Phonotactics of Gender in Mandarin Given Names: Patterns and Constraints

Fulang Chen and Michael Kenstowicz
Massachusetts Institute of Technology

1 Introduction

Linguists and psychologists have long been interested in sound symbolism. The phenomenon refers to the nonarbitrary association between sounds (of language or the natural environment) and their interpretation. It is actually part of the broader investigation of iconicity – 'the association between the physical properties of a stimulus and its interpretation or meaning' (Knoeferle et al. 2017). Sound symbolism is set against the observation that the relation between sound and meaning in human language is largely unpredictable and conventional as most famously expressed in Saussure's *l'arbitraire du signe*. Exceptions to this default state include special regions of the vocabulary in many languages such as onomatopoeia and ideophones, which often have distinctive phonotactics in addition to an intuitive imitative association with their referents. Researchers have demonstrated that adult subjects will tend to align nonsense syllables with pairs of antonyms in ways that cannot always be predicted from the phonological structure of their language but seem to accord with more universal regularities. Two of the most well-known cases are front vs. back vowels aligned with the small vs. large size dimension (Sapir 1929) and sonorant vs. obstruent consonants with rounded vs. angular shapes in the visual dimension (Köhler 1929, a.k.a. the *bouba-kiki* effect). For these more well-established sound-meaning associations researchers have sought to identify their basis. For example, front vs. back vowels and small vs. large size might reflect the relative volume of the oral cavity; another possibility is the acoustic correlates of these articulations in the second formant (Shinohara & Kawahara 2010). Ohala's (1984, 1994) Frequency-Code Hypothesis ties these phonetic factors together suggesting that the effect is based on a speaker's intuition that higher acoustic frequency naturally aligns with smaller volume or magnitude, while lower acoustic frequency naturally aligns with larger volume or magnitude.

Ohala's hypothesis has been applied to the associations between the social-biological category of gender and speech based on the proposition that (adult) females are on average smaller in stature, vocal tract length and anatomy, while males are larger in these dimensions. For example, research on English given names has discovered phonotactic patterns that correlate with gender. Specifically, female names are more likely to have a higher ratio of open syllables (Slater & Feinman 1985) and are more likely to contain high front vowels while male names favor back vowels (Cutler et al. 1990; Wright et al. 2005); for consonants female names are more likely to contain sonorants, especially /l, m, n/ that correlate with rounded visual shapes while male names favor obstruents, especially /p, t, k/ that correlate with sharp-edged figures (Wright et al. 2005; Sidhu & Pexman 2015). Recent studies suggest that some of these patterns are cross-linguistic (e.g. Sullivan 2018; Starr et al. 2018; Wong & Kang 2019; a.o.) and conform to the Frequency-Code Hypothesis. It is observed that higher-frequency sounds such as high, front, unrounded vowels, which have higher F2 (or F2-F1 difference), are more common in female names, possibly because such sounds imply smallness; low-frequency sounds such as low, back, rounded vowels, which have lower F2 (or F2-F1 difference), and grave (i.e. labial and velar) consonants, which also have lower F2, are more common in male names, possibly because such sounds imply largeness.

In this paper, we investigate the phonotactic patterns that correlate with gender in given names for Mandarin Chinese, a language that is phonotactically quite different from English. We also compare the phonotactic grammars of Mandarin female and male given names obtained from Maximum-Entropy phonotactic learning models.

* We thank the audiences at AMP and the MIT Phonology Circle for helpful comments on this paper.

Proceedings of *AMP 2021*

2 Background

2.1 *Mandarin syllables* As listed in Tables 1 and 2 below, Mandarin syllables are composed of 21 possible onset consonants (including the palatals /tɕʰ, tɕ, ɕ/, which are both [+coronal] and [+dorsal]; it is debated whether /tɕʰ, tɕ, ɕ/ are independent phonemes, or are allophones of the velars /kʰ, k, x/, the dentals /tsʰ, ts, s/, or the retroflex /tʂʰ, tʂ, ʂ/ (see Duanmu 2007)). There are no complex consonantal onsets. There are five distinct [+syllabic] phonemes (underlined in Table 2) and five allophones that form vocalic nuclei; we treat the apical vowels [ɿ] and [ʅ], which occur after dental and retroflex sibilants, respectively, as syllabic approximants [ɹ] and [ɻ] (retroflex). Finally, Mandarin has three distinct on-/off-glides /j, ɥ, w/ and three distinct coda consonants /n/, /ŋ/ and suffixal /ɻ/.

	[+labial]	[+coronal]	[+dorsal]
Obstruent	pʰ, p, f	tʰ, t, z, tʂʰ, tʂ, ʂ, tsʰ, ts, s, tɕʰ, tɕ, ɕ	tɕʰ, tɕ, ɕ, kʰ, k, x
Sonorant	m	n, l, ɻ	ŋ

Table 1. Distinct consonants in Mandarin syllables

	[+front, -back]		[-front, -back]		[-front, +back]	
	[-round]	[+round]	[-round]	[+round]	[-round]	[+round]
[+high, -low]	i	y				u
[-high, -low]	e, ɛ		ə		ɤ	o
[-high, +low]			a		ɑ	

Table 2. Distinct vowels in Mandarin syllables

Mandarin is a tonal language with four distinctive tones: high(-level) T1, (mid-)rising T2, low(-falling) T3, and (high-)falling T4. In Table 3, the digits alongside the syllable /tɕjaw/ represent the pitch values of each tone; in the traditional Chao (1930) notation, 5 indicates the top of the pitch space and 1 the bottom.

