Construction of Japanese Prefectural Assembly Minutes Datasets Across Three Electoral Terms:
Comparative Analysis of 2011, 2015, and 2019 Four-Year Periods

Keiichi Takamaru[1], Yuzu Uchida[2], Hokuto Ototake[3], Yasutomo Kimura[4]

[1]Utsunomiya Kyowa University, Japan [2]Hokkai-Gakuen University, Japan [3]Fukuoka University, Japan [4]Otaru University of Commerce, Japan

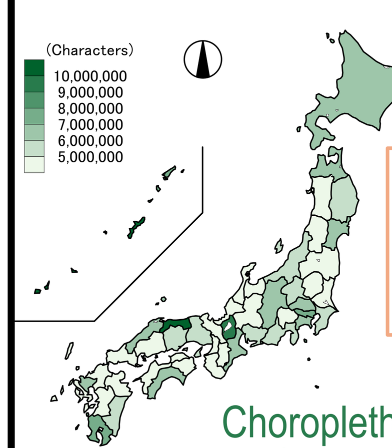
Background and Motivation

- Local government assembly minutes capture **authentic political discourse** at the subnational level.
- They offer valuable data for **corpus linguistics**, **dialectology**, and **political science**.
- However, **longitudinal cross-regional analysis** of such records has remained **limited**.
- We construct a **12-year**, **47-prefecture longitudinal corpus** spanning **three electoral terms**.

Dataset Construction

Data Source

47 prefectural assembly search systems



⚠ Formats differ across prefectures — custom crawler required for each

Choropleth map shows total characters spoken per prefecture (2019 term)

Crawl & Segment

Split into individual remarks by sentence-final periods

Normalize & Annotate

Speaker name / Meeting name / Date

⚠ Name notation inconsistencies
Manual verification and disambiguation required
Hiragana / Kanji notation variants — Stage / ring name vs. legal name

Assign Speaker IDs

Unique ID per member consistent across all three terms

Overall Size of our Datasets

Assembly minutes data
from all **47** prefectures
3 electoral terms, **12 yrs** longitudinal data
743M chars, **471M** tokens, **12.2M** remarks

Assembly Members DB

provided by Seijiyama, Election candidates DB

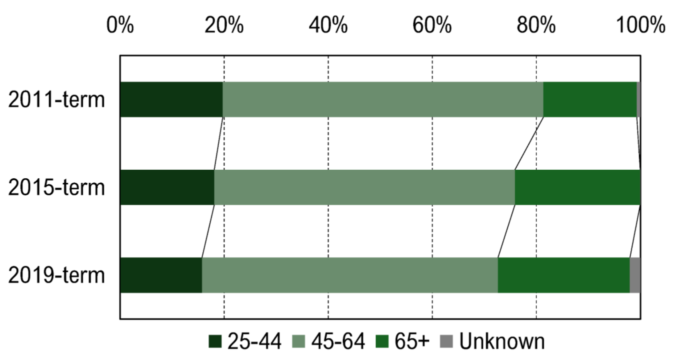
Birth Year / Gender / Name
Electoral district / Election results

Linking

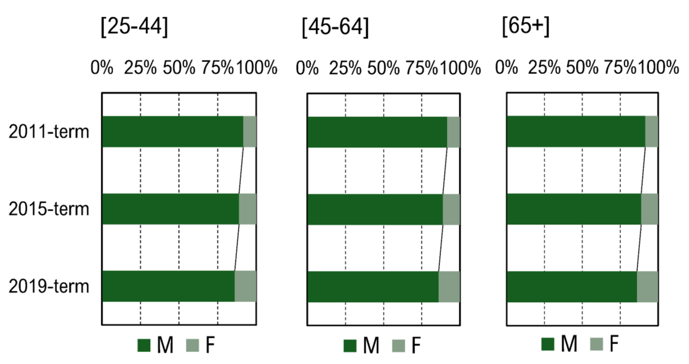
Comparative Analysis

Member composition

The proportion of members **aged 25–44 has declined** across terms, while those aged **65+ have increased**.



