Potayto or potahto, jagaimo or bareisho? Japanese dialect classification

Project Category: Natural Language

Elaine Chou, Stacey Huang, Ningrui Li

Author e-mail addresses: eschou@stanford.edu, sahuang@stanford.edu, ningruil@stanford.edu

1. Introduction

Dialects are subsets of a language that are often delineated by geographic or social boundaries, and they are characterized by a range of idiosyncratic features. As a result, they may or may not be mutually intelligible. For example, most English dialects are comprehensible to other native speakers, but Chinese dialects can be more dissimilar from one another than some officially recognized languages, e.g. Spanish from Portuguese or Italian.

Dialect differentiation and classification is difficult and often even contentious due to the complexities of language. Historical and social factors as well as questions of identity are equally important for defining boundaries between dialects [1,2,3]. In Japan, debate over dialect classification continues between linguists, geolinguists, and other scholars; historically, Japanese dialect classification has relied on limited components of the language (e.g. using only word pronunciations or grammar to distinguish dialects) [4,5].

In this study, several machine learning models are used to classify Japanese dialects using a broad set of linguistic features. These studies can serve as a starting point to ease the controversy surrounding dialect classification. Various linguistic features were derived from survey responses obtained from different regions in Japan, and these features were used as inputs to the learning models. We first evaluated the performance of several supervised learning models (logistic regression, Naive Bayes, support vector machine (SVM), and *k*-nearest neighbors (*k*-NN)) to predict dialect labels based on these linguistic features. We also assessed the dialect clustering generated by unsupervised learning methods, such as *k*-means clustering, Gaussian mixture models (GMMs), and hierarchical clustering.

2. Related Work

Dialect mapping in Japan began in the 8th century with primitive maps and limited dialect coverage. Starting in the late 1800s, Japanese linguists created more extensive maps by manually differentiating dialects based on linguistic aspects such as pronunciation, grammar, and type of polite speech used [4,5]. More recent work on classifying Japanese dialects has also considered the historical context in which language has developed in Japan [6,7]. Kawaguchi and Inoue studied the development of standard Japanese, conducting cluster analysis and factor analysis on frequency of word usage and century of appearance [6]. Inoue subsequently expanded on this work by considering railroad distance between cities and prefectures as an additional factor in dialect dissemination and similarity [7].

In many languages, including Irish, Dutch, and Bulgarian, research has been conducted on classifying dialects using linguistic distance metrics [8,9,10]. Nerbonne and Heeringa used hierarchical clustering to classify Dutch dialects based on Levenshtein distance, which corresponded to the number of character changes required to transform one word into another [8]. Heeringa and Braun conducted similar analysis but defined distance using the Almeida-Braun system, a type of weighted Levenshtein distance [9]. Nerbonne et al. found that bootstrapping methods increased the stability of unsupervised clustering techniques, which otherwise suffer from large sensitivity to minor changes in input features and the distance metric used [10].

Dialect classification using speech transcripts has also been conducted [11,12]. Huang and Hansen used GMM's for English and Spanish dialect classification and outperformed human listeners [11]. Saul and Rahim found that pairing factor analysis with Hidden Markov Models and then conducting Maximum Likelihood significantly improved speech recognition [12].

3. Dataset and Features

The dataset comes from the National Institute for Japanese Language and Linguistics (NINJAL), and it was collected by surveyors who traveled throughout Japan as part of the "Field Research Project to Analyze the Formation Process of Japanese Dialects" (FJPD) [13]. This dataset contains roughly a total of 146,000 responses to 211 different prompts, and they were accumulated across 544 locations throughout Japan.

The survey prompts were selected to assess several aspects of the Japanese language (Table 1), such as the pronunciation certain words, the nouns used to describe certain objects, the type of polite language that is used, and the rhythm of certain words. The linguistic aspect assessed by each prompt was included in the dataset, and some example prompts used in the survey are shown in Table 1. The survey responses were transcribed and recorded along with its corresponding location (geographic coordinates, prefecture, and city name).

Prompt	**Aspect of interest**	**Example responses**
What do you call a tuber like this?	Language use – 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. Sample prompts and responses, including studied linguistic feature of interest.

