に，三府県の近接は混同行列で三者間の誤分類が残る点とも整合的であり，関西圏内部の地域差が府県境界で明確に分断されるというより，表現分布が重なり合う連続体として現れている可能性を示す．

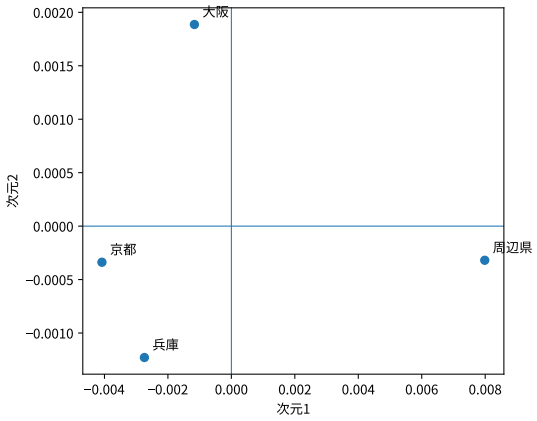

図 2　地域平均特徴量距離にもとづく MDS プロット

表 5　地域平均特徴量のユークリッド距離

	兵庫	京都	大阪	周辺県
兵庫	0.00000	0.00208	0.00351	0.01079
京都	0.00208	0.00000	0.00379	0.01208
大阪	0.00351	0.00379	0.00000	0.00942
周辺県	0.01079	0.01208	0.00942	0.00000

5　おわりに

本研究では，語彙・文末表現・形態素パターンからなる解釈可能特徴を用いて，関西方言の下位地域（京都・大阪・兵庫・周辺県）を推定した．ランダムフォレストにより 正解率 0.7826，F1 値 0.78 を得て，文末表現や「ねん」系・「て＋くれはる」などの定型パターンが有効な手掛かりであることを確認した．さらに，地域平均特徴ベクトルに基づく距離分析により，京都・大阪・兵庫が互いに近接し，周辺県が相対的に離れる傾向を確認した．今後は特徴量拡張，音声情報の統合により，精緻に分析する．

参考文献

[1] Marcos Zampieri, Preslav Nakov, and Yves Scherrer. Natural language processing for similar languages, varieties, and dialects: A survey. **Natural Language Engineering**, Vol. 26, No. 6, pp. 593–620, 2020.

[2] Kevin Heffernan. Kansaiben koopasu no shookai [an introduction to the corpus of kansai spoken japanese]. **Journal of Policy Studies**, No. 41, pp. 157–163, October 2012.

[3] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics.

[4] Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. Value: Understanding dialect disparity in nlu. arXiv preprint, 2022.

[5] Dipankar Srirag, Nihar Sahoo, and Aditya Joshi. Evaluating dialect robustness of language models via conversation understanding. arXiv preprint, 2024.

[6] 吉野幸一郎, 平山直樹, 森信介, 高橋文彦, 糸山克寿, 奥乃博. 日本語方言における音声対訳コーパスの構築. 言語処理学会 第 22 回年次大会 発表論文集, March 2016.

[7] Shinnosuke Takamichi and Hiroshi Saruwatari. CPJD corpus: Crowdsourced parallel speech corpus of Japanese dialects. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[8] 尹熙洙, 王竣磊, 岡田純子, 小木曽智信. 短単位版「関西弁コーパス」の構築と予備的分析. 言語処理学会 第 31 回年次大会 発表論文集, March 2025.

[9] Niyati Bafna, Emily Chang, Nathaniel R. Robinson, David R. Mortensen, Kenton Murray, David Yarowsky, and Hale Sirin. Dialup! modeling the language continuum by adapting models to dialects and dialects to models. arXiv preprint, 2025.

[10] Yo Sato and Kevin Heffernan. Creating dialect sub-corpora by clustering: a case in japanese for an adaptive method. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, Miyazaki, Japan, May 2018.

[11] Zedian Xiao, William Held, Yanchen Liu, and Diyi Yang. Task-agnostic low-rank adapters for unseen english dialects. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 7857–7870. Association for Computational Linguistics, December 2023.

[12] Leena Lulua and Ashraf Elnagar. Automatic arabic dialect classification using deep learning models. **Procedia Computer Science**, Vol. 142, pp. 262–269, 2018.