Female representation has **increased** in all age groups across terms, yet remained at **only 12%** even in the 2019 term.



Volume of speech

Term	Gender	24-44	45-65	65+
2011	M	43,583	44,626	31,299
	F	44,455	59,969	59,366
2015	M	45,814	45,378	30,469
	F	42,886	55,950	66,057
2019	M	46,921	45,448	32,262
	F	46,824	52,745	59,147

Female members speak more characters than male counterparts in almost all age groups and terms, with opposite age trends: **female speech volume increases with age**, while **male speech volume decreases**.

Non-speaking Rate

The proportion of members who never speak varies from under 2% for **younger females** to over 10% for **males aged 65+**.

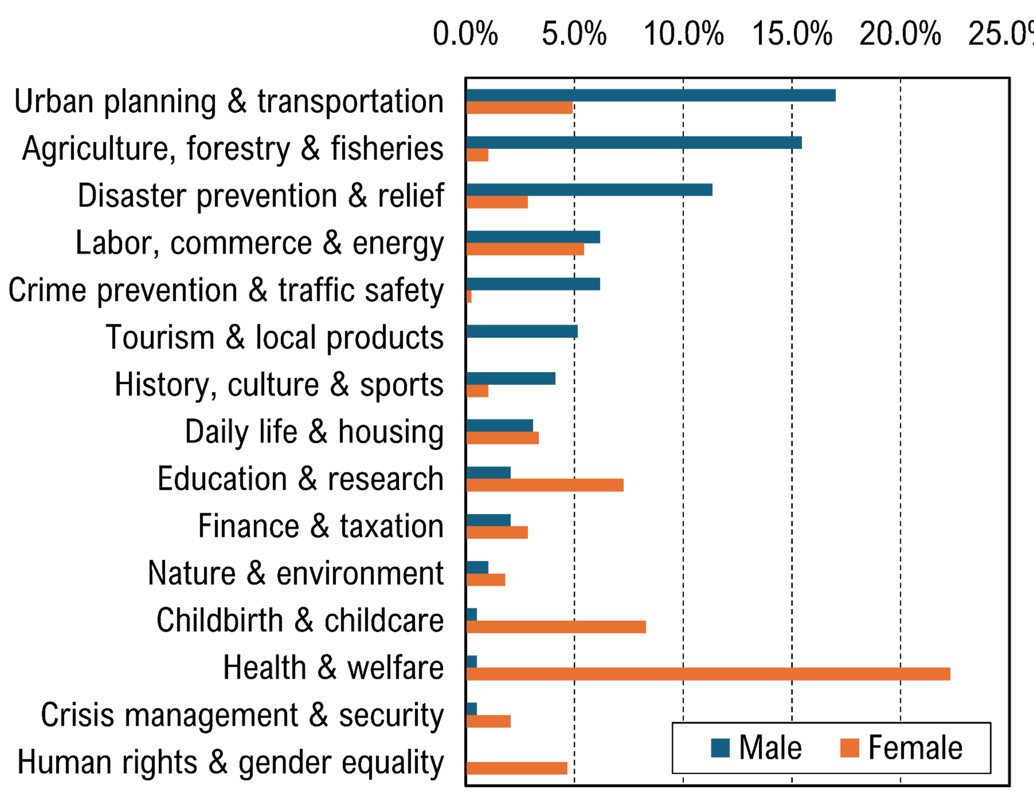
Term	Gender	24-44	45-65	65+
2011	M	9(1.7%)	76(4.6%)	62(12.5%)
	F	1(2.0%)	2(1.2%)	2(6.3%)
2015	M	9(2.0%)	38(2.7%)	71(11.3%)
	F	0(0.0%)	2(1.1%)	1(2.5%)
2019	M	7(1.9%)	36(2.6%)	76(11.7%)
	F	0(0.0%)	3(1.5%)	3(5.3%)

Retiring or defeated members show **significantly higher non-speaking rates** (up to 10.8%), whereas **newcomers and returning members** show **markedly lower rates** (as low as 0.3%) — a finding only possible through longitudinal multi-term analysis.