Feature vectors were created for each response based on the linguistic feature being assessed. Response transcriptions were inconsistently recorded using either the International Phonetic Alphabet (IPA) or Japanese kana, so a mapping was created for converting kana to IPA. In most cases, a dictionary was used to map prompt responses to a vector of 33 features based on the presence of certain linguistic features. This dictionary included vowels, certain consonants, and other linguistic features like glottal stops. Each dictionary entry mapped the different forms of each feature to an integer based on their similarity. For example, ë maps to a 1, ẽ maps to a 2, and ẹ maps to a 4, since ë is closer to ẽ than it is to ẹ.

Some geographical locations had multiple responses. In addition, certain respondents gave no response or had no word for the item or phenomenon for certain prompts. These entries were removed in pre-processing. Additionally, longer responses were separated into two parts, since our feature vector does not account for repetitions.

4. Methods

4.1 Supervised techniques

Survey responses were first manually labeled by cross-referencing each survey location with a dialect map created by Japanese linguist Hirayama Teruo [4]. Teruo split the Japanese language into six dialect groups shown in Fig. 1(a). The locations of all survey responses are shown in Fig. 1(b), and each response is labeled using Teruo's dialect map. There were only three survey responses corresponding to the Hachijo dialect, so it was not considered as part of this analysis.

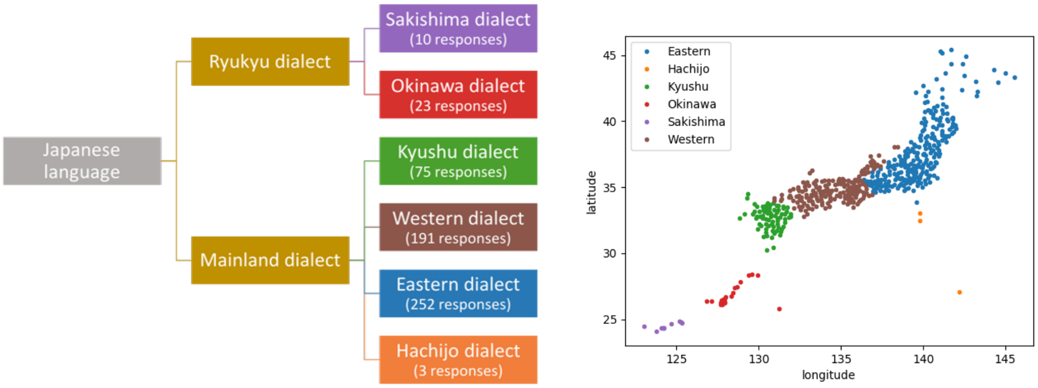

Figure 1. (a) Teruo's dialect tree for the Japanese language, along with the number of responses from each dialect region. (b) Map with dialect labels of locations of the survey responses.

The feature vectors from each prompt were concatenated into a single vector. Principal component analysis (PCA) was used on this vector to reduce dimensionality. 5-fold cross validation was used to choose how many dimensions to keep (50). The data were split into training (75%) and test (25%) sets, and following classifiers were explored: one-vs-rest logistic regression, Naive Bayes, SVM, and *k*-NN.

In logistic regression, the sigmoid function $h_\theta(x) = \frac{1}{1+e^{-\theta^T s}}$ is used as the hypothesis function to give the probability that an example is in one of two classes. Fitting the parameter θ in logistic regression can be interpreted as likelihood maximization on independent training examples and results in the batch gradient ascent update equation $\theta := \theta + \frac{\alpha}{m}\sum_{i=1}^{m} y^{(i)} - h_\theta(x^{(i)}))x^{(i)}$, where α is the learning rate, m is the number of examples, and $x^{(i)}$ is the i^{th} input vector with label $y^{(i)}$. For problems with $k > 2$ classes, one-vs-rest logistic regression fits k parameters θ_i ,$1 \le i \le k$ that each distinguish between class *i* and all other classes. Examples are then classified as $\hat{y} = \underset{i}{\operatorname{argmax}}\ h_{\theta_i}(x)$.