[13] Bruno Gonçalves and David Sánchez. Learning about spanish dialects through twitter. arXiv preprint, 2015.

[14] Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. A neural model for user geolocation and lexical dialectology. arXiv preprint, 2017.

[15] 入江さやか, 金明哲. 方言録音文字化資料における拍 bigram から見た方言分類 — 岐阜・愛知方言の所属は東か西か —. 計量国語学, Vol. 32, No. 1, pp. 1–18, June 2019.

A　BERT 分類の試行

本研究では，日本語事前学習モデルである BERT[1)] を用いた分類も試行した．

- データ分割は層化ランダム分割とし，test=0.30，残りを train/validation に分割した．
- 最大系列長は 128，256 の 2 つの設定で試した．
- バッチサイズは学習 8，評価 16 とし，学習エポック数は 200 とした．
- 評価指標は正解率および F1 値を用いた．

結果として，BERT 分類は安定して十分な性能を得られなかった．複数回試行の平均は正解率 0.5508（分散 0.0406），F1 値 0.4099（分散 0.0673）であった．

この要因として，話者数が大きくない条件では，高容量モデルの安定学習に必要なデータ量が不足し，分割の揺らぎが性能分散として表れやすいこと，そして関西方言の地域差は語彙意味というより終助詞・定型表現・形態素列など「局所的で形式的」な手がかりに現れやすい一方，BERT ではそれらを明示的に数え上げて強調する仕組みがなく，本研究規模のデータでは識別根拠が希薄になりやすいこと，が考えられる．このため本研究では，高精度化よりも「根拠の明示」と「距離の可視化」を優先し，語彙マーカーと形態素パターンの統合特徴に基づく軽量分類を主結果として報告する．

B　地域平均特徴ベクトルの説明

地域の「近さ」を議論するため，特徴ベクトル $f(x)$ の地域平均 $\mu_r = \mathbb{E}_{x\sim r}[f(x)]$ を計算する．表 6 は，地域差（分散）が比較的大きく，解釈もしやすい次元を抜粋して示したものである．文末助詞や頻出語彙（例：「めっちゃ」「ちゃう」「ほんま」）に加え，形態素列パターン（例：「ん/の+や」「ねん」「て+くれはる」）が地域ごとに異なる強度で現れており，平均ベクトルの差分が地域間距離の根拠として機能することが確認できる．さらに，地域平均特徴ベクトル間の距離 $d(r_i, r_j) = \|\mu_{r_i} - \mu_{r_j}\|_2$ は，単なる分類境界の有無ではなく，**どの地域どうしがどの程度似ているか**を連続的に評価する指標となる．距離が小さい場合，両地域は語彙・文末表現・形態素パターンの分布が近く，相互に混同されやすいことを意味する．一方，距離が大きい場合，平均的な表現選好が異なる方向に偏っており，分類器にとっても分離しやすい関係にあると解釈できる．

表 6　地域平均特徴ベクトルの地域ラベル別平均値（正規化頻度）

	兵庫	京都	大阪	周辺県
文末「な」	0.021762	0.020717	0.023031	0.031949
文末「で」	0.017466	0.016952	0.018541	0.019596
文末「やん」	0.002781	0.003243	0.005272	0.004931
文末「わ」	0.001972	0.001479	0.001926	0.003190
「めっちゃ」	0.001231	0.002643	0.002080	0.000271
「ちゃう」	0.001723	0.001721	0.001631	0.002534
「ほんま」	0.000567	0.000398	0.000725	0.000285
「はる」	0.000063	0.000574	0.000155	0.000013
形態素列「ん/の+や」	0.001721	0.001744	0.001644	0.001250
形態素列「ねん」	0.003867	0.004230	0.005467	0.004113
形態素列「て+くれはる」	0.000350	0.000136	0.000207	0.000289

また，50 種類以上の語彙および特徴量の組み合わせについて探索的に検証を行った．その多くは，対象コーパス内での出現頻度が極めて低く，正規化頻度がほぼ 0 または 0 となったため，地域平均特徴ベクトルの推定や分類器の学習に実質的な情報を与えなかった．低頻度特徴を含めた場合，距離推定や重要度算出においてノイズとして作用する可能性があることから，本研究では一定以上の出現頻度を満たす特徴量に限定して分析を行った．

1）https://huggingface.co/tohoku-nlp/bert-base-japanese-v2