What happens if an efficient information structure is disrupted? A preliminary study on personal names

Personal names serve to refer to specific individuals. While an intuitive way of achieving individual reference is to give each person a unique personal name, a system utilizing unique name words might require hundreds of millions of novel, pronounceable words, inevitably making many names long and confusable. Instead, language across the world make use of combinatoric name phrases (Ramscar et al., 2024; Van Langendonck, 2008; Wilkinson, 2015), which comprise a sequentially first name token (henceforth **prefix-name**) usually coming from a relatively small name-specific lexicon, and other tokens (henceforth **bynames**) typically taken from the rest of the lexicon (Table 1). This combinatoric structure is efficient, in that the entropy of prefix-names tends to be small, and this in turn reduces the information of the sets of bynames that are conditioned on them. Moreover, this structure allows very large sets of name phrases to be generated without increasing the overall size of lexicons (thus avoiding the lexical inflation that would result from unique names).

Critically however, naming behavior changed as countries started regulating how parents name their children to facilitate bureaucratic oversight (Ramscar et al., 2024). In most Western societies, this led to bynames becoming inherited, whereas in most East Asian societies, prefix-names became inherited, leading to changes in the distribution of different parts of name phrases. Fixing what had previously been flexible bynames in Western societies made it inevitable that more and more people would share identical names. To avoid this, as populations grew, new prefix-names were introduced, enlarging the prefix-name lexicon and lowering the relative frequency of existing prefix-names, causing Western names to become less efficient. In contrast, the distribution of Korean prefix-names remained largely unchanged (Kiet et al., 2008). Ramscar et al. (2024) show that prior to name regulation in the West, English prefix-names were distributed similarly to modern Korean prefix-names, and that the entropy of the former increased after regulations were introduced. **Accordingly, these changes predict names in Western societies will increasingly have come to be less efficient as compared to names in East Asian societies**. To estimate the relative impact of these different changes, we operationalized efficiency as a name's **surprisal**, computed from the internal name distribution of a person living at a given time, and assume that names with higher surprisal are less efficient, in that they require language users to process more information when names are used.

We based our estimate on two samples of Americans and Korean names, as there are prefix-name distribution data publicly available for these populations. For Americans, we sampled 15000 prefix-names of newborns in Indiana from a 50-year window, using the US baby name data published by the Social Security Administration[1]. For Koreans, due to limited data, we sampled 15000 prefix-names from the 1985, 2010, and 2015 census data[2]. Then to estimate the efficiency of actual Korean and US names (by definition, real names are impossible to anonymize), we calculated the average surprisal of names in publicly available datasets of popular sports stars. For Americans, we extracted the prefix-names of Superbowl players from 1966 and 2018 and calculated the surprisal based on the name distribution simulated at each corresponding year. For Koreans, we extracted the prefix-names of Asian Cup and the FIFA World Cup players between 1980 and 2022 and calculated the surprisal based on the name distribution simulated at the closest census year.

Figure 1a shows the prefix-name frequency distribution of 15000 sampled names from USA and South Korea census data. Korean prefix-names are far more concentrated than American prefix-names: the prefix-name 김(*Kim*) accounted for more than 15% of the names sampled, whereas the most frequent American name William had fewer than 5%. Because of such difference in distributions, American prefix-names are on average much more surprising than Korean prefix-names (Figure 1b), even though both name systems are equally capable of uniquely referring to specific individuals (100% of the American player names and 99.7% of the Korean player names were unique to a single individual). In addition, the average surprisal for American prefix-names has steadily increased over time (β=0.015, p=0.005), whereas that for Korean ones has remained unchanged (β=0.002, p=0.803).

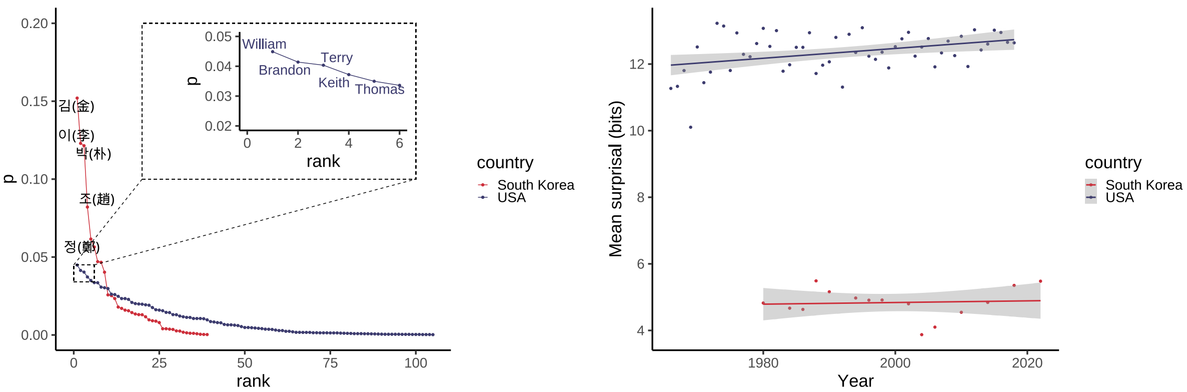
The simulation results are in line with our prediction: due to the different ways in which name regulations have impacted name structure, Western names have become increasingly less efficient than East Asian names; unlike East Asian names, Western names efficiency decreases as populations grow. While these information changes suggest that Western names have become harder to remember and use, their actual behaviorial consequences remain to be tested. We plan to conduct a behavioral experiment where we compare the accuracy in the recall and use of names in their respective populations.

[1] https://www.ssa.gov/oact/babynames/limits.html

[2] https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1IN15SD&conn_path=I2

Society	Full Name	Prefix-name	Byname
Western	David Smith	David	**Smith**
Ancient China	子喜(*Zi Xi*)	**子(*Zi*)**	喜(*Xi*)
Size of lexicons		small	large

Table 1: Historical name structures in the West and ancient China: in the West, the name consisted of a name-specific lemma (e.g. 'David', Van Langendonck, 2008) followed by a description of the person (e.g. 'Smith' as his occupation); in ancient China, the name consisted of a prefix (e.g. the gender prefix '*Zi*') followed by a word for everyday things (Wilkinson, 2015). These two name systems have a commonality: the first part comes from a relatively small size of lexicons, whereas the second part comes from a relatively large size of lexicons. The bolded part of each name was set to be inherited from parents to children by naming laws.



(a)

(b)

Figure 1: **(a)** The simulated distribution of 15000 sampled Korean (red) and 15000 sampled American (blue) prefix-names in 1985. The 5 most popular prefix-names in each population are labeled on the graph. **(b)** The efficiency of Korean (red) and American (blue) names, operationalized by the average surprisal of Korean World Cup and American Superbowl player prefix-names, as a function of year. Shaded area marks the 95% confidence interval. American name data was taken from the US baby name data published by Social Security Administration, and the Korean name data was taken from the 1985, 2000, and 2015 census. The American data is available every year between 1966 and 2018, whereas the Korean data is available every 2 years between 1980 and 2022.

References

Kiet, H. A. T., Baek, S. K., Jeong, H., and Kim, B. J. (2008). Korean family name distribution in the past. *Journal of the Korean Physical Society*, 51(5):1812–1816.

Ramscar, M., Chen, S., Futrell, R., and Mahowald, K. (2024). Cross-cultural structures of personal name systems reflect general communicative principles. *in review*.

Van Langendonck, W. (2008). *Theory and typology of proper names*. De Gruyter Mouton.

Wilkinson, E. (2015). *Chinese History: a new manual*. Harvard University Press.