Naive Bayes uses the naive assumption that all features are independent. Applying Bayes rule results in: $P(y|x_1,\dots,x_n) = \frac{P(y)P(x_1,\dots,x_n|y)}{P(x_1,\dots,x_n)} = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1,\dots,x_n)}$, where x_i is the i^{th} component of an input vector x. The classification rule is $\hat{y} = \underset{y}{\operatorname{argmax}}\, P(y)\prod_i P(x_i|y)$. We tested Naive Bayes on the original full-dimension data with multinomial conditional probability distributions and Laplace smoothing, as well as on the dimension-reduced data with Gaussian distributions, because PCA caused the feature values to no longer be discrete.

SVMs maximize the margin between classes by performing the optimization $\min_{w,b} \frac{1}{2}|w|^2 + C\sum_{i=1}^{m}\xi_i$ s.t. $y^{(i)}(w^T\phi(x^{(i)}) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1,\dots,m$, where $\phi(x)$is a kernel and C is the regularization cost for points crossing the support vectors. The SVM used in this study had a radial basis function (Gaussian) kernel $K(x,z) = \exp(-\gamma|x-z|^2)$ with C = 10 and γ = 0.001; these hyperparameters were estimated using an exhaustive grid search with 5-fold cross validation.

For *k*-NN, the Euclidean distances between each example in the test set and all examples in the training set are computed. The test set is assigned the majority label of its *k* closest training set examples. The optimal number of neighbors was estimated to be $k = 3$ using 5-fold cross validation.

4.2 Unsupervised techniques

After using supervised learning to verify sufficient data volume and effective feature choices, we investigated unsupervised techniques: *k*-means, GMMs, and hierarchical clustering. *k*-means groups the data into *k* clusters by first randomly choosing data points to be cluster centroids μ_j. Each example is assigned to its closest centroid $c^{(i)} = \underset{j}{\operatorname{argmin}}\left\|x^{(i)} - \mu_j\right\|^2$and each centroid is redefined as the mean of data points assigned to it, $\mu_j = \frac{\Sigma_{i=1}{}^{m} 1\{c^{(i)}=j\}x^{(i)}}{\Sigma_{i=1}{}^{m} 1\{c^{(i)}=j\}}$. This process is repeated until convergence.

Gaussian mixture models assume that each point $x^{(i)}$is drawn from one of *k* multivariate Gaussian distributions randomly chosen from latent variable $z^{(i)} \sim$ *multinomial*. Using the expectation-maximization (EM) algorithm, $z^{(i)}$ is first estimated by choosing $w^{(i)} = p(z^{(i)} = j|\, x^{(i)}; \phi, \mu, \Sigma)$, then the parameters are updated according to $\phi_j = \frac{1}{m}\Sigma_{i=1}{}^{m} w_j{}^{(i)}$, $\mu = \frac{\Sigma_{i=1}{}^{m} w_j{}^{(i)}\, x^{(i)}}{\Sigma_{i=1}{}^{m} w_j{}^{(i)}}$, and $\Sigma_j = \frac{\Sigma_{i=1}{}^{m} w_j{}^{(i)}\,(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T}{\Sigma_{i=1}{}^{m} w_j{}^{(i)}}$. These steps are repeated until convergence.

Hierarchical clustering requires only a distance matrix between points rather than the points themselves, so the clusters could be formed directly using the responses, rather than using feature vectors derived from the responses. Following a similar method used by Nerbonne and Heeringa [8], for each survey prompt, a distance matrix was formed by computing the Levenshtein distance between response pairs. To combine the prompts, the sum of log distances across prompts is computed (taking the log suppresses larger distances). With the distance matrix as input, hierarchical clustering iteratively combines the two clusters that result in the smallest increase in distance until all clusters merge. Merging methods considered in this study are nearest point, farthest point, UPGMA, and Ward.

5. Experiments, Results, and Discussion

5.1 Supervised techniques

1. **Baseline.** A simple baseline is to always predict the most common class, which results in an error of ~50%.
2. **Different classifiers.** Logistic regression performed the best out of the four classifiers tested (confusion matrices shown in Fig. 2, training and test errors shown in Table 2).

