# 回答者の生年に基づく語の意味変化の検出

峯尾海成[1]　門戸巧[2]　佐藤道大[2]　山岸祐己[2]　谷口ジョイ[2]
[1] 静岡理工科大学大学院理工学研究科　[2] 静岡理工科大学情報学部
{2221027.mk, 2018119.mt, 1918061.sm, yamagishi.yuki, taniguchi.joy}@sist.ac.jp

## 概要

本研究は，静岡県全域で用いられる方言「まめったい」の使用・理解に関する大規模調査から得られた時系列データを分析することで，意味やイメージにどのような変化が生じているのか，またその生成時期を推定することを主な目的としている．多項分布型レジームスイッチングおよび多群出現順位統計量によって分析を行った結果，「体を動かしてよく働く」という意味での使用・理解は衰退傾向にあり，代わって「性格的に几帳面だ」という意味での使用が増加していることが明らかになった．また，語のイメージは，肯定的なものから否定的なものへと変化していた．

## 1　語の意味変化

本稿は，静岡方言「まめったい」という語の意味がどのように変化しているかを問う意味変化論に関するものである．語の意味やイメージが変化する過程にはさまざまな要因が関わっており，進行中の変化を捉えることは不可能であるとされていた [1]．そのため，これまでの研究においては「既に意味変化が完了した語」が扱われてきた．本研究において筆者らが進行中の意味変化を可視化するために焦点を当てたのが，静岡全域で用いられる「まめったい」という形容詞である．「まめったい」という語は，(1) 体を動かしてよく働く (2) 健康だ (3) 落ち着きがない (4) 性格的に几帳面だ，のように複数の意味をもち [2]，多義性が認められる上，複数の語義が併存している．加えて，静岡方言研究会 [3] の調査によれば，「まめったい」は，共通語の「まめだ」と競合関係にあることが指摘されており，共通語が「まめったい」の意味領域に影響を与えている可能性がある．本研究は，静岡方言の「まめったい」という語をひとつの事例とし，時系列データを扱うことを想定した手法を応用することにより，過去に生じた意味やイメージの変化のみならず，進行中の変化についても明らかにすることを目的としている．本研究では以下 2 点の仮説を検証する．

1. 静岡方言「まめったい」は過去に意味および語のイメージに変化が生じており，その変化は現在も進行中である．
2. 提案手法である多項分布型レジームスイッチングおよび多群出現順位統計量によって，これまで困難であった進行中の意味変化の可視化が可能である．

## 2　提案手法

### 2.1　多項分布型レジームスイッチング検出

多項分布型レジームスイッチング検出は，タイムラインを生成することで，複雑に変化するデータを単純化する手法である．本研究では，データが多項分布に従うと仮定し，多項分布型レジームスイッチング検出をデータの単純化手法として用いることで，静岡方言「まめったい」において，過去に生じた意味変化とその時期を推定する．代表的な変化点検出手法 [4, 5] では，単一情報の傾向変化を可視化することに特化しているが，多項分布を仮定したレジームスイッチング手法 [6] においては，複数情報を扱うことを前提としており，変化が生じた時期の把握が容易であることが示されている．ここでは，その基本技術を応用し，自動でレジーム数を決定する方法についても述べる．

ある調査データを $\mathcal{D} = \{(s_1, t_1), \cdots, (s_N, t_N)\}$ とする．ここで，$s_n$ と $t_n$ は，$J$ カテゴリの状態と $n$ 番目の観測時刻，すなわち，各回答における選択肢と回答者の生年をそれぞれ表す．$|\mathcal{D}| = N$ を観測数とすると，$t_1 \leq \cdots \leq t_n \leq \cdots \leq t_N$ となる．$n$ はタイムステップとし，$\mathcal{N} = \{1, 2, \cdots, N\}$ をタイムステップ集合とする．また，$k$ 番目のレジームの開始時刻を $T_k \in \mathcal{N}$，$\mathcal{T}_K = \{T_0, \cdots, T_k, \cdots, T_{K+1}\}$ をスイッチングタイムステップ集合とし，便宜上

