International Journal of Population Data Science

Journal Website: www.ijpds.org

Data Note: Alternative Name Encodings - Using Jyutping or Pinyin as tonal representations of Chinese names for data linkage

Joseph Lam[1,*], Mario Cortina-Borja[1], Robert Aldridge[2], Ruth Blackburn[1], and Katie Harron[1]

Submission History

Submitted:	29/10/2024
Accepted:	10/02/2025
Published:	xx/xx/20xx

[1] Great Ormond Street Institute of Child Health, University College London, London, UK
[2] Institute for Health Metrics and Evaluation, University of Washington, Seattle, USA

Abstract

Accurate data linkage across large administrative databases is crucial for addressing complex research and policy questions, yet linkage errors—stemming from inconsistent name representations—can introduce biases, predominantly for names not given in English. This data note examines the impact of romanisation on linkage accuracy, focusing on Chinese names and comparing standardised systems (Jyutping and Pinyin) with the non-standardised Hong Kong Government Cantonese Romanisation (HKG-romanisation). We identify three primary issues: language-specific variations in romanisation, the loss of tonal information inherent to tonal languages, and discrepancies in name order conventions. Using a dataset of 771 Hong Kong student names, our analysis reveals that standardised romanisation systems enhance the uniqueness and consistency of name representations, thereby improving linkage precision and recall compared to HKG-romanisation. Specifically, Jyutping and Pinyin achieved over 95% recall in blocking strategies, whereas HKG-romanisation only reached 68.8%. Incorporating tonal information further improved recall. These findings underscore the necessity of adopting standardised, tone-sensitive romanisation systems and flexible database designs to reduce linkage errors and promote data equity for under-represented groups. We advocate for the implementation of phonetic encodings in databases, alongside language-specific pre-processing protocols, to ensure more inclusive and accurate data linkage processes.

Keywords
data linkage; romanisation; linkage errors; data equity

*Corresponding Author:
Email Address: joseph.lam.18@ucl.ac.uk (Joseph Lam)

https://doi.org/10.23889/ijpds.v6i1.2935

Context

Differential linkage errors through romanisation

Data linkage is used increasingly across large administrative databases to enhance data resources to allow addressing complex research and policy questions [1]. Accurate data linkage is important as linkage errors can undermine the quality of the linked data by introducing selection bias, and limiting researchers' ability to accurately represent the targeted populations [2]. In data linkage, names are often used as key identifiers to decide if records from multiple data sources belong to the same person.

Romanisation refers to the transformation of names in local languages to a commonly operable language, which is English in most developed countries in the Global North. Romanisation can be viewed as a process of encoding that selectively retains information operable in (Western) databases (e.g. only containing Latin-based alphabets), and drops information deemed irrelevant or too costly to retain, such as using American Standard Code for Information Interchange (ASCII) for encoding. Unicode is an improvement over ASCII with greater flexibility on type of characters it could encode, but still does not provide tonal representations. Other examples relate to data system designs, people with multiple surnames (e.g. common in Spain and Latin-American countries) might only be allowed to provide one or use hyphenated versions, or letters not in the English alphabet and diacritical marks might be ignored or misrepresented. Romanisation is not always standardised, e.g. there are over 10 different systems for Arabic [3].

When romanised names are not consistently represented in databases, linkage errors are more likely to occur (i.e. false matches, where records belonging to different people are linked together, or missed matches, where records belonging to the same person remain unlinked). As large-scale linked administrative data are used more readily by health professionals and policy makers to generate evidence and drive decisions, this disparity in linkage rates perpetuates health and social inequities for under-represented groups, such as migrants and people from ethnic minoritised communities [4, 5]. For example, a historical linkage of census data from the United States matched only 3.6% of male Chinese migrants between 1880-1900 compared to 16.3% of English migrants [6]. A recent linkage of asylum and resettled refugees with census data in the United Kingdom found a substantial difference in linkage rates by language and country of origin [7]. There is a growing proportion of births in England (37%) and London (66%) to families where one or both parents were born outside the UK; between 2013-2017, 53% of singleton child births in New York City were born to non-US-born mothers [8]. Use of poorly romanised and processed names in data systems will selectively and continually under-represent subpopulations in results, if current data linkage 'blind spots' or structural barriers in data systems to inclusion in linked datasets are not actively addressed.

Building on Postel's work on Chinese-specific pre-processing for historical linkage [9], this data note describes three problems faced by, but not unique to representing Chinese characters for data linkage. We compare standardised romanisation of Cantonese (Jyutping) [10], Mandarin (Pinyin) [11], with the non-standardised Hong Kong Government Cantonese Romanisation system (HKG-romanisation) [12], as a case example. We then propose a solution model for alternative name encodings, by demonstrating the advantages of challenging current standard practices in global information systems to improve linkage and promoting data equity [13].