T1	H	tɕjaw55	‘teach’
T2	LH	tɕjaw35	‘chew’
T3	L(LH)	tɕjaw21(4)	‘mix’
T4	HL	tɕjaw53	‘call’

Table 3. Tones in Mandarin

2.2 *Mandarin given names* Mandarin given names are predominantly monosyllabic/monomorphemic (and represented by one written character) or disyllabic/bimorphemic (represented by two characters) and constitute an open class, where any lexical item is in principle available to serve as a given name. This contrasts with the situation in English where names are largely selected from a closed class of dedicated lexical items.[1] As a result the choice of any Mandarin given name implicates the semantic features associated with the morpheme and its written character. While English names may have semantic characteristics based on their etymology (e.g. Leonard = ‘lion-heart’), these features presumably play a much weaker role compared to Chinese. Also, due to extensive (segmental) homophony – for example, the characters 萍 ‘duckweed’ and 平 ‘even’, which are more frequently used in female and male names, respectively,[2] have the same pronunciation /pʰiŋ35/ – one might expect the choice of a Mandarin given name to be based primarily on the semantic associations of the individual homophone. Hence, one might wonder whether the phonotactic patterns that correlate with gender in English and other languages would play any role at all in the choice of Mandarin given names. Our study addresses this question (cf. Starr et al. 2018; Wong & Kang 2019). In addition, since Mandarin given names are chosen from the lexicon as a whole, one might wonder to what extent their phonotactic properties align with phonotactic properties of the overall lexicon. We consider this question in section 3.4.

[1] The class, while closed in the grammatical sense, can and does expand through linguistic borrowing.

[2] In our corpus, the character 萍 has 216 occurrences in female names and 11 occurrences in male names; the character 平 has 94 occurrences in female names and 186 occurrences in male names.

3 Corpus study

3.1 *Method* Our study is based on data from an online corpus[3] composed of the Mandarin given names (monosyllabic/monomorphemic or disyllabic/bimorphemic) of 25,856 persons, categorized and balanced for self-identified gender (12,928 of each gender). Compared with some other recent studies on Chinese given names, the corpus we investigate is twice as large as the Mandarin corpus (N = 11,206) used by Starr et al. (2018) and much larger than the Cantonese corpus (N = 288; 144 of each gender) used by Wong & Kang (2019). In order to be able to compare our findings with these earlier studies, we also selected the top 5,000, 500, and 150 most representative disyllabic names for each gender in order to approximate the size of their data sets. Specifically, we assume that more representative names for each gender are composed of characters that are more frequently used in the corresponding gender's names. Hence, we established a ranking for each gender's names in the corpus, based on the combined number of occurrences of the component characters in the corresponding gender's names. For example, the name 小平 (/ɕjaw21 pʰiŋ35/) occurs as both a female and a male name; the component characters 小 'small' and 平 'even' have 247 and 94 occurrences, respectively, in the female names set, and 135 and 186 occurrences, respectively, in the male names; thus, we assigned the female name 小平 the number 247 + 94 = 341, and the male name 小平 135 + 186 = 321. Based on the number assigned to each name, we ranked each gender's names; then, based on this ranking, we obtained the top 5,000, 500, and 150 distinct disyllabic names for each gender.

We coded each Mandarin given name in the entire corpus and in our samples for six predictors that are based on the phonotactic patterns that previous research has found to correlate with gender in English given names (cf. Sullivan 2018; Starr et al. 2018; Wong & Kang 2019): *open-syllable proportion* (the number of open syllables divided by the total number of syllables in the name), *high-nucleus proportion* (the number of [+high] nuclear vowels divided by the total number of nuclear vowels in the name), *back-nucleus proportion* (the number of [+back] nuclear vowels divided by the total number of nuclear vowels in the name), *round-nucleus proportion* (the number of [+round] nuclear vowels divided by the total number of nuclear vowels in the name), *obstruent-onset proportion* (the number of obstruent onsets divided by the total number of onset consonants in the name), and *non-coronal-onset proportion* (the number of [-coronal] onsets divided by the total number of onset consonants in the name). The denominator in these calculations varied between 1 and 2 depending on how many syllables/onsets are contained in the given name.

We assessed the significance of these phonological predictors with univariate and multivariate logistic regression models (using the *glm* function in *R*); the gender (female = 0; male = 1) associated with each name was the dependent variable (with 'female' as the reference level), and one (univariate) or all (multivariate) of the predictors were the independent variables (cf. Sullivan 2018; Wong & Kang 2019).

3.2 *Results* Tables 4 and 5 list the top 10 most frequently used written characters in Mandarin female and male names, respectively, in our corpus. It is easy to see that the semantic content of the characters plays a role in the choice of these given names. The characters 丽 'beautiful', 芳 'fragrant', 玲 'tinkling', 兰 'orchid', and 梅 'plum' are predominantly used in female names, while 国 'nation', 军 'military', and 伟 'great' are predominantly used in male names. The characters 华 'China, splendid' and 文 'literary' are frequently used in both genders' names. For the top 10 characters in each gender's names, the six phonological predictors mentioned above in 3.1 do not play any decisive role in the choice of Mandarin given names. In terms of syllable structure, exactly half of the top 10 characters in both genders' names are pronounced with an open syllable. In terms of vowel segments, four of the top 10 characters in female names vs. two of the top 10 characters in male names are pronounced with a high, front nuclear vowel; none of the top 10 characters in female names vs. two of the top 10 characters in male names are pronounced with a non-high, back nuclear vowel. In terms of consonant segments, four of the top 10 characters in female names vs. seven of the top 10 characters in male names are pronounced with an obstruent onset; and four of the top 10 characters in female names vs. only one of the top 10 characters in male names is pronounced with a sonorant onset.