$T_0 = 1$, $T_{K+1} = N + 1$ とする．すなわち，$T_1, \cdots, T_K$ は推定される個々のスイッチングタイムステップであり，$T_k < T_{k+1}$ を満たすとする．そして，$\mathcal{N}_k$ を $k$ 番目のレジーム内のタイムステップ集合とし，各 $k \in \{0, \cdots, K\}$ に対して $\mathcal{N}_k = \{n \in \mathcal{N}; T_k \leq n < T_{K+1}\}$ のように定義する．なお，$\mathcal{N} = \mathcal{N}_0 \cup \cdots \cup \mathcal{N}_K$ である．

いま，各レジームの状態分布が $J$ カテゴリの多項分布に従うと仮定する，$\boldsymbol{p}_k$ を $k$ 番目のレジームにおける多項分布の確率ベクトルとし，$\mathcal{P}_K$ はそれら確率ベクトルの集合，つまり $\mathcal{P}_K = \{\boldsymbol{p}_0, \cdots, \boldsymbol{p}_K\}$ とすると，$\mathcal{T}_K$ が与えられたときの対数尤度関数は以下のように定義できる．

$$L(\mathcal{D}; \mathcal{P}_K, \mathcal{T}_K) = \sum_{k=0}^{K} \sum_{n \in \mathcal{N}_k} \sum_{j=1}^{J} s_{n,j} \log p_{k,j}. \tag{1}$$

ここで，$s_{n,j}$ は $s_n \in \{1, \cdots, J\}$ を

$$s_{n,j} = \begin{cases} 1 & \text{if } s_n = j; \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

のように変換したダミー変数である．各レジーム $k = 0, \cdots, K$ と各状態 $j = 1, \cdots, J$ に対する式 (1) の最尤推定量は $\hat{p}_{k,j} = \sum_{n \in \mathcal{N}_k} s_{n,j} / |\mathcal{N}_k|$ のように与えられる．これらの推定量を式 (1) に代入すると以下の式が導ける．

$$L(\mathcal{D}; \hat{\mathcal{P}}_K, \mathcal{T}_K) = \sum_{k=0}^{K} \sum_{n \in \mathcal{N}_k} \sum_{j=1}^{J} s_{n,j} \log \hat{p}_{k,j}. \tag{3}$$

したがって，スイッチングタイムステップの検出問題は，式 (3) を最大化する $\mathcal{T}_K$ の探索問題に帰着できる．

適当な条件下で式 (3) を最大化しようとすると，最適解を求めるための計算量が $O(N^K)$ となってしまうため，ある程度大きい $N$ に対して $K \geq 3$ となってしまうと，実用的な計算時間で解くことができない．したがって，任意の $K$ について解くために，貪欲法と局所探索法を組み合わせた方法 [6] を用いる．なお，本実験では貪欲法アルゴリズムの終了条件として最小記述長原理 (MDL) [7] を採用し，事前にレジーム数を設定することなく自動で終了させる．

## 2.2 多群出現順位統計量に基づく時系列データの変換

データカテゴリの時系列的変化を明示し，それらを複数カテゴリ間で比較するため，出現順位を用いた統計量によるデータ変換を行う．この手法は，Mann–Whitney の U 検定 [8] を基盤とし，多群を扱えるよう拡張したものであり，データの出現頻度の傾向変化を z-score として表現する．静的な分析手法による指標を，動的な視点で捉えられるよう可視化するため，長期的な変化が捉えやすいと言える．また，各カテゴリの z-score は，他のカテゴリすべてを基準としているため，複数間のカテゴリにおける比較が容易である．

