Deriving gender from phonemes: Evidence from Mandarin names using the Random Forest algorithm

Chun Hau Ngai[1] & Alexander Kilpatrick[2]
[1]*Indiana University, Bloomington,*
[2]*Nagoya University of Commerce and Business*
chngai@iu.edu, Alexander_Kilpatrick@nucba.ac.jp

This study examine the whether systematic sound pattern reported in English name held in Mandarin Chinese (hereafter: Mandarin). Unlike English, Mandarin names are disyllabic, each represented by a logograph, while strongly associated with a semantic concept. Psycholinguistic studies have suggested that phonology has a lesser role in the naming of Mandarin characters in comparison to Indo-European languages (Zhang et al., 2009); thus, rendering Mandarin a stronger test to the systematic sound patterns previously reported in English.

Methods: 212 most common Mandarin given names (115 female & 97 male) from Bao & Cai (2021) was examined. The presence or absence of a phoneme was coded as binary measure. The random forest (Breiman, 2001), a machine learning algorithm, was used to examine systematic sound patterns in Mandarin names. The accuracy of a random forest is determined by feeding its testing subset (two-third data) into the model and observing the error rate which is the percentage of samples that the algorithm was unable to accurately classify. Feature importance was examined to evaluate the contribution of phonemes to classification.

Results & Conclusion: The random forest was able to accurately classify 80.28% of the testing subset sample into their allocated gender category, suggesting phonology alone is adequate in predicting gender adaptation (OBB = 19.71%). Details on feature importance could be found in the long abstract. Overall, phonological patterns previously reported have also been found to be predictive of gender identification. Sonorants and high front vowel, with the addition of /ɕ/, were important to the identification of female names. Obstruents and low back vowels, on the contrary, were predictive of male names. Contrary to van de Weijer et al. (2020), our results suggest that tones do play a role in the identification of gender.

References: • Bao, H.-W.-S., & Cai, H.-J. (2021). Psychological and behavioral effects of personal names in real world: Evidence and theories. *Advances in Psychological Science*, *29*(6), 1067. • Breiman, L. (2001). Random Forest. *Machine Learning*, 5–32. • van de Weijer, J., Ren, G., van de Weijer, J., Wei, W., & Wang, Y. (2020). Gender identification in Chinese names. *Lingua*, *234*. • Zhang, Q., Chen, H. C., Stuart Weekes, B., & Yang, Y. (2009). Independent effects of orthographic and phonological facilitation on spoken word production in mandarin. *Language and Speech*, *52*(1), 113–126.