Data issues

Three problems with processing Chinese characters and names

Language-specific romanisation

Theoretically, Jyutping or Pinyin should more accurately represent pronunciations compared to the HKG-romanisation system. However, not all variations in romanised representations arise from flaws in the romanisation system. Instead, there are historic-, country- and language-specific variations in how some Chinese characters are represented in different geographic regions and countries. This variation is informative for linkage, because it can help distinguish between different people with similar names. Figure 1 shows the variations in representation of a common surname "林" in Cantonese, Mandarin, Vietnamese, Malaysian/Singaporean, Indonesian, Japanese, Korean, and other dialects. "林" is most commonly pronounced as "Lin" by Mandarin speakers. These variations of "林" are not a result of inconsistent romanisation, but a reflection of how that character is pronounced locally, and how such pronunciations change over time within the same region. Using a single unified romanisation system for Chinese characters from all countries would mean these language-specific and temporal-specific distinctions are lost. Knowing the country of origin and language system of the individuals may help to identify the best romanisation approach to retain most relevant information.

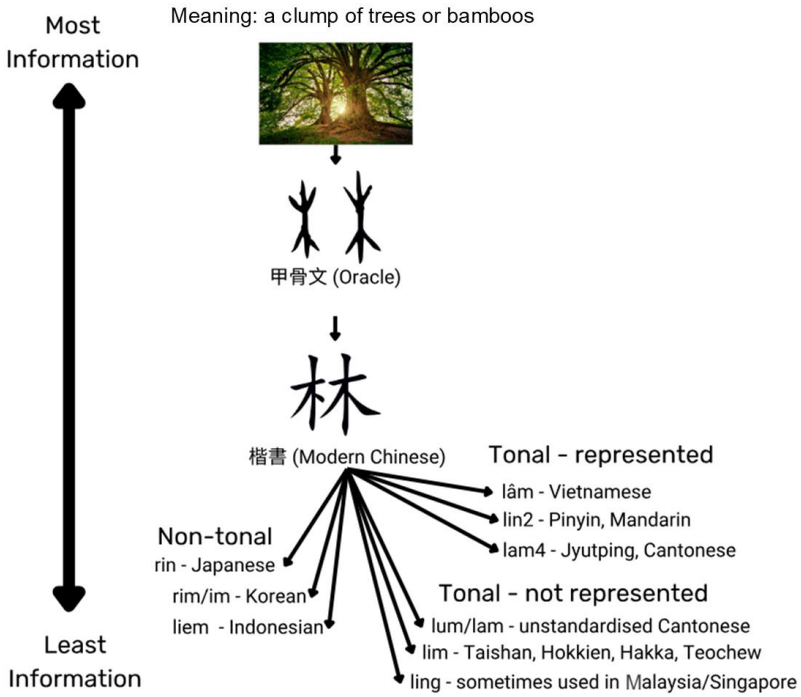
Non-tonal representation of a tonal language

Modern databases were developed for Indo-European languages in mind using Latin-based alphabets. With rare exceptions such as Panjabi (Punjabi) [14], Indo-European languages do not differentiate tones in how they are spoken. This means that even if local languages were perfectly represented by roman alphabets, romanised names will still be less specific and distinctive than their original names as tonal information is not retained. This loss of granularity degrades the identifying information available and could lead to linkage errors. This is a particular problem for HKG-romanisation and Pinyin, as intonations are not represented in the former, and not properly recorded in the latter as diacritics (notations on top of each word that are necessary for tones) are often ignored. Jyutping and a properly recorded Pinyin system are more sensitive to tonal changes, and a more consistent representation of Chinese characters than HKG-romanisation.

Name orders

A common issue for character-based languages is the misplacement of character orders. For example, Chinese surname and forename are often inverted due to different

Figure 1: Illustration of different ways the character "林" is romanised and pronounced by countries and languages. Photo by Johannes Plenio: https://www.pexels.com/photo/two-brown-trees-1632790/



naming conventions; forename characters are sometimes misplaced as middle names [15]. The latter is more prevalent in HKG-romanisation and Jyutping, as each character is separated by a space; and less so for Pinyin as there is no space between forename characters. The inability to segment which characters belong to surname or forename fields means that linkages may be inaccurate. Previous attempts, e.g. the Abramitzky, Boustan, Eriksson (ABE) method [6], to pre-process multi-part names have resorted to clustering multi-part names to their first character, e.g. clustering "Chin Fung", "Chin Hing", "Chin Lung" as "Chin". Note that these are all HKG-romanisation; linking Cantonese-based romanised names would be predominantly impacted using the ABE method. Chinese surnames are already less specific for linkages than English, with larger clusters of people sharing the same common surnames (Appendix 1). The ABE method further lowers the specificity of Chinese forenames, resulting in linkages that are more prone to false matches. Postel demonstrated that proper segmentation, indexing and ordering of Chinese characters could substantially improve linkage rates for Chinese names [9].