Use Cases and Downstream Applications

Gender and Age Analysis in Political Discourse

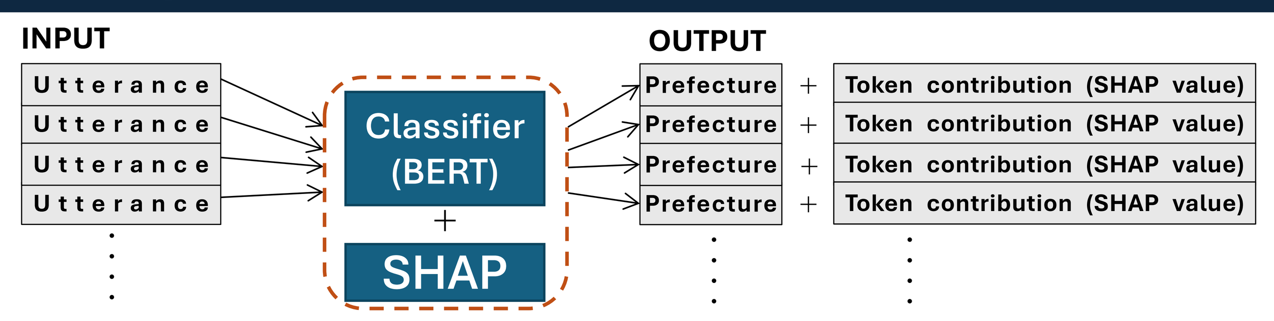
Using log-likelihood ratios, we extracted gender-specific political vocabulary from the corpus. **Female** members used significantly more terms related to **health & welfare, childcare, and education**, while **male** members favored **urban planning, agriculture, and disaster prevention**.



Distribution of political characteristic terms by gender

Yuzu Uchida, Keiichi Takamaru, Hokuto Ototake, and Yasutomo Kimura. 2019. Visualization of activity of local assembly members based on log-likelihood ratio and political vocabulary ratio (in Japanese). *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, 31(2):662–671.

Computational Dialectology with eXplainable AI



A BERT-based classifier was trained to predict the prefecture of origin from assembly utterances. XAI (SHAP) then identified which tokens contributed most to each prediction — revealing dialect words and region-specific expressions as key features.

はぐってていただきまして、62ページ、5目
農地費733万2,000円でございます。

Hagutte *itadakimashite, 62 peeji, 5-moku noochi-hi 733 man 2000 en de gozaimasu.*
Please **turn to** page 62, and the land expenses for item 5 are 7,332,000 yen.

