Potayto or potahto, jagaimo or bareisho? Japanese dialect classification

Elaine Chou (eschou), Stacey Huang (sahuang), and Ningrui Li (ningruil)

Introduction

Dialects are subsets of a language delineated by geographic/social boundaries, and may or may not be mutually intelligible. Classifying dialects is often difficult and even contentious.

Japanese dialect classification has historically relied on limited components of the language [1, 2]. We aim to create a machine learning model that classifies dialects using more comprehensive measures. Such studies could serve as a starting point towards easing controversy surrounding dialect classification.

Dataset

We used the dataset from the "Field Research Project to Analyze the Formation Process of Japanese Dialects" (FPJD) study done by the National Institute for Japanese Language and Linguistics (NINJAL). Responses to 211 prompts were collected from 554 locations. These prompts assessed how dialects in those regions differed in grammatical structure, pronunciations, words used for common nouns, and so on.

Sample prompts	Aspect of interest	Example answers
What do you call a tuber like this?	Noun	Jagaimo (じゃがいも) Bareisho (馬鈴薯) Imokko (芋っこ)
When saying, "It's 10 o'clock and they haven't come yet", how would you say "haven't come yet"?	Grammar (negative conjugation)	Konai (来ない) Kunee (くねえ) Kinaka (きなか)

Table 1: Example prompts from the survey.

A feature vector was created for each prompt response. We included 33 features that correspond to different pronunciations of key vowels and consonants, as well as linguistic features, such as glottal stops.

The feature vectors for each prompt were combined, and its first fifty principal components were used for analysis.

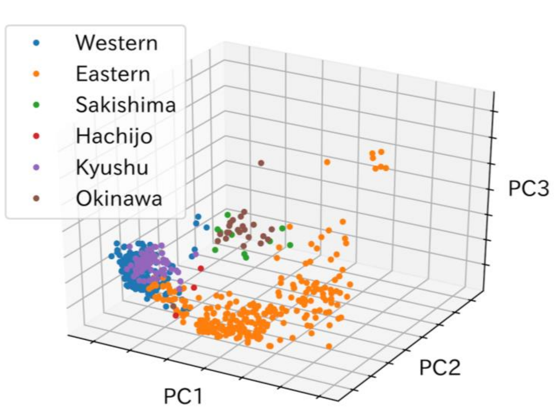

Fig. 1: First three principal components of feature vectors (labelled by dialect region).