In the following section, we will compare the utility of Jyutping, Pinyin and HKG-romanisation in representing Chinese characters, sensitive to language-specific romanisation, tonal representation and name orders.

Proposed solution

We scraped online student class lists from schools in Hong Kong that provided both Chinese and English Names (n = 774). We only included names that had a Mandarin or Cantonese origin, based on the provided Chinese and (romanised) English names (n = 771). Records providing English names with no space within forename characters, or use of non-accented Pinyin as English names reflect a Mandarin origin.

We derived Jyutping and Pinyin using the pinyin_jyutping package [16] in Python 3.8 [17],which used an online open-source Cantonese dictionary CC-Canto [18], for each character of the Chinese name. Jyutping and Pinyin tones are represented using numbers: 1-6 for Cantonese and 1-5 for Mandarin. We provided a short introduction to Chinese names and romanisation systems in Appendix 1. We manually entered the records that the package failed to translate, and stored pronunciations and tones in separate columns. Raw name list, cleaned data and codes are available from a University College London Research Data Repository [19].

Of the 771 included names, a large majority (97.7%) had a 3-character full name, most (83.9%) had a romanised English forename, and most names were given based on Cantonese (97.4%) (Table 1).

We compared how closely three different systems of romanisation (HKG-romanisation, Jyutping, Pinyin) represented the original Chinese characters, in terms of uniqueness, which provided some information on the utility of these systems for balancing sensitivity and specificity of linkages.

Our sample of 771 individuals all had unique Chinese full names and shared 123 unique surnames (Table 2). The top five most frequent surnames accounted for over 30% of surnames. HKG-romanisation resulted in 152 unique surnames with 29 extra surnames than Chinese. Both Jyutping and Pinyin reduced the number of unique representations of names. The HKG-romanisation system represented different Chinese characters using the same codes, for example, "Chiu" is used to represent "趙" (Jyutping: Ziu6, Pinyin: Zhao4) and "邱" (Jyutping: Jau1, Pinyin: Qiu1). The HKG-romanisation also represented the same Chinese characters using different codes,

Table 1: Descriptive characteristics of Chinese and English names in the study dataset

Descriptive characteristics		Count	%
Total *n*		771	100
Chinese surname			
	1 Character	771	100
Chinese forename			
	1 Character	17	2.3
	2 Characters	754	97.7
English forename			
	Romanised only (e.g. Chin Hang)	647	83.9
	English only (e.g. Johnny)	27	3.5
	Both English and romanised (e.g. Chin Hang Johnny)	97	12.6
Language			
	Cantonese	751	97.4
	Mandarin	20	2.6

for example, "周" is represented as "Chow", "Chau", or "Chiau", where Jyutping would consistently represent it as "Zau1", and Pinyin "Zhou1". Both Jyutping and Pinyin represented the same characters consistently, but both represented different characters using the same codes, for example, "Wong4" for both "黃" and "王" in Jyutping, and "Yan2" for both "顏" and "嚴" in Pinyin. Pinyin without tones had all the problems of Pinyin, plus the inability to differentiate tones, hence was less specific.

There were 743 unique combinations of forenames in Chinese, corresponding to 641 unique Jyutping representations, and 679 Pinyin representations (Table 2). In our data of predominantly Cantonese-based names, we observed more unique Pinyin forename combinations than Jyutping. Similar to naming clusters in other languages, Hong Kong people tend to use variations of similarly sounding names. For example, forename "Paak3Hei1" occurred six times, but each represented a unique Chinese forename. Pinyin would represent these six names in four different ways. In this case, Pinyin becomes more specific than Jyutping in differentiating people in our sample. We expect the specificity of Pinyin to be lower in a majority Mandarin-speaking sample, or in a larger population with more variations in names. Both Jyutping and Pinyin seem to be able to increase linkage quality in dealing with pronunciation-based errors and rare names.