多項分布レジームスイッチングの問題設定と同様に，ある調査データのタイムステップ（回答者の生年）集合と，それらが有するカテゴリ（選択肢）集合をそれぞれ $\mathcal{N}$ と $\mathcal{J}$ とする．つまり，それぞれの要素数は $N = |\mathcal{N}|$ と $J = |\mathcal{J}|$ とし，各要素は整数と同一視されるとする．すなわち，$\mathcal{N} = \{1, \cdots, n, \cdots, N\}$ および $\mathcal{J} = \{1, \cdots, j, \cdots, J\}$ である．なお，オブジェクト $n$ は最古のものが 1，最新のものが $N$ となるよう，出現順に並んでいるものとする．このとき，タイムステップ $n$ がカテゴリ $j$ を有する場合は 1，それ以外の場合は 0 となっている $J$ 行 $N$ 列の行列を $Q$ ($q_{j,n} \in \{0, 1\}$) とすると，オブジェクト $n$ が有するカテゴリ数は $d_n = \sum_{i=1}^{J} q_{i,n}$，タイムステップ $n$ までのカテゴリ $j$ の出現数は $I_{j,n} = \sum_{i=1}^{n} q_{j,i}$ のように表せる．いま，オブジェクトに付随してカテゴリが出現するとし，以降では，オブジェクト出現からカテゴリ出現へと視点を変える．このとき，オブジェクト $n$ が唯一のカテゴリのみ有する $d_n = 1$ の場合では，オブジェクト $n$ に付随して出現したカテゴリ $j$ の出現順位は $r_n = I_{n-1} + 1$ であるが，複数のカテゴリを有する $d_n > 1$ の場合では，平均順位を考えなければならないため，その出現順位は $r_n = I_{n-1} + (1 + d_n)/2$ となる．ここでの目的は，タイムステップとカテゴリの集合が与えられたとき，出現順位の値が大きい（新しい），または逆に小さい（古い）タイムステップが有意に多く含まれるカテゴリを定量的に評価する指標の構築である．

Mann–Whitney の二群順位統計量 [8] を多群に拡張し，カテゴリの出現順位に適用する方法について述べる．いま，カテゴリ $j$ に着目すれば，このカテゴリに属するタイムステップ集合 $\{n \in \mathcal{N} : q_{j,n} = 1\}$ と，このカテゴリに属さないタイムステップ集合 $\{n \in \mathcal{N} : q_{j,n} = 0\}$ の二群に分割することができる．よって，Mann–Whitney の二群順位統計量に従い，次式により，タイムステップ $n$ までのカテゴリ $j$ に対し z-score $z_{j,n}$ を求めることができる．

$$z_{j,n} = \frac{u_{j,n} - \mu_{j,n}}{\sigma_{j,n}}. \tag{4}$$

ここで，統計量 $u_{j,n}$, 出現順位の平均 $\mu_{j,n}$，および，その分散 $\sigma^2_{j,n}$ は次のように計算される．

$$u_{j,n} = \sum_{i=1}^{n} nq_{j,i} - \frac{I_{j,n}(I_{j,n}+1)}{2}, \tag{5}$$

$$\mu_{j,n} = \frac{I_{j,n}(n-I_{j,n})}{2}, \tag{6}$$

$$\sigma^2_{j,n} = \frac{I_{j,n}(I_n-I_{j,n})}{12}\left((I_n+1) - \sum_{i=1}^{n}\frac{d_i^3-d_i}{I_n(I_n-1)}\right). \tag{7}$$

先程と同様，各オブジェクトが複数のカテゴリを有し得ないケースでは，式 (7) の $d_i$ を含む項，すなわち平均順位を扱うための補正値の計算は不要である．この多群順位統計量は，基本的には 2 クラス分類器の SVM (Support Vector Machine) [9] を多クラス分類器に拡張するときに利用される one-against-all と類似した考え方となる．

以上より，式 (4) で求まる z-score $z_{j,n}$ により，オブジェクト $k$ までの各カテゴリ $j$ が，出現順位の値が大きい（新しい），または逆に小さい（古い）オブジェクトを有意に多く含むかを定量的に評価することができる．すなわち，この $z_{j,n}$ が正の方向に大きければ大きいほど，タイムステップ $n$ の直近での出現が有意に多いということであり，カテゴリ $j$ の勢力が伸びていることになる．逆に，$z_{j,n}$ が負の方向に大きいということは，過去に比べて勢力が衰えていることになる．また，式 (4) で求まる z-score $z_{j,n}$ の計算量は全てのオブジェクトと全てのカテゴリについて算出した場合でも $O(NJ)$ と高速であり，オンライン処理においても新たに追加されたオブジェクトごとに $O(J)$ の計算量しかかからない．

## 3 調査概要

本調査は，静岡方言を母方言話とする 1,544 名を対象としている．調査は，質問紙，あるいはウェブ調査により行い，性別，生年，出身地域，家庭内における方言使用の有無といった基本情報を収集した．冒頭に「まめったい」という語の使用・理解についての設問を設け，「使用する」「使用しないが理解できる」と回答した 1,195 名については，1 節で述べた 4 つの意味についてそれぞれの用例を示し、その使用、理解について回答を依頼した。