Supervised Approaches

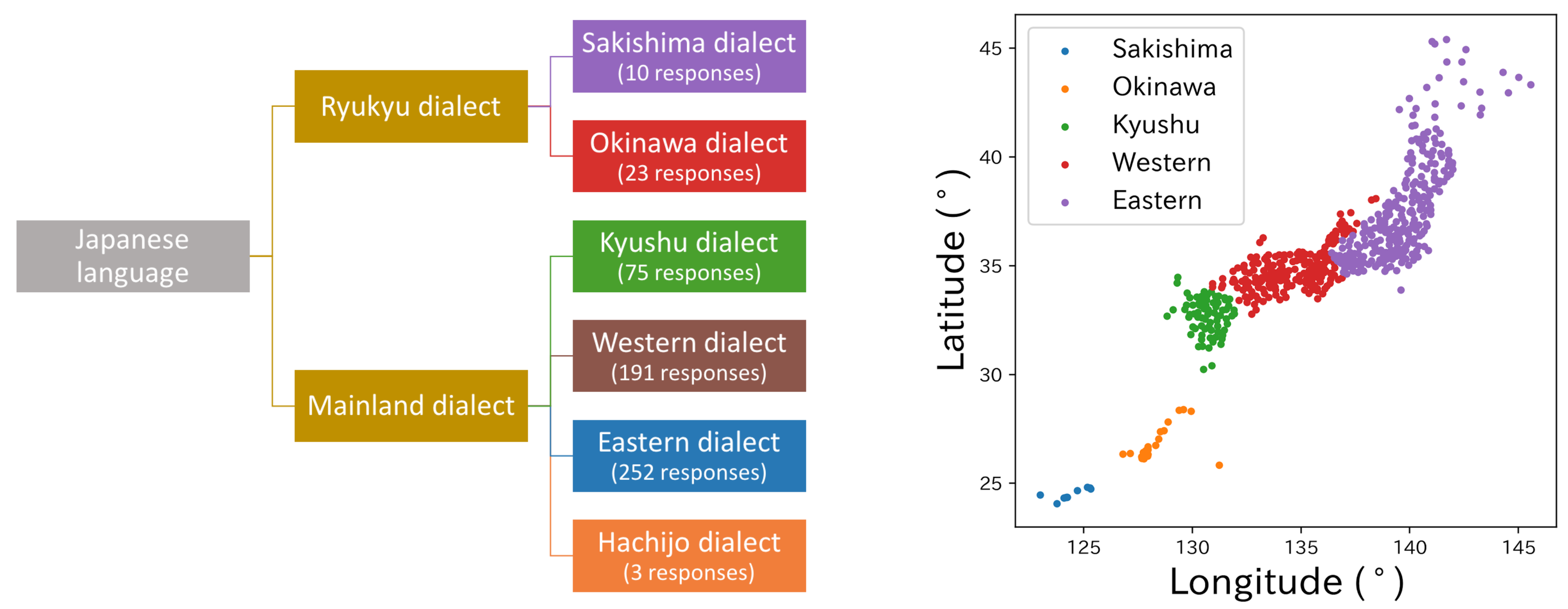

Fig. 2: Examples labeled using a dialect map created by Japanese linguist Hirayama Teruo [2]. Responses from the Hachijo region were not considered because there were only three of them.

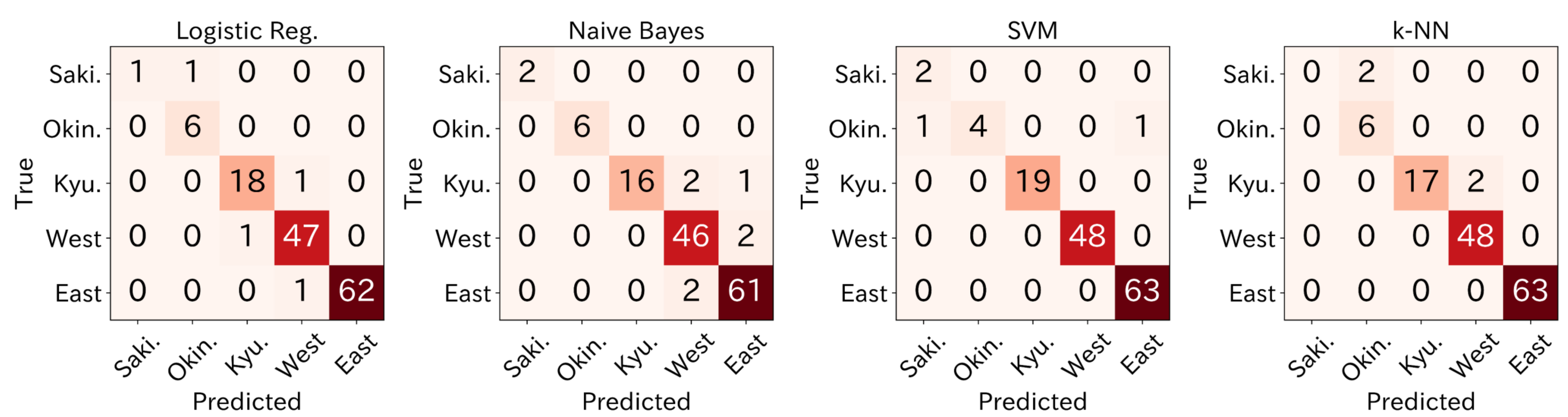

Fig. 3: Confusion matrices for various classifiers (train/test = 75%/25%). Hyperparameters for each model were chosen using an exhaustive grid search with 5-fold cross validation.