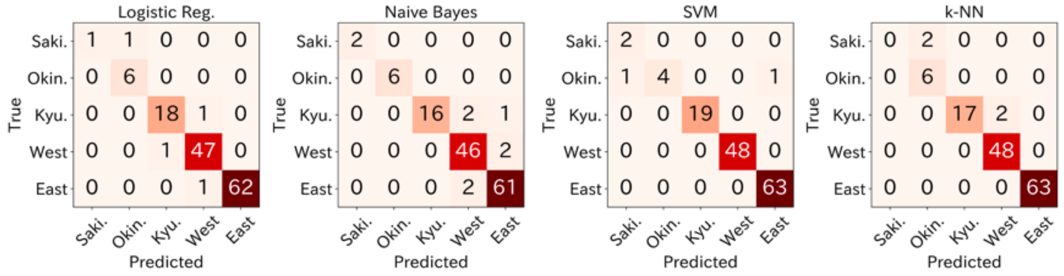

Figure 2. Test set confusion matrices for (a) logistic regression, (b) Naive Bayes, (c) SVM, and (d) *k*-NN.

	Logistic Regression	**Naive Bayes**	**SVM**	***k*-NN**
Training / test errors (%)	0% / 1.74%	4.82% / 7.10%	0.05% / 2.10%	2.01% / 2.75%

Table 2. Training / test errors (%) for each classifier.

3. **Ablative analysis.** We analyzed the impact of certain prompts for predicting dialect classes using an ablative analysis (Fig. 3). Test errors were averaged over 20 runs for each case. Prompts assessing language lead to the greatest accuracy improvement across all methods.

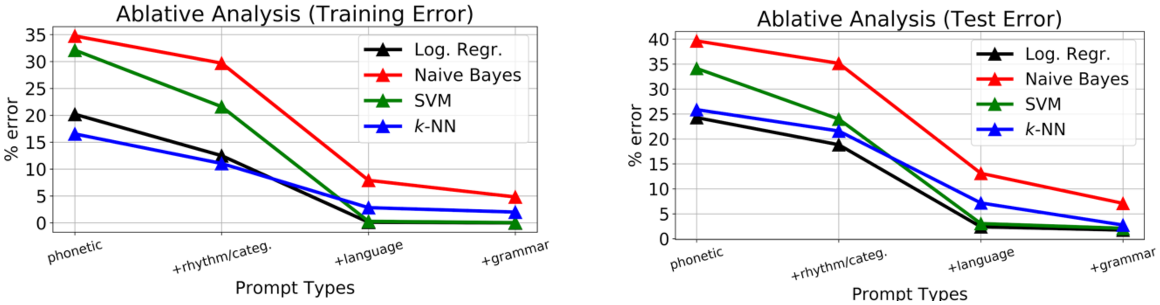

Figure 3: Ablative analysis for training and test data for each classifier type. At each step, prompts addressing different linguistic features were added.

4. **Top misclassified examples.** We ran the SVM classifier 100 times, and the top ten most commonly misclassified examples were marked (Fig. 4). Misclassified examples occurred most commonly around region borders. However, there is clear ambiguity around dialect families around region boundaries, where linguistic mixing occurs, so these labels are themselves also ambiguous.

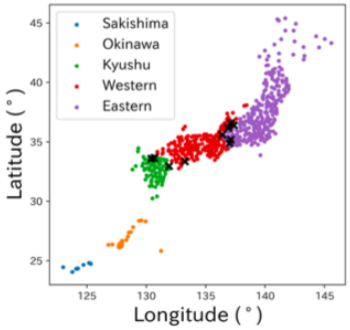

Figure 4: Dialect map with x's marking ten examples that were most commonly misclassified by SVM.

5.2 Unsupervised techniques

1. **Fuzzy *k*-means clustering.** Clustering results for *k*-means are shown in Fig. 5(a). Visually, the formed clusters are distinct and generally fall along political boundaries, such as prefecture borders, and geographical boundaries, like mountains or islands. Fuzzy *k*-means was used to assess the confidence of each example belonging to a cluster (Fig. 5(b)). Examples close to region boundaries tended to be less confidently labeled.