Example of output

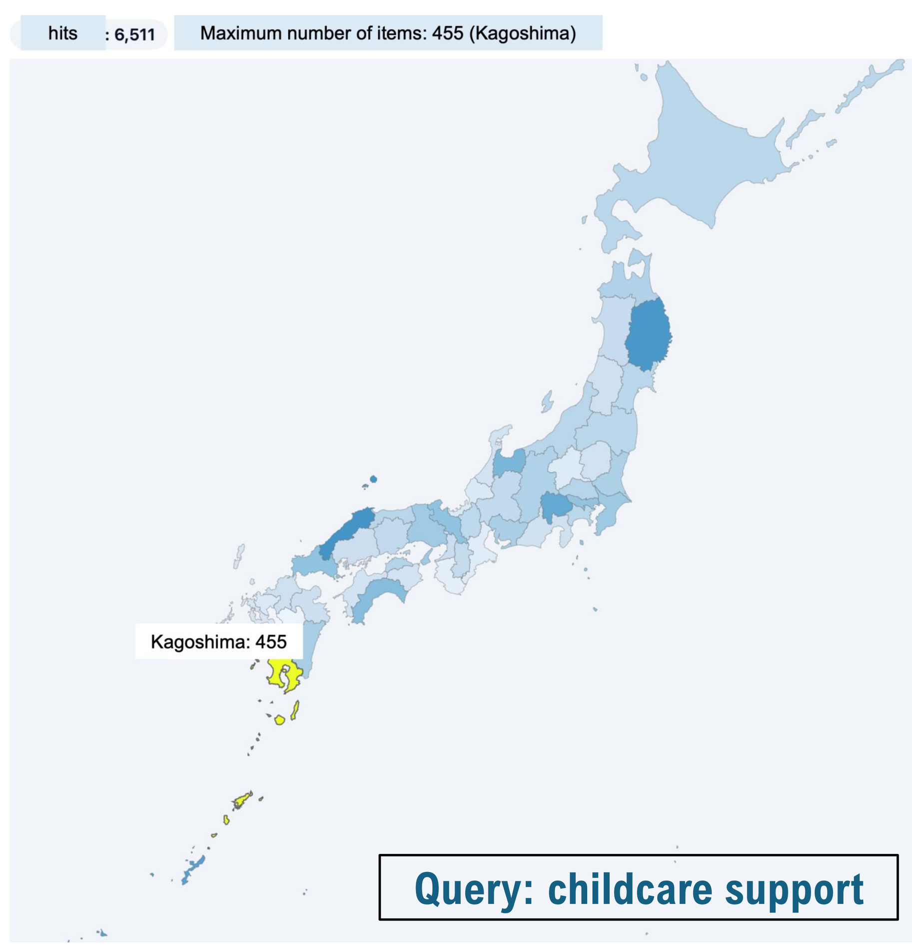
A word ***hagutte***('to turn a page') — embedded in a formulaic parliamentary expression — was automatically identified as a regional marker in **Tottori Prefecture**. It is thought to be **a dialect feature** used without awareness in formal settings.

Prefecture	Freq.
Tochigi	1
Chiba	1
Niigata	5
Tottori	224
Okayama	9
Total	236

Keiichi Takamaru and Hokuto Ototake. 2023. An exploratory survey of dialects in minutes of local assemblies: A dialectological study using machine learning and XAI (in Japanese). *Studies in Dialects*, 9:27–51.

Giimiru: a Web-based Search and Visualization System

Giimiru provides public access via five functionalities: **keyword search**, **semantic similarity search**, **geographic visualization**, **time-series analysis**, **and cross-tabulation** — supporting research in sociolinguistics, political science, and econometrics.



Example of geographic visualization

Hokuto Ototake, Hiroki Sakaji, Keiichi Takamaru, Akio Kobayashi, Yuzu Uchida, and Yasutomo Kimura. 2018. Web-based system for Japanese local political documents. *International Journal of Web Information Systems*, 14(3):357–371.

Conclusions and Future Issues

- We constructed a longitudinal cross-regional corpus of Japanese prefectural assembly minutes spanning 12 years and three electoral terms across all 47 prefectures.
- Comparative analysis revealed systematic changes in member composition and speech patterns: the proportion of younger members has declined, female representation has gradually increased, and female members consistently speak more characters than male counterparts across all age groups.
- Longitudinal data spanning three terms enabled analyses impossible with single-term datasets, such as tracking non-speaking rates across career stages.
- The corpus supports diverse applications in computational linguistics, political science, and dialectology. We plan to continue expanding the dataset and integrating it into Giimiru v2 to promote interdisciplinary use.

Data Availability

The datasets are accessible for research purposes upon request. Public query access is also available through the web-based interface **Giimiru**.

http://local-politics.jp/

This work was supported in part by JSPS KAKENHI Grant Numbers 25K04045 and 24K15196, and by JST RISTEX Grant Number JPMJRS25L2.