[3] https://github.com/NLPBLCU/Chinese-Celebrities-Names

Character			Number of occurrences	
			in female given names	in male given names
丽	li:53	‘beautiful’	401	11
英	iŋ55	‘flower’ (archaic), ‘outstanding’	361	53
红	xuŋ35	‘red’	317	96
晓	ɕjaw21	‘dawn’	301	150
华	xwa35	‘China’, ‘splendid’	277	234
芳	faŋ55	‘fragrant’	273	35
玲	liŋ35	‘tinkling’	270	5
玉	y53	‘jade’	265	123
兰	lan35	‘orchid’	264	7
梅	mej35	‘plum’	253	7

Table 4. Top 10 most frequently used characters in Mandarin female names

Character			Number of occurrences	
			in female given names	in male given names
明	miŋ35	‘bright’	128	299
文	wən35	‘literary’	233	288
国	kwo35	‘nation’	59	241
华	xwa35	‘China’, ‘splendid’	277	234
建	tɕjɛn53	‘build’	76	206
志	tʂʅ53	‘aspiration’	82	193
军	tɕyn55	‘military’	30	191
伟	wej21	‘great’	56	189
德	tɤ35	‘virtue’	71	187
平	pʰiŋ35	‘even’	94	186

Table 5. Top 10 most frequently used characters in Mandarin male names

Table 6 summarizes the number of distinct characters that represent the names in the entire corpus and the samples as well as the proportion of names in the corpus and the samples that contain at least one of the top 10 most frequently used characters. While most names in the corpus as a whole and in the top 5,000 sample do not contain any of the top 10 characters, almost every name in the top 500 and every name in the top 150 samples contains at least one of the top 10 characters. Thus, the phonotactic properties of the top 10 characters in each gender’s names are likely to be over-represented in the smaller-sized samples.

	Corpus (12,928)	Top 5,000	Top 500	Top 150
Number of distinct characters in female given names	1808	989	172	44
Number of distinct characters in male given names	1801	887	188	59
% female given names containing the top 10 most frequently used characters	21.48%	34.28%	94.00%	100%
% male given names containing the top 10 most frequently used characters	16.58%	33.96%	95.80%	100%

Table 6. Distinct characters and top 10 most frequently used characters in corpus and samples

In both the entire corpus and the samples we find that all six predictors for gender trend in the same direction as reported in the corpus studies of English names referenced above. The descriptive statistics summarized in Table 7 show that for Mandarin, female names have a higher proportion of open syllables and high vowel nuclei, while male names have a higher proportion of back vowel nuclei, round vowel nuclei, obstruent onsets, and non-coronal onsets.

		Corpus (12,928)		Top 5,000		Top 500		Top 150	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Open syllable %	F	49.35%	36.04%	49.69%	33.36%	53.30%	33.71%	55.00%	35.59%
	M	44.98%	35.05%	41.17%	32.81%	37.00%	30.54%	32.67%	31.19%
High nucleus %	F	41.73%	36.30%	40.54%	33.75%	52.40%	34.16%	54.33%	32.22%
	M	37.18%	35.00%	38.34%	33.28%	38.10%	31.71%	42.00%	32.81%
Back nucleus %	F	13.69%	25.94%	11.72%	22.38%	8.30%	19.15%	7.67%	18.98%
	M	20.43%	29.53%	21.72%	28.45%	18.30%	27.23%	18.33%	25.53%
Round nucleus %	F	16.91%	27.76%	16.20%	25.61%	13.80%	24.31%	14.33%	25.47%
	M	21.09%	29.71%	22.86%	29.10%	19.50%	28.40%	20.67%	27.89%
Obstruent onset %	F	74.68%	37.91%	74.88%	36.02%	54.80%	40.00%	49.00%	40.74%
	M	84.51%	30.57%	85.46%	28.19%	72.60%	34.81%	68.67%	34.54%
Non-coronal onset %	F	27.48%	37.69%	30.47%	37.43%	32.00%	40.24%	35.33%	41.59%
	M	30.16%	37.36%	31.66%	36.62%	48.70%	39.07%	55.00%	38.32%

Table 7. Means and standard deviations

Figures 1 and 2 below help visualize the effects of varying sample sizes. We observe that the predictors are largely stable across sample sizes; exceptions are that for female names the high-nucleus proportion as a predictor for gender is more pronounced in the positive direction in the smaller-sized samples, while the obstruent-onset proportion as a predictor for gender is more pronounced in the negative direction in the smaller-sized samples; thus the smaller-sized samples tend to exaggerate these gender markers.

