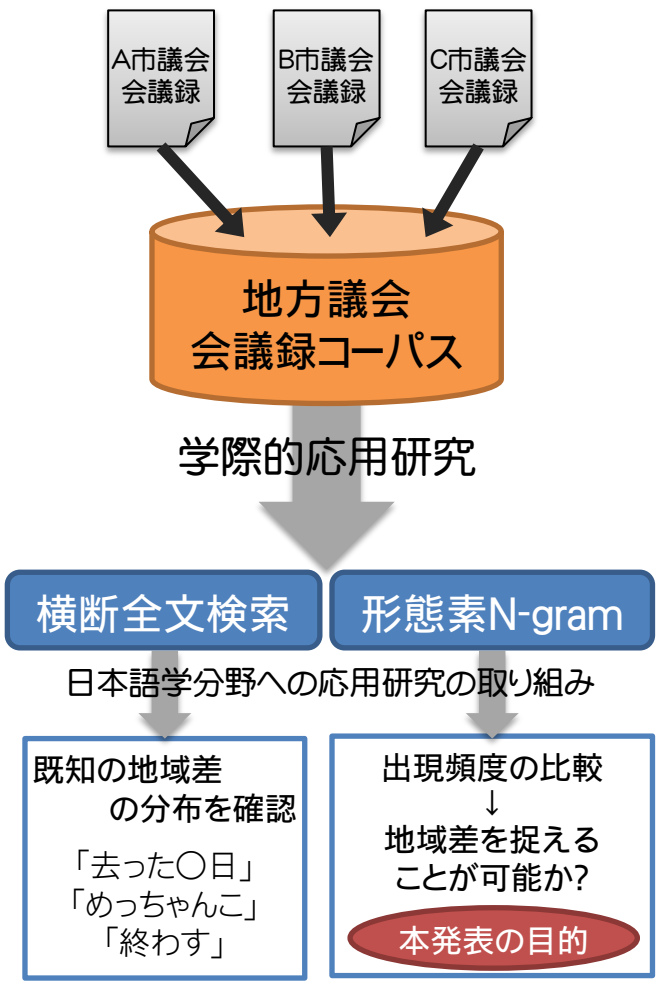
P4-23

形態素N-gramを用いた地方議会会議録コーパスの地域変異検出の試み
―文末表現を例に―

高丸圭一(宇都宮共和大)・乙武北斗(福岡大)・渋木英潔(横浜国大)・木村泰知(小樽商大)・森辰則(横浜国大)



形態素N-gramから文末表現の地域差を捉える試み

① 全国405自治体の2010年の地方議会会議録

② MeCabを用いて形態素解析
解析辞書にはUnidicを使用

③ ひらがなで構成された4-gramを分析
第1～第3形態素⇒ひらがな
第4形態素⇒句点

④ 出現頻度の総数50以上の4-gramを対象
総数 4,341,447
異なり数 1,331

⑤ 自治体ごとに出現確率を計算

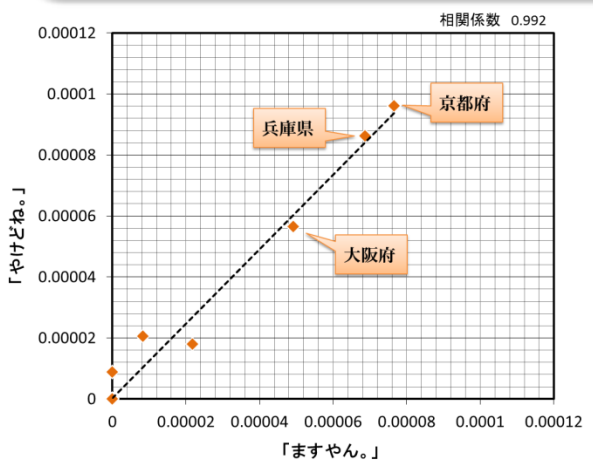
$$\text{出現確率} = \frac{\text{そのフレーズの出現頻度}}{\text{第4形態素が句点である4-gramの総出現頻度}}$$

