Appendix 1. Introduction to Chinese names and Chinese romanisation systems

In modern Chinese, a full name starts with a surname, followed by a forename. There are two writing systems: Traditional Chinese, used by people in Taiwan, Hong Kong and Macau; and Simplified Chinese, used mainly by people in China and by people of Chinese heritage in other South Asian countries. Traditional and Simplified Chinese differ only by writing but share the same pronunciations.

Chinese is a character-based language. Surnames typically consist of one to two characters, and up to nine characters [1]. The Grand Dictionary of Chinese Surnames collated 11,969 Surnames, where over 90% of the Chinese population shared 120 common surnames and all of them consist of one character [1]. The top five surnames (Wang, Li, Zhang, Liu, Chen) account for over 30% of the population [2]. Forenames usually consist of one to two characters, with no upper limit. In 2020, over 90% of Chinese full names consisted of 3 characters, around 6% of full names had 2 characters, and 3% had 4 or more characters. Modern Chinese surnames are patrilineal, where infants have their father's surnames. However, about 8% of newly registered babies had their mother's surnames [2].

Standard Chinese, or Mandarin, is spoken by about 80% of the population in China; Yue Chinese, or Cantonese, is the second most spoken Chinese language with over 80 million native speakers in Hong Kong, Macau, and Guangdong, Guangxi provinces. Cantonese is also predominantly spoken by ethnic Chinese communities in Vietnam, and early Chinese migrants in Europe and North America. There are other dialects of the Chinese language. This piece focuses on Cantonese and Mandarin as they represent most Chinese speakers.

Hong Kong government cantonese romanisation (HKG-romanisation) - Cantonese

Hong Kong has used romanised Cantonese to represent places and official names since the British colonial periods (1800s) to present [3]. The romanisation system use by the Hong Kong Government is based on three legacy systems developed in the 19th century British missionaries in Hong Kong and China and it remains the official system. Linguists have since highlighted substantial inconsistencies in representing consonants, vowels, diphthongs and syllabic consonants within the HKG-romanisation system [4]. For example, the vowel "ei" represented with International Phonetic Alphabet can be represented as "ei", "ee", "ay", "ai" or "i" using HKG-romanisation. The same problem manifests in names. For example, a common surname "楊" can be represented as "yang", "young", "yep", "yong", "yeung", "yeang", "yung". HKG-romanisation does not capture tones. Each character is separated by a blank space.

Jyutping – Cantonese

Jyutping is a romanisation system developed in 1993 by the Linguistic Society of Hong Kong [5]. Jyutping represents Cantonese with 6 tones. It can represent all modern Cantonese sounds and tones with alphanumeric characters without any diacritics or other symbols: the same example surname "楊" is represented as "joeng4". Each character is separated by a blank space.

Pinyin – Mandarin

Pinyin was developed in 1950s by a group of Chinese linguists as part of a National Reform of the Chinese Written Language [6]. Pinyin represents Mandarin with 4 tones plus a neutral tone. Tones are written as diacritics, ā, á, ǎ, à, and a – neutral tones are presented without any accent marks. As per the Official Basic Rules of the Chinese Phonetic Alphabet Orthography [7], surname and forename are separated by a blank space, but sur/forename with multiple characters are not separated by a blank space. For ease of comparison, Pinyin tones are represented using numbers 1,2,3,4 and 5.