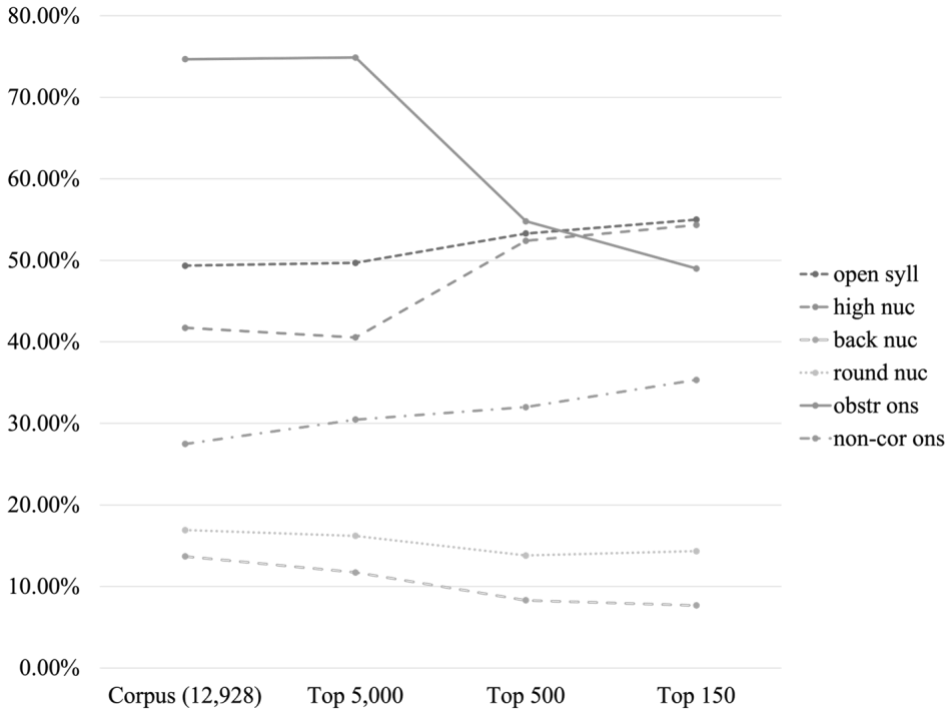

Figure 1. Predictors for gender in Mandarin female names across sample sizes

For male names the open-syllable proportion and the non-coronal-onset proportion as predictors for gender are more pronounced in the smaller-sized samples, while the obstruent-onset proportion as a predictor for gender is less pronounced in the smaller-sized samples.

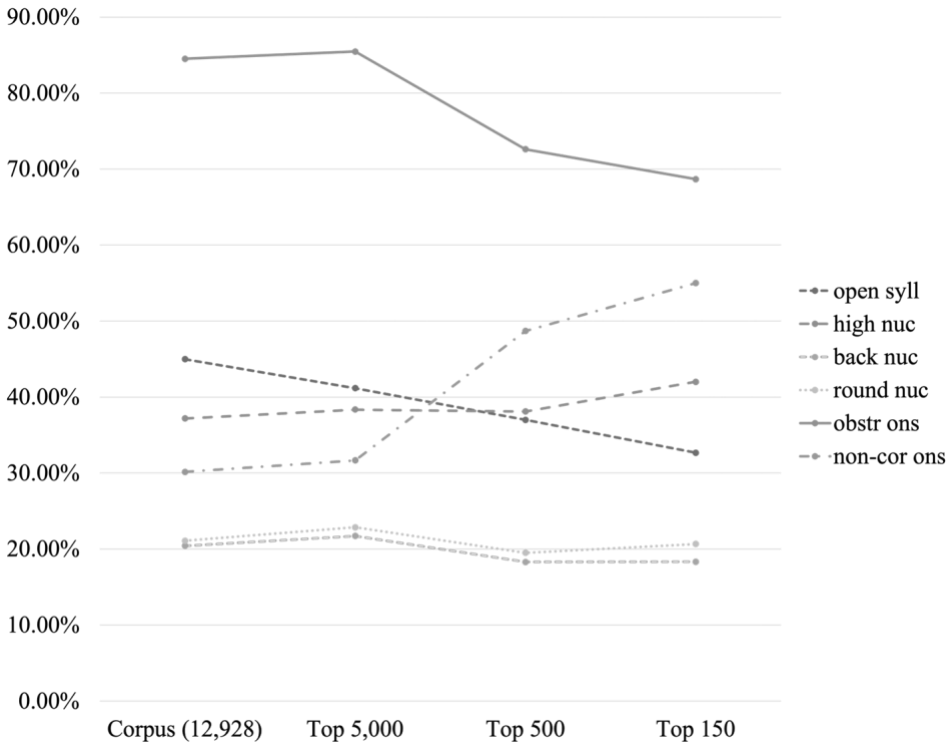

Figure 2. Predictors for gender in Mandarin male names across sample sizes

The z-scores for the univariate and multivariate logistic regression models of these predictors are reported in Table 8; here positive signs indicate biases toward male names. We observe that all six predictors are significant under both univariate and multivariate analyses on names in the corpus and the samples, with just a few exceptions: the back-nucleus proportion, the round-nucleus proportion, and the obstruent-onset proportion are not significant predictors for gender under the multivariate analysis on names in the top 150 sample; the round-nucleus proportion is not a significant predictor for gender under the multivariate analysis on names in the top 500 and top 150 samples; and the non-coronal-onset proportion is not a significant predictor for gender under both univariate and multivariate analyses in the top 5,000 sample.