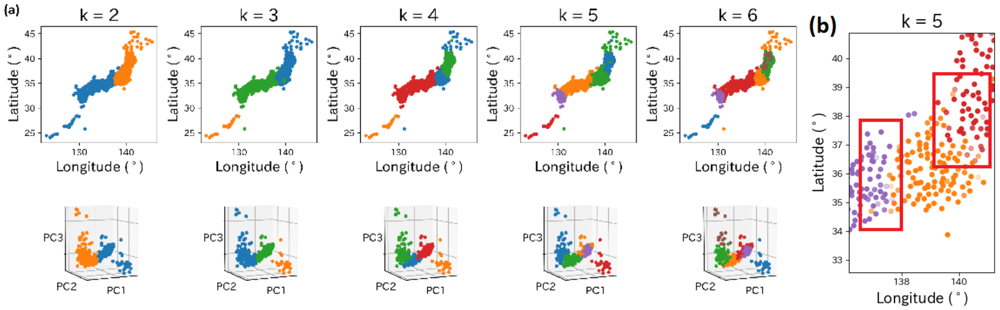

Figure 5: (a) *k*-means clustering results (top: geographical labeling of clusters, bottom: clustering in PCA space). (b) Results of fuzzy *k*-means clustering (less confident labels correspond to lighter colors).

2. **GMM likelihood on held-out data.** Before running GMM, we split the data into training and test sets as in supervised learning, then ran likelihood tests on the test set. GMM clustering results for $k = 5$ are shown in Fig. 6(a); data points with scores (log probabilities of belonging to a cluster) below -500 were indicated with an x. As before, these points occur more commonly among boundaries, as well as in the city of

Sapporo, which may have additional dialect mixing due to its current migration patterns and history. On average, ~99% of the maximum likelihood of test samples equaled 1. The posterior likelihood and log likelihood scores are plotted in Fig 6(b).

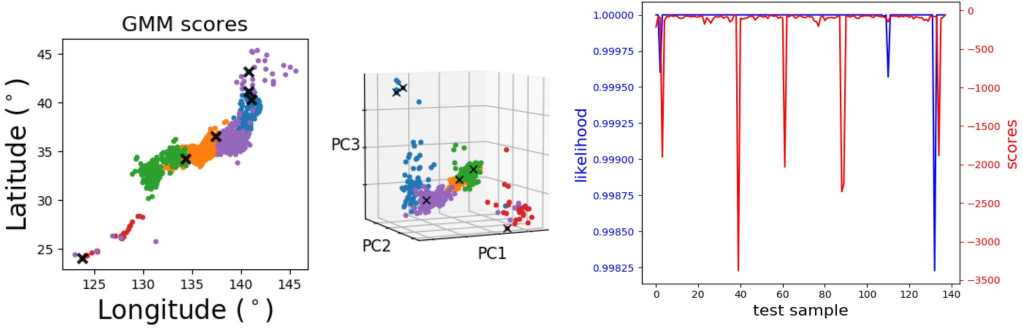

Figure 6: (a) GMM label map and PCA plot with x's marking data examples with low scores. (b) Plot of the likelihood and score for each test sample.

3. **Distance metric in hierarchical clustering.** Hierarchical clustering results varied dramatically depending on the distance metric used to merge clusters. Ward's method produced the most balanced tree and meaningful clusters (Fig. 7). The dendrogram resulting from hierarchical clustering reflects the major structure of many Japanese dialect trees, first splitting the Ryukyu Islands from the mainland, then dividing the mainland into East and West.