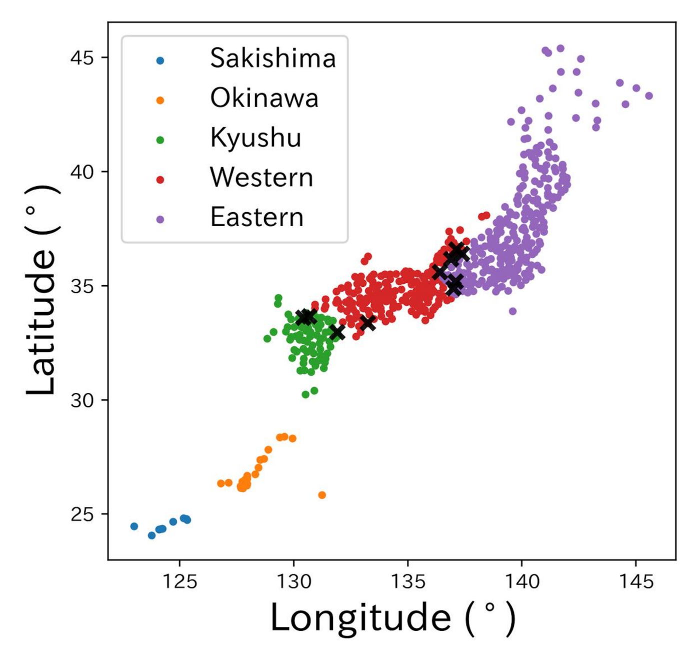

Fig. 4: Dialect map with x's marking ten examples that were most commonly misclassified by an SVM.

SVM (RBF kernel, $\gamma = 0.001$, $C = 10$) performed the best out of the four classifiers, with a test error of 1.45%.

Misclassified examples occurred mostly around region borders (Fig. 4). However, the "borders" are ambiguous as there is likely linguistic mixing between regions.

Prompts addressing language appeared to be the most useful for differentiating between dialect regions (Fig. 5).

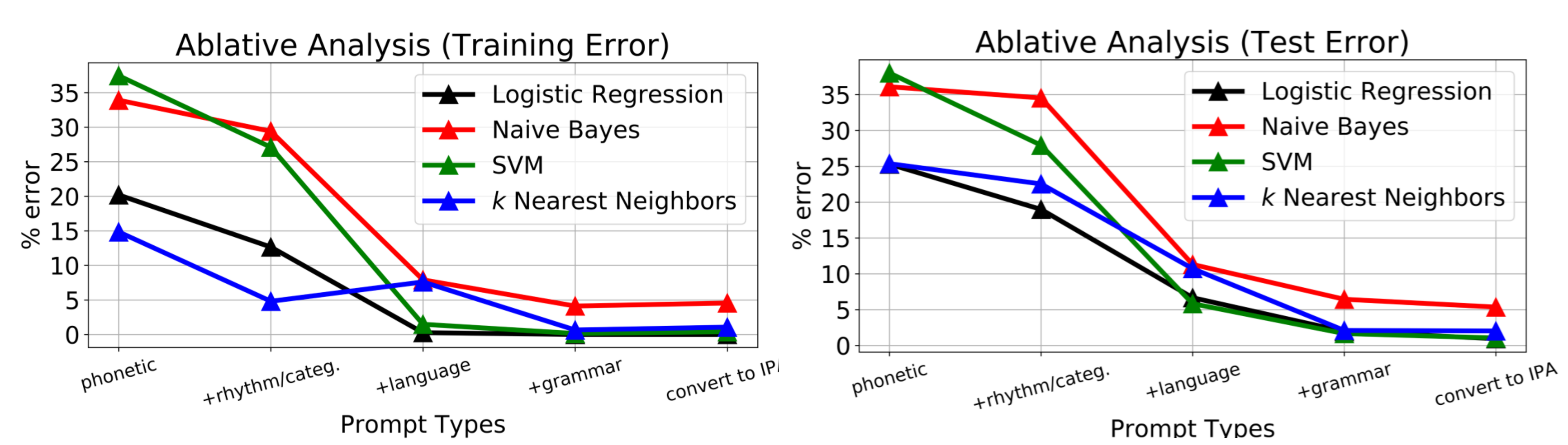

Fig. 5: Ablative analysis for each classifier type. At each step, prompts addressing different linguistic features were added. At the last step, all responses were converted from katakana to the International Phonetic Alphabet (IPA) before featurization.