	Corpus (12,928)		Top 5,000		Top 500		Top 150	
	Univ.	Mult.	Univ.	Mult.	Univ.	Mult.	Univ.	Mult.
Open syllable %	-9.837 ***	-12.940 ***	-12.69 ***	-14.481 ***	-7.584 ***	-8.747 ***	-5.292 ***	-5.876 ***
High nucleus %	-10.231 ***	-13.076 ***	-3.28 **	-7.714 ***	-6.595 ***	-6.323 ***	-3.192 **	-3.955 ***
Back nucleus %	19.17 ***	12.496 ***	18.80 ***	13.013 ***	6.387 ***	4.070 ***	3.888 ***	1.323
Round nucleus %	11.604 ***	5.665 ***	11.976 ***	3.406 ***	3.370 ***	0.443	2.030 *	0.677
Obstruent onset %	22.45 ***	18.245 ***	15.88 ***	12.184 ***	7.184 ***	2.241 *	4.293 ***	1.172
Non-coronal onset %	5.737 ***	2.930 **	1.607	1.217	6.444 ***	4.181 ***	4.086 ***	2.706 **

Table 8. Z-scores for univariate and multivariate logistic regression models
(Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1)

3.3 *Other predictors* In addition to the six aforementioned predictors, we indicate here how some other gender predictors that have been investigated in recent studies apply in our data. Sullivan (2018)

found that among sonorants, /l, m, n/ that correlate with rounded visual shapes (see Sidhu & Pexman 2015) are more common in both English and French female names. Similarly, we find that for Mandarin more female names than male names contain at least one of the /l, m, n/ onsets (see Table 9).

		Corpus (12,928)	Top 5,000	Top 500	Top 150
% given names containing /l, m, n/ onsets	F	28.79%	33.96%	58.60%	63.33%
	M	19.95%	21.98%	40.20%	48.67%

Table 9. /l, m, n/ onsets in Mandarin given names

Starr et al. (2018) reported that /p, t, k/, which have shorter voice-onset times than their aspirated counterparts, are more common in male names for both Cantonese and Mandarin; /p, t, k/ are also known to correlate with sharp figures (see Sidhu & Pexman 2015) and are more common in male names for both English and French (Sullivan 2018). We replicate these findings for Mandarin, where more male names than female names contain at least one of the /p, t, k/ onsets (see Table 10).

		Corpus (12,928)	Top 5,000	Top 500	Top 150
% given names containing /p, t, k/ onsets	F	10.17%	8.96%	3.20%	0.00%
	M	20.77%	24.22%	26.00%	24.67%
% given names containing /p^h, t^h, k^h/ onsets	F	8.92%	8.00%	5.00%	2.67%
	M	11.99%	10.62%	6.00%	4.00%

Table 10. /p, t, k/ and /p^h, t^h, k^h/ onsets in Mandarin given names

Starr et al. (2018) also reported that for both Cantonese and Mandarin, the palatals /tɕʰ, tɕ, ɕ/, which have higher F2, are more common in female names, whereas the retroflex /tʂʰ, tʂ, ʂ/, which have lower F2, are more common in male names. We replicate their findings for Mandarin, where more female names than male names contain at least one of the /tɕʰ, tɕ, ɕ/ onsets, while more male names than female names contain at least one of the /tʂʰ, tʂ, ʂ/ onsets (see Table 11). These findings support the Frequency-Code Hypothesis.

		Corpus (12,928)	Top 5,000	Top 500	Top 150
% given names containing /tɕʰ, tɕ, ɕ/ onsets	F	40.40%	44.96%	33.80%	30.67%
	M	37.93%	40.72%	32.80%	29.33%
% given names containing /tsʰ, ts, s/ onsets	F	6.88%	6.96%	2.00%	0.67%
	M	7.99%	6.48%	2.60%	2.00%
% given names containing /tʂʰ, tʂ, ʂ/ onsets	F	17.67%	16.56%	7.00%	31.33%
	M	28.89%	30.18%	21.40%	47.33%

Table 11. /tɕʰ, tɕ, ɕ/, /tsʰ, ts, s/ and /tʂʰ, tʂ, ʂ/ onsets in Mandarin given names

It is worth observing that in Tables 9 through 11 the percentages in the top 500 and Top 150 columns exaggerate the effect of the phonological predictors compared to the larger Top 5,000 and entire corpus columns. This is likely due to the fact that almost every name in the top 500 and every name in the top 150 samples contains at least one of the top 10 characters (see Table 6).

Starr et al. (2018) also found that /ŋ/, which has lower F2 than the alveolar nasal /n/, is more common in male names for both Cantonese and Mandarin. We replicate this finding for Mandarin, where more male names than female names contain the /ŋ/ coda (see Table 12). This finding also accords with the Frequency-Code Hypothesis.

		Corpus (12,928)	Top 5,000	Top 500	Top 150
% given names containing /ŋ/ coda	F	39.87%	42.54%	48.80%	50.00%
	M	50.02%	58.70%	61.80%	64.67%
% given names containing /n/ coda	F	44.40%	47.34%	36.80%	31.33%
	M	43.59%	44.34%	47.40%	47.33%

Table 12. /ŋ/ and /n/ codas in Mandarin given names

According to the Frequency-Code Hypothesis, high tones imply smallness, and low tones largeness; equating smallness with femaleness and largeness with maleness, it is predicted that high tones should be more common in female names and low tones more common in male names. Ohala (1994: 327) further suggests that rising F0 is associated with 'deference, politeness, submission, lack of confidence', and falling F0 with 'assertiveness, authority, aggression, confidence, threat'; thus, rising tones should be more common in female names and falling tones more common in male names. But Wong & Kang (2019) found that for Cantonese, female names are more likely to begin with a low or rising tone, while male names are more likely to begin with a high or falling tone; hence at least a portion of their tonal findings do not conform to the Frequency-Code Hypothesis. We find that for Mandarin, the high(-level) T1 is more common in male names in the corpus as a whole and in the top 5,000 sample, while it is more common in female names in the smaller samples; the (mid-)rising T2 is more common in male names, while the low(-falling) T3 and (high-)falling T4 are more common in female names (see Table 13).[4] Thus, both Wong & Kang's (2019) and our findings suggest that tones do not play a major role in the choice of gendered names; presumably, the choice of given names is already maximally restricted when the semantic content of the written characters and the segmental features of the syllables are taken into consideration. If these factors have priority over tone in choosing a name then there will be little opportunity left for tonal factors to play a role.[5]