References

1. Yuan Y, Du R. The Grand Dictionary of Chinese Surnames. Beijing: Education and Science Press; 1996.
2. The Ministry of Public Security of the People's Republic of China. 《二〇二〇年全国姓名报告》发布_部门政务_中国政府网 [Internet]. 2021 [cited 2024 Aug 7]. Available from: https://www.gov.cn/xinwen/2021-02/08/content_5585906.html.
3. Kataoka S, Lee C. A system without a system: Cantonese romanization used in Hong Kong place and personal names. Hong Kong Journal of Applied Linguistics. 2008;11(1):79–98. Available from: https://repository.eduhk.hk/en/publications/a-system-without-a-system-cantonese-romanization-used-in-hong-kon.
4. Kataoka S. Finding Order in Disorder: An Investigation into the Hong Kong Government Cantonese Romanization. Newsletter of Chinese Language. 2014;93: 9–25.
5. Jyutping Cantonese Romanization Scheme - The Linguistic Society of Hong Kong [Internet]. 1993 [cited 2024 Aug 19]. Available from: https://lshk.org/jyutping-scheme/
6. Chen LL. Hanyu Pinyin. In: The Routledge Encyclopedia of the Chinese Language. Routledge; 2016.
7. GB/T 16159-2012 汉语拼音正词法基本规则 Basic rules of the Chinese phonetic alphabet orthography [Internet]. 2012 [cited 2024 Aug 19]. Available from: https://pinyin.info/rules/GBT16159-2012.html.

Supplementary Appendix 1: Surname blocking strategy comparison

Blocking rule	Recall	Precision	Reduction
block_rule1	1.000	0.746	0.966
block_rule2	0.996	0.827	0.969
block_rule3	0.996	0.761	0.967
block_rule4	0.688	0.725	0.976
block_rule5	1.000	0.574	0.956
block_rule6	1.000	0.522	0.951
block_rule7	1.000	0.076	0.664
block_rule8	0.996	0.504	0.950
block_rule9	0.996	0.318	0.920
block_rule10	1.000	0.073	0.654

Appendix 2. Surname blocking strategy comparisons

Ground Truth is defined by Chinese Surname. True positive is defined as the proportion of candidate pair being classed as the same block as the ground truth. Recall is defined as the proportion of true positives being captured across all comparisons. Precision is defined as the proportion of true matches divided by the number of candidate pairs after blocking. Reduction ratio is defined as the proportion of non-matching pairs eliminated by the blocking rule.

Blocking Rules
Rule 1: Full Jyutping + tone
Rule 2: Full Pinyin + tone
Rule 3: Full Pinyin only (no tone)
Rule 4: HKG-Romanisation

Rule 5: First 2 characters of Jyutping + tone (Surname)
Rule 6: First 2 characters of Jyutping (no tone)
Rule 7: Just the Jyutping tone

Rule 8: First 2 characters of Pinyin + tone (Forename)
Rule 9: First 2 characters of Pinyin (no tone)
Rule 10: Just the Pinyin tone

Example of the blocking rules applied.

Chinese name	surname_jyutping	forename_jyutping	surname_pinyin	forename_pinyin	surname_HKG	forename_HKG
張沛霖	zoeng1	pui3lam4	zhang1	pei4lin2	CHEUNG	PUI LAM
張恩宜	zoeng1	jan1ji4	zhang1	en1yi2	CHEUNG	YAN YI
黃雨錡	wong4	jyu5kei4	huang2	yu3qi2	HUANG	YU I JOYCE
黎恩彤	lai4	jan1tung4	li2	en1tong2	LAI	YAN TUNG
梁楷汶	loeng4	kaai2man4	liang2	kai3wen4	LEUNG	KAI MAN

Supplementary Appendix 2: Five examples names to demonstrate blocking rules

Chinese name	block_rule1	block_rule2	block_rule3	block_rule4	block_rule5	block_rule6	block_rule7	block_rule8	block_rule9	block_rule10
張沛霖	zoeng1	zhang1	zhang	CHEUNG	zo1	zo	1	zh1	zh	1
張恩宜	zoeng1	zhang1	zhang	CHEUNG	zo1	zo	1	zh1	zh	1
黃雨錡	wong4	huang2	huang	HUANG	wo4	wo	4	hu2	hu	2
黎恩彤	lai4	li2	li	LAI	la4	la	4	li2	li	2
梁楷汶	loeng4	liang2	liang	LEUNG	lo4	lo	4	li2	li	2