Unsupervised Approaches

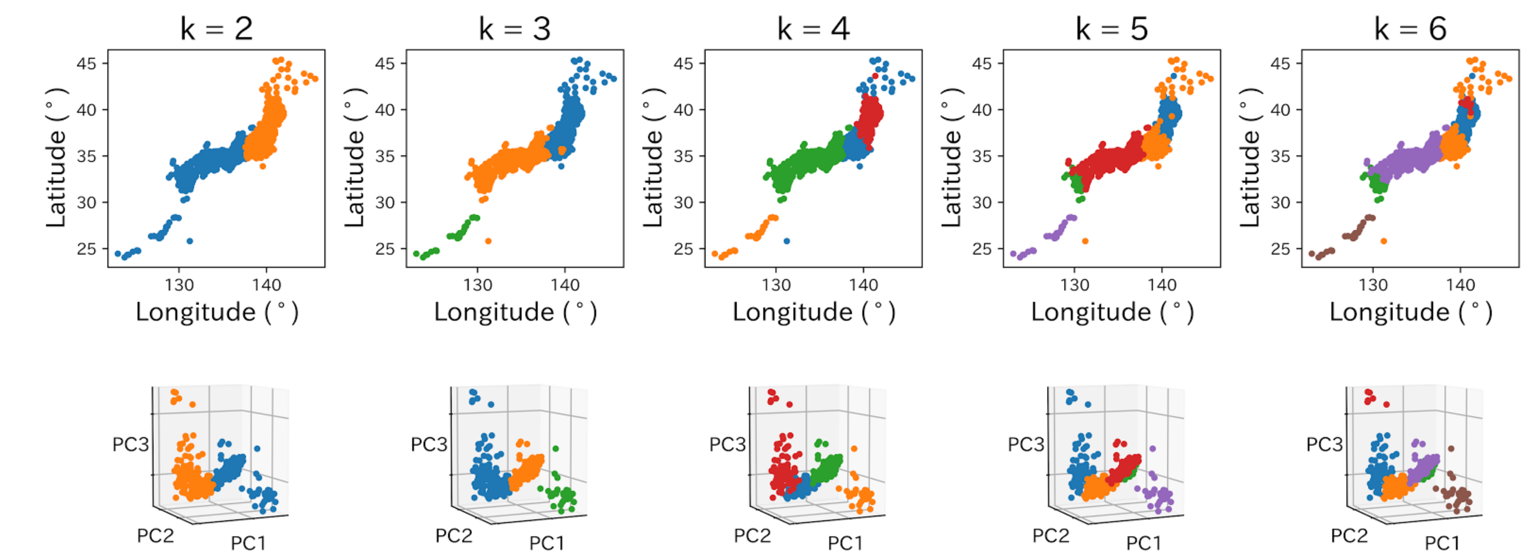

Fig. 6: *k*-means clustering results (top: geographical labels, bottom: clustering in PCA space).