		Corpus (12,928)	Top 5,000	Top 500	Top 150
% given names containing T1 syllable	F	43.90%	45.46%	38.60%	34.67%
	M	46.80%	47.64%	33.00%	28.67%
% given names containing T2 syllable	F	54.74%	57.38%	51.40%	54.00%
	M	54.69%	62.14%	89.80%	96.00%
% given names containing T3 syllable	F	25.26%	28.10%	22.00%	22.67%
	M	26.45%	27.44%	19.80%	18.00%
% given names containing T4 syllable	F	46.23%	49.52%	63.00%	58.67%
	M	44.47%	41.76%	31.20%	25.33%

Table 13. T1, T2, T3 and T4 syllables in Mandarin given names

It is worth mentioning that tone height and tone contour in Mandarin are found to correlate with size in the direction predicted by the Frequency-Code Hypothesis once the lexical restrictions are removed. Lapolla (1994) conducted an experiment where five Mandarin speakers were asked to classify as 'small' vs. 'big' a set of CV nonsense syllables that varied their onset consonants in terms of the features [±nasal] and [±grave] and their nuclear vowels as [±grave] as well as the four Mandarin tones. He reports (p. 139) that for grave-initial consonants 52 responses were labeled 'small' vs. 71 'big'; for the grave vowels 51 were labeled 'small' vs. 80 as 'big' showing the expected trends for the F2-based segmental features. The classification as 'small' vs. 'big' as a function of tone is indicated below:

	T1	T2	T3	T4
'small'	35.2	16.2	16.2	20.9
'big'	13.3	16.2	15.2	50.5

Table 14. Tone-size correlation in Mandarin nonsense syllables

The results in Table 14 indicate that the high(-level) T1 was biased towards the 'small' response while the

[4] In the smaller-sized samples, a much greater proportion of male names contain at least one T2 syllable – this is because 78.80% and 92.67% of male names in the top 500 and top 150 samples, respectively, contain at least one of the characters 明 /miŋ35/, 文 /wən35/, 国 /kwo35/, 华 /xwa35/; a much greater proportion of female names contain at least one T4 syllable – this is because 36.40% and 41.33% of female names in the top 500 and top 150 samples respectively contain the character 丽 /li:53/.

[5] A similar asymmetry between segmental and tonal factors is found in Mandarin loanword adaptation: when there is a choice between changing a segmental feature or a tonal property in order to bring the loan into agreement with the Mandarin lexicon, it is faithfulness to the segmental factor that typically wins out (Wu 2006; Glewwe 2021; a.o.).

(high-)falling T4 displayed the opposite preference. These findings accord with the Frequency-Code Hypothesis (see Shang & Styles (2017) for the behavior of Mandarin tones with respect to the *bouba-kiki* effect).

Lastly, Starr et al. (2018) reported that reduplication is more common in female names for both Cantonese and Mandarin. We replicate their finding for Mandarin, where more female names than male names involve segmental reduplication (see Table 15).

		Corpus (12,928)	Top 5,000	Top 500	Top 150
Number of disyllabic given names with segmental reduplication	F	416	71	17	12
	M	93	27	3	1

Table 15. Segmental reduplication in Mandarin disyllabic given names

3.4 *Given names vs. Overall lexicon* Our study has revealed that female and male names in Mandarin display distinct phonotactic tendencies. Since Mandarin given names are chosen from the lexicon as a whole, one might wonder to what extent the phonotactic tendencies displayed by Mandarin given names align with phonotactic tendencies of the broader lexicon. To address this question, we compare the phonotactic tendencies displayed by Mandarin given names with the tendencies displayed by the 3,913 monosyllabic and 34,233 disyllabic words listed in the Chinese Lexical Database (CLD) (Sun, Hendrix & Baayen 2018).

We observe that for segmental features (both vocalic and consonantal), female names are more polarized compared to male names, which more closely approximate the lexicon as a whole. Table 16 shows this pattern for the gender predictors for the vocalic nucleus and the consonantal onset discussed in section 3.2. For the high-nucleus proportion the mean of 41.73% for female names is farther away from the mean of 30.13% for the overall CLD than the mean of 37.18% for male names. For the back-nucleus proportion, the round-nucleus proportion and the obstruent-onset proportion, the mean for female names is about 5% or 10% lower than the means for male names and the overall CLD.

	Female names (12,928)		Male names (12,929)		CLD	
	Mean	SD	Mean	SD	Mean	SD
Open syllable %	49.35%	36.04%	44.98%	35.05%	59.59%	37.08%
High nucleus %	41.73%	36.30%	37.18%	35.00%	30.13%	34.50%
Back nucleus %	13.69%	25.94%	20.43%	29.53%	23.15%	31.45%
Round nucleus %	16.91%	27.76%	21.09%	29.71%	21.75%	30.92%
Obstruent onset %	74.68%	37.91%	84.51%	30.57%	84.75%	30.38%
Non-coronal onset %	27.48%	37.69%	30.16%	37.36%	29.42%	36.84%

Table 16. Phonotactic tendencies in the Mandarin given names and overall lexicon

As seen in the first section of Table 17, the same pattern holds for the additional gender predictors for the onset consonant investigated in section 3.3. Exceptions to this general pattern occur with the dorsal nasal coda which is far more common in male names than in female names and the overall CLD.