最後に，「まめったい」を用いた短文作成を依頼し（852 名が回答），上記に基づく意味分類を行った上で，肯定的な意味で使用されているか，あるいは否定的に用いられているか，という観点から分類を行った．

## 4 実験結果とまとめ

以下は，「まめったい」の使用・理解に関する調査結果を，上記の 2 手法で可視化したものである．多項分布型レジームスイッチングにより「まめったい」の使用・理解に見られる変化を単純化・可視化した結果，1956 年，1980 年，1995 年（調査協力者の生年）でデータ構造の変化が見られ，その使用は全体として減少傾向にあることが判明した（図 1）．また，この傾向は 1950 年生まれから見られ，1970 年生まれから顕著なものになっていることが示唆された（図 2）．また，「まめったい」の使用・理解の低下については，z-score の最終値 $z_{j,N}$ の絶対値の大きさから，今後も継続する可能性が高いと思われる．

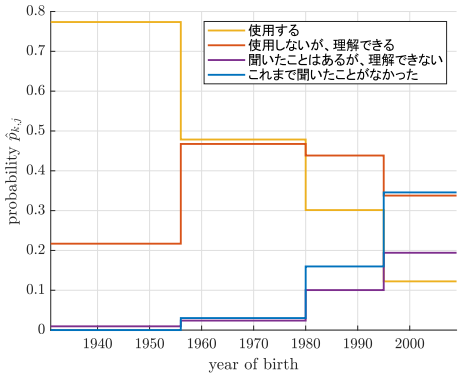


**図 1** 多項レジームスイッチング検出 (使用・理解)

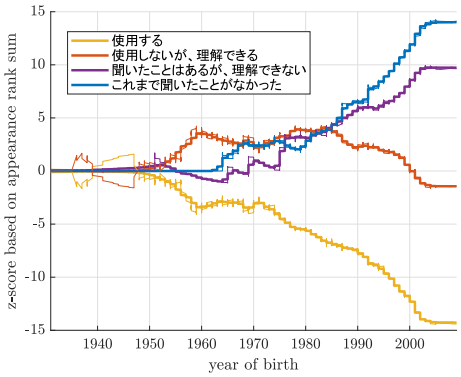


**図 2** 多群出現順位統計量 (使用・理解)

図 3 は，「まめったい」を用いた短文作成に基づく意味分類における多項分布型レジームスイッチングの検出結果であり，1980 年代半ば生まれにデータ構造が変化していることが示されている．短文作成に基づく意味分類においては「体を動かしてよく働く」という意味での使用が減少傾向にあり，「性

格的に几帳面だ」という意味での使用が増加傾向にあることが明らかとなった（図 3）．また，「健康だ」および「落ち着きがない」という意味については，検出されたスイッチング以降ほぼ出現していないことがわかる．多群出現順位統計量の結果から，この変化傾向は 1970 年生まれから際立っていることが明らかとなった（図 4）．

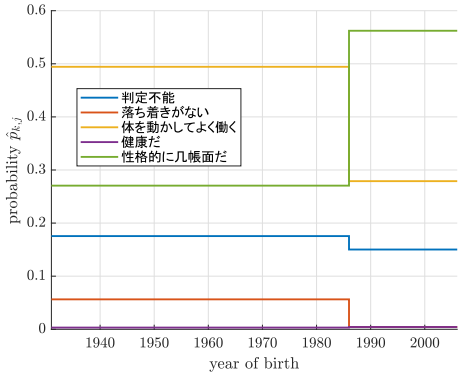


**図 3**　多項レジームスイッチング検出 (意味分類)

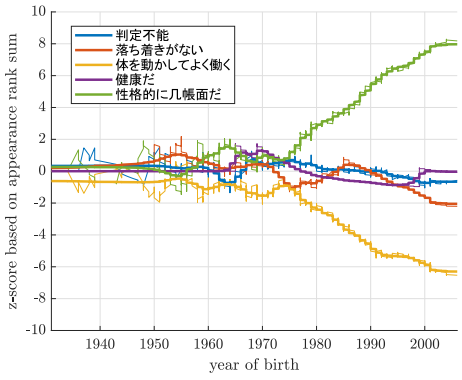


**図 4**　多群出現順位統計量 (意味分類)

最後に，「まめったい」を用いて作成された短文が肯定的な意味，あるいは否定的な意味で使用されているかについて分析を行ったところ，レジームスイッチングの結果からは，スイッチングが検出されなかった（図 5）．そのため，年代全体を通した分布に大きな変化はないと推察される．一方，多群出現順位統計量の結果（図 6）からは，出現傾向の変化が確認でき，特に 1970 年代半ば生まれからは，肯定的な意味から否定的な意味で使用されるようになっている．