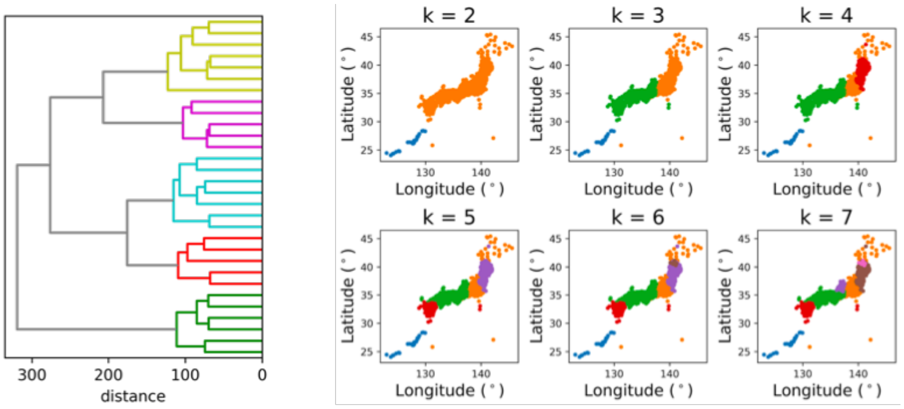

Figure 7: Truncated dendrogram (left) and geographical labeling (right) for hierarchical clustering directly on unfeaturized IPA responses using Ward's method.

In all unsupervised approaches, the differences in the dialects between Sapporo, the major city in Hokkaido, and the rest of the northern island were found. This matches historical patterns of colonization of Hokkaido that originated from the Tokyo area in the late 19th century, while Sapporo has continued to be settled by people from the Tohoku region of northern Japan [14].

6. Conclusions and Future Work

Machine learning models can efficiently synthesize the linguistic richness in dialects, reducing some of the inherent difficulty in dialect classification. For supervised learning, logistic regression performed the best, which may suggest that the data is divided into classes with the probability of belonging to any class given by sigmoid functions. While supervised techniques successfully classified among geographic and political 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. For example, our current feature vector does not consider repetition of symbols or multiple types of symbol occurrences. Finally, while this study only analyzed text transcriptions, these methods could also be extended to audio speech samples.

Contributions

Elaine Chou: Literature review, data labels, formatting/interpreting prompt responses, kana to IPA conversion, initial supervised learning, ablative analysis, PCA analysis, distance metrics, hierarchical clustering, GMM analysis

Ningrui Li: Data parsing, hyperparameter optimization for SVM and and k-NN using 5-fold cross validation, evaluation metrics, results visualization, k-means, Python mastery

Stacey Huang: Dataset, literature review, development of feature vectors and dictionary, conversion of kana/kanji script, data cleaning, normalization, GMM

References

[1] C. Mallison and W. Wolfram, "Dialect accommodation in a bi-ethnic mountain enclave community: More evidence on the development of African American English," Language in Society, vol. 31, no. 5, pp. 743–775, 2002.

[2] R. Solheim, "Dialect development in a melting pot: The formation of a new culture and a new dialect in the industrial town of Høyanger," Nordic Journal of Linguistics, vol. 32, no. 2, pp. 191–206, 2009.

[3] R. E. Callary, "Phonological change and the development of an urban dialect in Illinois," Language in Society, vol. 4, no. 2, pp. 155–169, 1975.

[4] 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

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

[6] 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.

[7] F. Inoue, "Multivariate analysis, geographical gravity centers and the history of standard Japanese forms," Area and Culture Studies, vol. 68, pp. 15-36, 2004.

[8] J. Nerbonne and W. Heeringa, "Measuring Dialect Distance Phonetically," Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON-97), pp. 11-18, 1997.

[9] W. Heeringa and A. Braun, "The use of the Almeida-Braun system in the measurement of Dutch dialect distances," Computers and the Humanities, vol. 37, pp. 257–271, 2003.

[10] J. Nerbonne et al., "Projecting dialect distances to geography: bootstrap clustering vs. noisy clustering," *Data Analysis, Machine Learning, and Applications,* pp. 643-654, 2008.

[11] R. Huang and J. Hansen, "Unsupervised discriminative training with application to dialect classification," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, pp. 2444 - 2453, 2007.

[12] L. K. Saul and M. G. Rahim, "Maximum likelihood and minimum classification error factor analysis for automatic speech recognition," IEEE Transactions on Speech and Audio Processing 8.2, pp. 115-125, 2000.

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

[14] F. Inoue, "Sprachraum and infrastructure: Abstracting geographical space via railway distance", *Language and Space: An International Handbook of Linguistic Variation*, vol. 2, part 1, pp. 542-560.