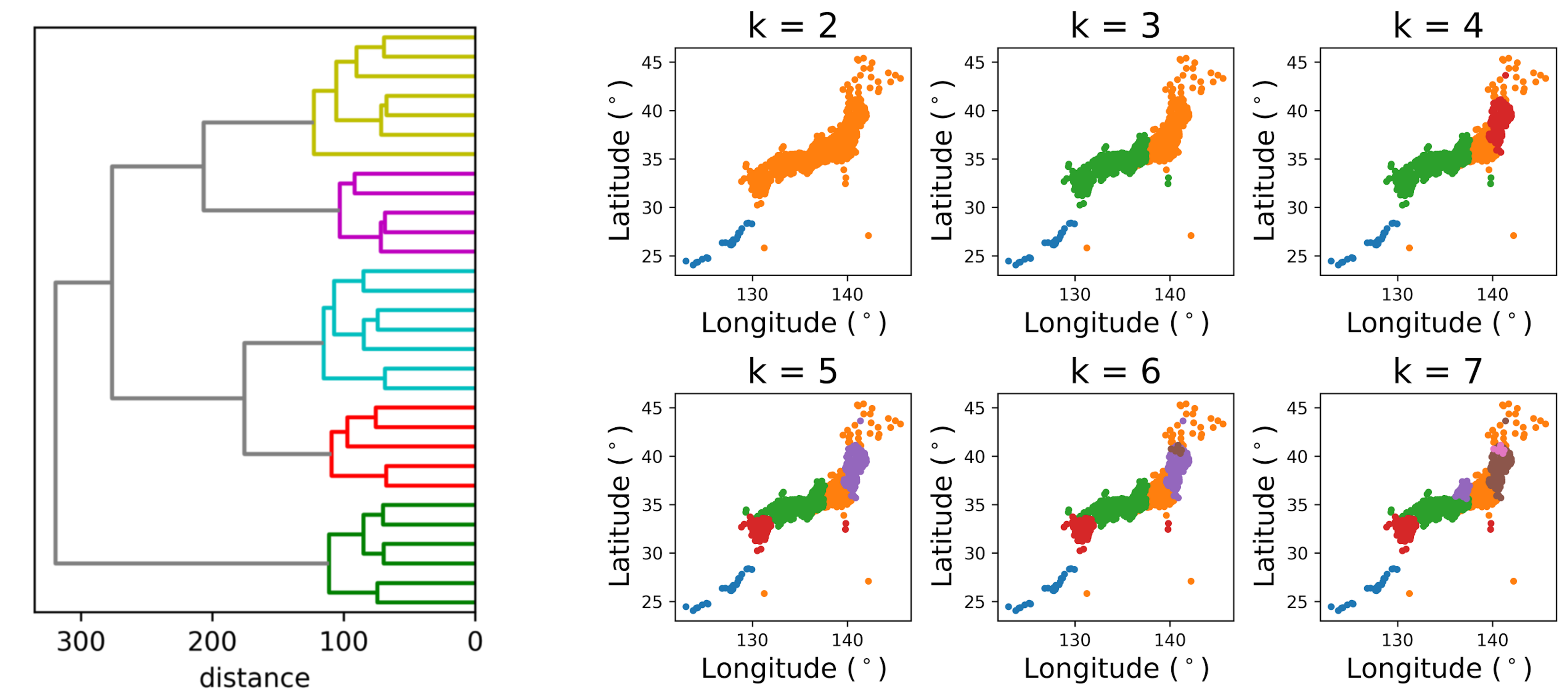

Fig. 7: Truncated decision tree (left) and geographical labeling (right) for hierarchical clustering directly on unfeaturized IPA responses.

Unsupervised learning was able to discern differences in dialects between Sapporo (major city) and the rest of Hokkaido (Fig. 6 and 7).

Conclusions

Machine learning models can efficiently synthesize the linguistic richness in dialects, reducing some of the inherent difficulty in dialect classification.

While supervised techniques successfully classified among geographic (island) and political (provincial) boundaries, the unsupervised approaches were able to also pick up subtler linguistic differences, providing evidence that dialect regions are not always associated with geographical or political boundaries.

Exploring more sophisticated features could uncover important aspects that are currently overlooked in dialect classification. While this study only analyzed transcriptions, these methods could also be extended to audio speech samples.

References

[1] S. Abe, "The classification and division of Japanese dialects", Jinbun 13(21-55), Gakushuin University, 2015. Retrieved from http://ci.nii.ac.jp/naid/110009889230

[2] Y. Kawaguchi and F. Inoue, "Japanese Dialectology in Historical Perspectives," Revue belge de Philologie et d'Histoire, vol. 80, no. 3, pp. 801-829, 2002.

[3] T. Onishi, "Mapping Japanese dialects," Dialectologia, Special Issue I, pp. 137-146, 2010. Retrieved from http://www.raco.cat/index.php/Dialectologia/article/viewFile/242107/324719.

[4] National Institute for Japanese Language and Linguistics (Corpora and Databases). Retrieved from https://www.ninjal.ac.jp/english/database/subject/diversity/.