意味変化の要因は複雑かつ多層的であるため，特定することは困難であるが，競合する共通語「まめだ」が，方言形「まめったい」の意味領域に影響を与えている可能性は排除できない．共通語「まめだ」が「まめったい」の主要な意味（＝体を動かしてよく働く）を侵食するような形で意味変化を引き起こしていることが今回の分析によって示唆された．

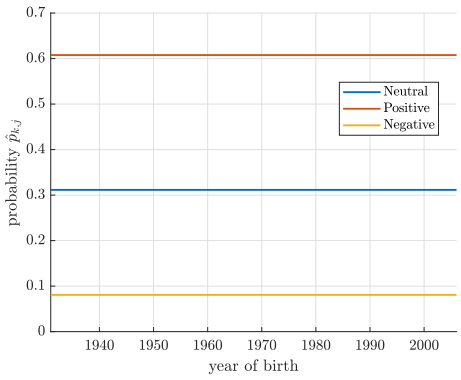


**図 5**　多項レジームスイッチング検出 (語イメージ)

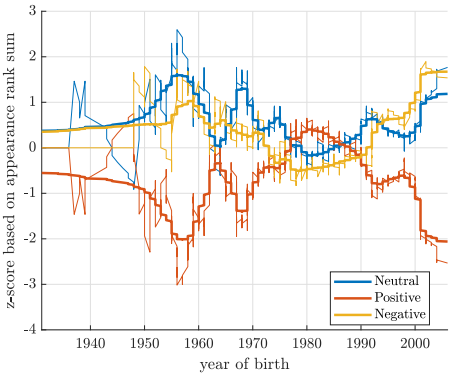


**図 6**　多群出現順位統計量 (語イメージ)

## 5　終わりに

本研究では，共通語「まめだ」と競合関係にある静岡方言「まめったい」の意味および語イメージの変化について，2 つの提案手法を用いて単純化及び可視化を行った．

分析の結果「まめったい」の使用・理解については低下傾向にあり，変化は今後も続くことが示唆された．語の意味については，「体を動かしてよく働く」という意味から「性格的に几帳面だ」という意味へと変化していた．これに伴い，語イメージも肯定的なものから否定的なものへと変化していた．また，同手法によりこうした変化の生成時期についても推定が可能であった．以上，本研究で用いたレジームスイッチング検出，および，多群出現順位統計量という 2 つの提案手法によって，これまで捉えることが困難であった「進行中の意味変化」の推定が可能であることがわかった．

## 謝辞

本調査にご協力いただいた方々に感謝いたします．本研究は，静岡理工科大学グループ研究推進支援費の助成を受けています．

## 参考文献

[1] Leonard Bloomfield. **Language**. H. Holt and Company, 1933.

[2] 東条操 (編). 全国方言辞典. 東京堂, 1951.

[3] 静岡方言研究会. 図節静岡県方言辞典. 吉見書店, 1951.

[4] J. Kleinberg. Bursty and hierarchical structure in streams. In **Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)**, pp. 91–101, 2002.

[5] Rebecca Killick, Paul Fearnhead, and I.A. Eckley. Optimal detection of changepoints with a linear computational cost. **Journal of the American Statistical Association**, Vol. 107, pp. 1590–1598, 12 2012.

[6] Yuki Yamagishi and Kazumi Saito. Visualizing switching regimes based on multinomial distribution in buzz marketing sites. In **Foundations of Intelligent Systems - 23rd International Symposium, ISMIS 2017**, Vol. 10352 of **Lecture Notes in Computer Science**, pp. 385–395. Springer, 2017.

[7] J. Rissanen. Modeling by shortest data description. **Automatica**, Vol. 14, No. 5, pp. 465–471, September 1978.

[8] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. **Ann. Math. Statist.**, Vol. 18, No. 1, pp. 50–60, 03 1947.

[9] Vladimir N. Vapnik. **The Nature of Statistical Learning Theory**. Springer-Verlag New York, Inc., New York, NY, USA, 1995.