Classification of Kansai dialects using phonetic distance

Some Japanese dialectologists have attempted to classify Japanese dialects using vocabulary, suffix, vowel shift, and proportion of standard Japanese use (e.g., Inoue, 2001; Kindaichi, 1964; Tojo, 1953). They descriptively revealed some variations in the dialects and classified them. In order to further understand how the dialects vary, we implemented a phonetics-based metric and examined how Kansai dialects spoken by old generations are classified using a density-based clustering method called HDBSCAN (Campello et al., 2013)

We implemented *aline* distance (Downey et al., 2008) in order to measure phonetic distances between words in standard Japanese and Kansai dialects. The advantage of aline algorithm is that it considers weight of phonetic features (e.g., dental, palatal, nasal, front, and central) when it compares pairs of words, and generates distance scores. Downey et al. (2008) demonstrated to what extent words in languages spoken in eastern Indonesia (i.e., Rindi and Sumba) phonetically diverged from their cognates (proto-Austronesian words). Here, we examined whether aline distance becomes a metric for classifying dialectal variation.

We designed a questionnaire containing 76 questions in order to elicit written forms of words, phrases, and sentences. We distributed it to 791 Japanese speakers in Kansai areas (i.e., Hyogo, Kyoto, Mie, Nara, Osaka, Shiga, and Wakayama) by post and received answers from them. We converted the written data to International Phonetic Alphabets, and calculated aline distance using the alineR package (Downey et al., 2017) in R language (R Core Team, 2021). We removed all participants who missed answering any question in the questionnaire, leaving 491 participants.

We initially ran a principle component analysis (PCA) and identified that ten words (*watashi*, *ore*, *shindoi*, *gokiburi*, *hikigaeru*, *benjo*, *higanbana*, *katazukeru*, *konai*, and *hisashiburi*) repetitively appeared in the retained PCA components. Hence, we selected these words and ran PCA again. The results of Horn's parallel analysis demonstrated that four PCA components to be retained in further analyses. The loadings of principle components (contribution > 10%) in each components did not overlap among the four principle components except PC4.

In order to identify the optimal HDBSCAN model (tuning the hyper parameter, *minPts*), we assessed internal cluster metrics (e.g., Calinski Harabasz, CDbw, Dunn, and Silhouettte) and external cluster metrics (e.g., Czekanowski Dice, Folkes Mallows, and Jaccard) with 500 bootstrap samples. The results demonstrated that the optimal minPts was 31, and that there were three clusters among the Kansai-dialect speakers. One group (n = 164) resided in the Japan-Sea side of Hyogo, Kyoto, and Shiga prefectures as well as Mie prefectures. Another group (n = 100) resided in

the southern part of Osaka and Nara, and Mie prefecture. The other group (n = 36) generally resided western part of Kansai area.

This study identified lexical items, which play a role in phonetically investigating dialectal variation, and demonstrated that Kansai dialects can be classified into three groups. In addition, this study suggests that phonetic distance (aline) can be another metric to understand how dialects vary in a language.

(486 words)

References

Campello, R., J. G. B., Moulavi, D., & Sander, J. (2013). Densiity-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 160–72. Springer. doi:10.1007/978-3-642-37456-2_14

Downey, S. S., Sun, G., & Norquest, P. (2017). alineR: an R package for optimizing feature-weighted alighments and linguistic distances. *The R Journal*, **9** (1), 138-152.

Downey, S. S., Hallmark, B., Cox, M. P., Norquest, P., & Lansing, J. S. (2008). Computational feature-sensitive reconstruction of language relationships: developing the ALINE distance for comparative historical linguistic reconstruction. Journal of Quantitative Linguistics, 15(4), 340-369.

Inoue, F. (2001). *Keiryōgakuteki hōgengaku*. Tokyo, Meijishoin.

Kindaichi, H. (1964). *Watashi no hōgenkukaku*. In M. Tojo (Ed.), Nihongo Hōgenkukaku, Tokyo, Tokyodo shuppan.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Tōjo, M. (1953). Nihon hōgengaku. Tokyo, Hirokawakōbunkan.