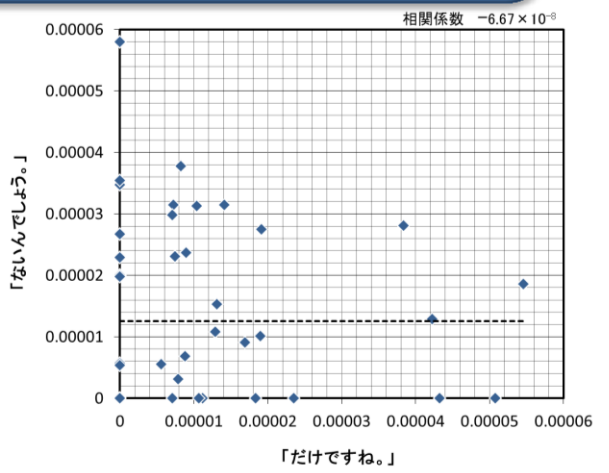
出現傾向(どの都道府県で多く出現するか)を明らかにする

↓

フレーズ間の相関係数

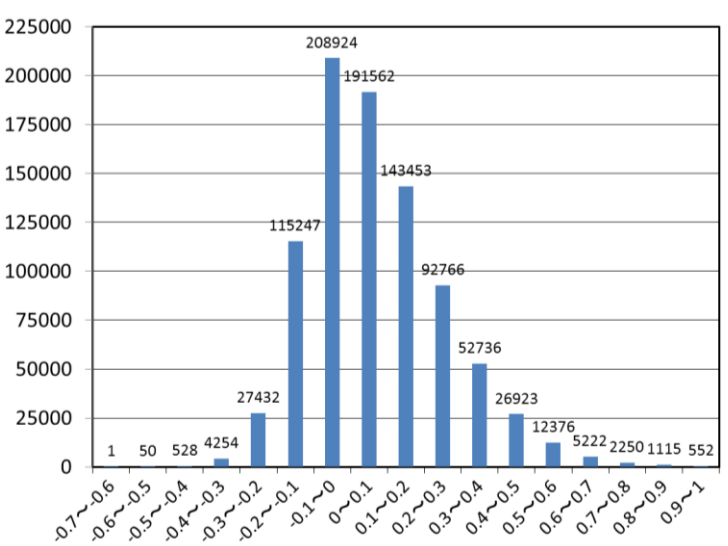
1,331フレーズの組み合わせ(885,115組)について,
47都道府県の出現確率をパラメータとして相関係数を求めた





	4-gram	全国計	最大	最小
1	て／おり／ます／。	866,450	佐賀県 (0.1861)	和歌山県 (0.0687)
2	で／ござい／ます／。	652,018	高知県 (0.1883)	和歌山県 (0.0259)
3	で／あり／ます／。	323,651	富山県 (0.1504)	神奈川県 (0.0230)
4	て／い／ます／。	172,671	秋田県 (0.0537)	長崎県 (0.0097)
5	て／いただき／ます／。	110,858	奈良県 (0.0371)	鹿児島県 (0.0033)
6	ませ／ん／か／。	109,905	鳥取県 (0.0620)	佐賀県 (0.0067)
7	を／いたし／ます。	87,083	鳥取県 (0.0419)	青森県 (0.0030)
8	いたし／まし／た／。	70,804	鹿児島県 (0.0306)	石川県 (0.0061)
9	て／ござい／ます／。	62,891	東京都 (0.0258)	富山県 (0.0000)
10	あり／まし／た／。	61,340	和歌山県 (0.0233)	鹿児島県 (0.0041)

出現頻度の高いフレーズ(上位10個)

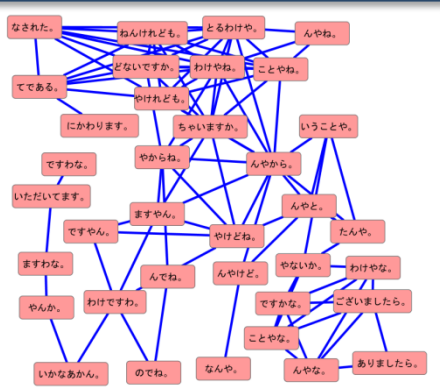


フレーズ間相関係数の分布

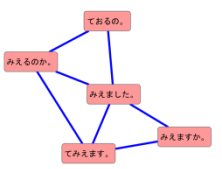
相関係数の高いフレーズ対 → 出現地域の傾向が類似

フレーズ間のネットワーク表現

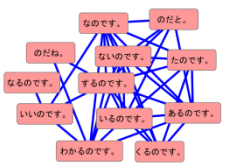
相関係数が0.85以上のフレーズ対をネットワーク図で表現
696フレーズ対(異なりフレーズ数:207)



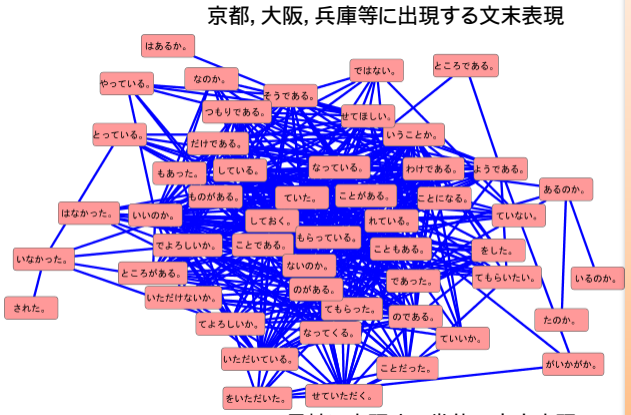
京都, 大阪, 兵庫等に出現する文末表現



愛知, 岐阜, 三重, 滋賀に出現する尊敬表現「～みえる」



宮城に出現する「～のです」



長崎に出現する常体の文末表現

まとめ

- 都道府県の出現確率をパラメータとして求めた相関係数によって, 地域差を捉えることができた。
- 地域差には, 発言の地域差(方言)と整文方法の偏りが含まれる。
- データの性質上, 元の文脈や話者の属性を確認しやすい。
- 今後, 本手法で発見される地域差について, 詳細な分析を進める。