	Female names (12,928)	Male names (12,929)	CLD
Consonantal onset			
% words/given names containing /l, m, n/ onsets	28.79%	19.95%	19.74%
% words/given names containing /p, t, k/ onsets	10.17%	20.77%	26.65%
% words/given names containing /p^h, t^h, k^h/ onsets	8.92%	11.99%	16.00%
% words/given names containing /$tɕ^h$, tɕ, ɕ/ onsets	40.40%	37.93%	31.07%
% words/given names containing /ts^h, ts, s/ onsets	6.88%	7.99%	12.03%
% words/given names containing /$tʂ^h$, tʂ, ʂ/ onsets	17.67%	28.89%	29.63%
Consonantal coda			
% words/given names containing /ŋ/ coda	39.87%	50.02%	36.39%
% words/given names containing /n/ coda	44.40%	43.59%	31.91%

Tones			
% words/given names containing T1 syllable	43.90%	46.80%	39.23%
% words/given names containing T2 syllable	54.74%	54.69%	39.02%
% words/given names containing T3 syllable	25.26%	26.45%	30.92%
% words/given names containing T4 syllable	46.23%	44.47%	53.36%

Table 17. Other phonotactic tendencies in the Mandarin given names and overall lexicon

In contrast with segmental features, the open-syllable proportion (in Table 16) and tones (in Table 17) do not display the same pattern, possibly because they are not readily analyzed as binary contrasts the way Jakobsonian segmental features are. A binary feature like [±back] references two poles of a single dimension and for this reason may have the same cognitive status as the female vs. male poles of gender. Prosodic contrasts like open vs. closed syllables involve subset-supraset relations and hence seem to be organized on a hierarchical basis. And Chinese tones, at least from an autosegmental perspective, are analyzed in terms categories like shape and register in addition to a simple [±high tone] distinction.

4 Phonotactics learner

To probe the interaction of the predictors for gender more closely, we ran phonotactic learning models (Hayes & Wilson 2008), which discover phonotactic constraints and their weights from a set of phonological forms, based on the principle of Maximum Entropy, on each gender's names in our corpus. We set the maximum constraint size to 1 (unigram), and the maximum O/E for constraints to 0.3. The grammars obtained from the models are presented in Table 18.

Female (12,928)			Male (12,928)		
Constraint		Weight	Constraint		Weight
*[+son,-ant,-syl]	*[ɻ]	3.798	*[+son,-ant,-syl]	*[ɻ]	4.447
*[+cor,+syl]	*[ɳ, ɽ]	2.05	*[-cons,+cor]	*[ɳ]	2.125
*[-hi,+back]	*[ɑ̲, ɤ̲, o̲]	2.047	*[+asp,+ant,-dor]	*[s^h, t^h]	1.78
*[-son,-cont,-cor]	*[p^h, p̲, k^h, k̲]	2.038			
*[+asp,-lab,-dor]	*[s^h, ʂ̲, t^h]	1.989			

Table 18. Phonotactic grammars for Mandarin female and male given names

We also ran the same model with the same settings on the 3,913 monosyllabic and 34,233 disyllabic words listed in the CLD and found only one constraint, *[+son,-ant,-syl] (*[ɻ]), with a weight of 3.422.

Taking these results at face value, the grammar for female names contains more marked segments (underlined in Table 18) than the grammar for male names, which we might interpret as another indication that female names are more marked segmentally compared to male names. Specifically, the marked segments for female names consist of more low acoustic frequency sounds – non-high, back vowels /ɑ, ɤ, o/, which have lower F2 than high, front vowels; and more obstruents, specifically the grave oral stops /p^h, p, k^h, k/, which have lower F2 than the coronal /t^h, t/, and the retroflex sibilant /(t)ʂ^h/, which has lower F2 than the palatal /(t)ɕ^h/, and the apical vowel /ɿ/, which occurs after retroflex sibilants – that imply largeness according to the Frequency-Code Hypothesis. However, higher acoustic frequency sounds that imply smallness according to the Frequency-Code Hypothesis are not penalized in the Max-Ent grammar for male names.

This asymmetry in the gender-feature association, we suggest, may arise in the gender-size association and/or the size-feature association. Figure 3 shows the association of these three binary dimensions which we have been assuming.

Gender	Female	Male
	|	|
Size	Small	Large
	|	|
Features	[+high]	[-high]
	[-back]	[+back]
	etc.	etc.

Figure 3. Associations between gender, size and phonological features

One possibility is that the association of largeness with male is stronger than the association of small with female; thus, the association of low acoustic frequency sounds (that imply largeness according to the Frequency-Code Hypothesis) with male is stronger than the association of high acoustic frequency sounds (that imply smallness according to the Frequency-Code Hypothesis) with female. In our corpus, we find that the character 小 'small' is often used in both genders' names, while 大 'large', 宏 'grand', 洪 'vast', 浩 'vast' are only frequently used in male names (see Table 19); this supports possibility 1.

Character			Number of occurrences	
			in female given names	in male given names
小	ɕjaw21	'small'	247	135
大	ta53	'large'	28	85
宏	xuŋ35	'grand'	37	112
洪	xuŋ35	'vast'	35	104
浩	xaw53	'vast'	4	83

Table 19. Characters with smallness/largeness semantics in Mandarin given names

The other possibility is that the association of low acoustic frequency sounds with largeness (and hence male) is stronger than the association of high acoustic frequency sounds with smallness (and hence female). More cross-linguistic study is needed to determine whether the associations seen in Figure 3 are weighted and if so whether the asymmetry seen in the gender-feature association arises in the gender-size association or the size-feature association.

5 Summary

Our study adds value to the literature on the phonological correlates of gender in given names. We have shown that many of the phonological predictors for gender in English trend in the same direction for Mandarin. This finding is significant because the two languages differ in several important respects. First, Mandarin has a different segmental inventory with contrasts in features such as aspiration, palatality, and retroflexion whose acoustic correlates nevertheless support Ohala's Frequency-Code Hypothesis. Second, Mandarin given names are drawn from a much broader lexical space compared to English; consequently the name implicates and is typically chosen for the semantic features associated with the particular lexical item and the written character it is linked to. The fact that phonotactic properties of the given name still correlate with gender in the expected direction is thus rather remarkable. Several other results of this study are worthy of mention here. It appears that the phonological connection with gender in Mandarin given names is strongest for segmental features (both vocalic and consonantal). Tonal features do not align with gender in the way predicted by the Frequency-Code Hypothesis probably because segmental features take priority leaving little opportunity for tone to play a decisive role. Compared to male names, female names are farther away from the Mandarin lexicon as a whole for many of the segmental gender predictors investigated here. Furthermore, the Maximum-Entropy phonotactics learner discovered that certain low acoustic frequency sounds that imply largeness according to the Frequency-Code Hypothesis are penalized for female names, while higher acoustic frequency sounds that imply smallness according to the Frequency-Code Hypothesis are not marked in the grammar for male names. We suggest that this asymmetry seen in the gender-feature association may arise in the gender-size association and/or the size-feature association.

References

Chao, Yuen-Ren. (1930). A system of tone letters. *Le Maitre Phonetique, 30*: 24–27.

Cutler, Anne, Mcqueen, James, & Robinson, Ken. (1990). Elizabeth and John, sound patterns of men's and women's names. *Journal of Linguistics, 26*(2): 471–482.

Duanmu, San. (2007). The phonology of standard Chinese. OUP Oxford.

Glewwe, Eleaor. (2021). The phonological determinant of tone in English loanwords in Mandarin. *Phonology, 38*(2): 203–39.

Hayes, Bruce, & Wilson, Colin. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry, 39*(3): 379–440.

Knoeferle, Klemens, Li, Jixing, Maggioni, Emanuela, & Spence, Charles. (2017). What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings. *Scientific Reports, 7*(1): 1–11.

Köhler, Wolfgang. (1929). Gestalt Psychology: an introduction to new concepts in modern psychology. New York: Liveright.

Lapolla, Randy. (1994). An experimental investigation into phonetic symbolism as relates to Mandarin Chinese. In Hinton, Leanne, Nichols, Johanna, and Ohala, John J. (eds), *Sound Symbolism*, 130–147. Cambridge & New York: Cambridge University Press.

Ohala, John Jerome. (1984). An ethological perspective on common cross-language utilization of F0 of voice. *Phonetica, 41*: 1–16.

Ohala, John Jerome. (1994). The frequency code underlies the sound-symbolic use of voice pitch. In Hinton, Leanne, Nichols, Johanna, and Ohala, John Jerome. (eds), *Sound Symbolism*, 325–347. Cambridge & New York: Cambridge University Press.

Sapir, Edward. (1929). A study of phonetic symbolism. *Journal of Experimental Psychology, 12*: 225-239.

Shang, Nan & Styles, Suzy J. (2017). Is a high tone pointy? Speakers of different languages match Mandarin Chinese tones to visual shapes differently. *Frontiers in Psychology, 8*: 2139.

Shinohara, Kazuko. & Kawahara, Shigeto. (2010). A cross-linguistic study of sound symbolism: the images of size. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*: 396–410.

Sidhu, David M., & Pexman, Penny M. (2015). What's in a name? Sound symbolism and gender in first names. *PLoS ONE, 10*(5): 1–22.

Slater, Anne Saxon, & Feinman, Saul. (1985). Gender and the phonology of North American first names. *Sex Roles, 13*: 429–440.

Starr, Rebecca Lurie, Yu, Alan C.L., & Shih, Stephanie S. (2018). Sound symbolic effects in Mandarin and Cantonese personal names and Pokémon names. Paper presented at the 1st conference on Pokémonastics, Keio University, Tokyo.

Sullivan, Lisa. (2018). Phonology of gender in English and French given names. University of Toronto MA thesis.

Sun, Ching Chu, Hendrix, Peter, Ma, Jianqiang. & Baayen, Rolf Harald. (2018). Chinese Lexical Database (CLD): A large-scale lexical database for simplified Mandarin Chinese. *Behavior Research Methods*, https://doi.org/10.3758/s13428-018-1038-3.

Wong, Kristen Wing Yan & Kang, Yoonjung. (2019). Sound symbolism of gender in Cantonese first names. *Proceedings of ICPhS, 19*: 2129–2133.

Wright, Saundra K., Hay, Jennifer, & Bent, Tessa. (2005). Ladies first? Phonology, frequency, and the naming conspiracy. *Linguistics, 43*(3): 531–561.

Wu, Hsiao-Hung Iris. (2006). Stress to tone: a study of tone loans in Mandarin Chinese. *MIT Working Papers in Linguistics, 52*: 227–253.