Implication for linkage

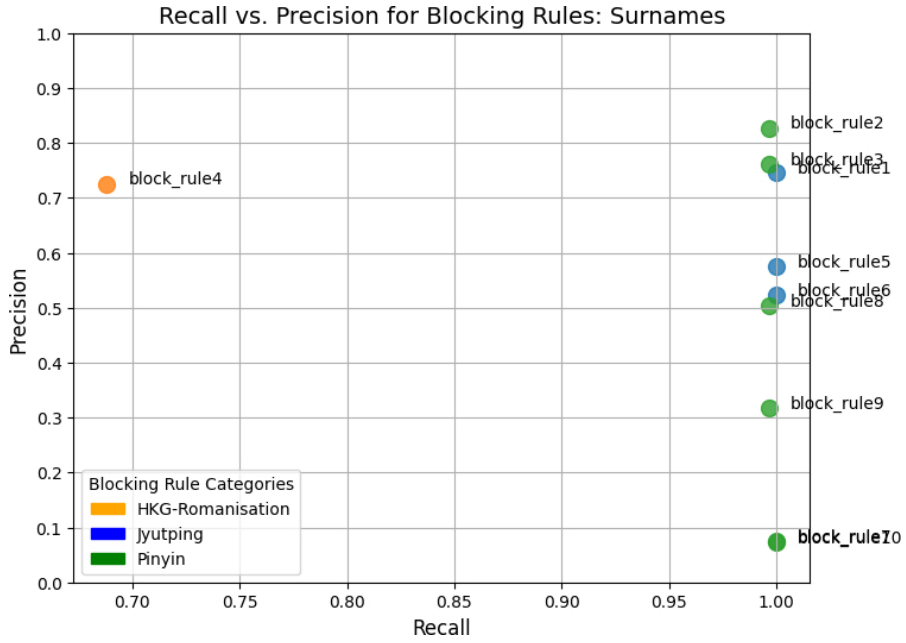
We compared how different name encoding strategies impact on blocking. Since forename and surname combination almost uniquely identifies every individual in our dataset, we only used blocking on surnames as an example. We evaluated whether different encodings of the same surname would be put into the same block and produced recall and precision statistics based on the ground truth (Chinese surnames). Blocking based on full HKG-romanisation performed the worst in terms of recall at 68.8%, whilst both Jyutping and Pinyin (with tonal information) achieved over 95% recall, at comparable or

Table 2: Count of unique values at each field using Chinese characters, Jyutping, Pinyin, Pinyin_notone (without tonal information) and HKG-romanisation for 771 names. In brackets, degree of loss or addition of unique values, using different romanisation methods compared to original Chinese characters (%)

Unique count	Chinese	Jyutping	Pinyin	Pinyin_notone	HKG-romanisation
Surname (1 char)	123	117 (−4.9%)	120 (−2.4%)	108 (−12.2%)	152 (+23.6%)
1-to-1		35	39	34	27
1-to-1 (occurred >1)		66	67	54	24
Many-to-1		16	14	20	11
1-to-Many (contain duplications)		–	–	–	49
Forename (1-2 char)	743	641 (−13.7%)	679 (−8.6%)	642 (−13.4%)	687 (−7.5%)
1-to-1		555	608	557	600
1-to-1 (occurred >1)		8	13	11	3
Many-to-1		77	57	74	60
1-to-Many (contain duplications)		–	–	–	14
Full name	771	767 (−0.5%)	769 (−0.3%)	763 (−1.0%)	770 (−0.1%)

We also described how many names have 1-to-1, 1-to-many, and many-to-1 correspondence.

Figure 2: Comparison of precision and recall by blocking rules based on different romanisation systems. Blocking rules are described in detail in Appendix 2



superior precision (Figure 2). Full comparison of other blocking rules is described in Appendix Table 1 and 2 (Appendix 2). Block rules 5,6 and 8,9 compares blocking using first two characters of Jyutping and Pinyin respectively with and without tonal information. Incorporating tonal information improved precision in both cases.

Informative missingness

A further promise of tonal representations of Chinese names is the potential of using tones to impute missing characters. Frequencies of tonal combinations of Jyutping and Pinyin names likely follow a Zipf's distribution [20], where a few combinations represent most names and a long tail of combinations have very few counts. In our sample, the top 10 tonal combinations represented 38.3% names in Jyutping and 45.0% in Pinyin (Figure 3). By further incorporating vowel information, statistically estimating missing tonal information should help calibrate non-tonal romanisation systems. For example, in a 3-character name where the first 2 characters have the Pinyin tone "2-3" and the last character is missing, we can expect it is at least twice as likely for the third character to have a "2" tone than a "4" tone. Combining tonal information with frequency of vowel combinations for each character, we could be potentially re-encoding romanised Chinese names. Further work using a large-scale Chinese names database would contribute to this regard.

Generalisable lesson

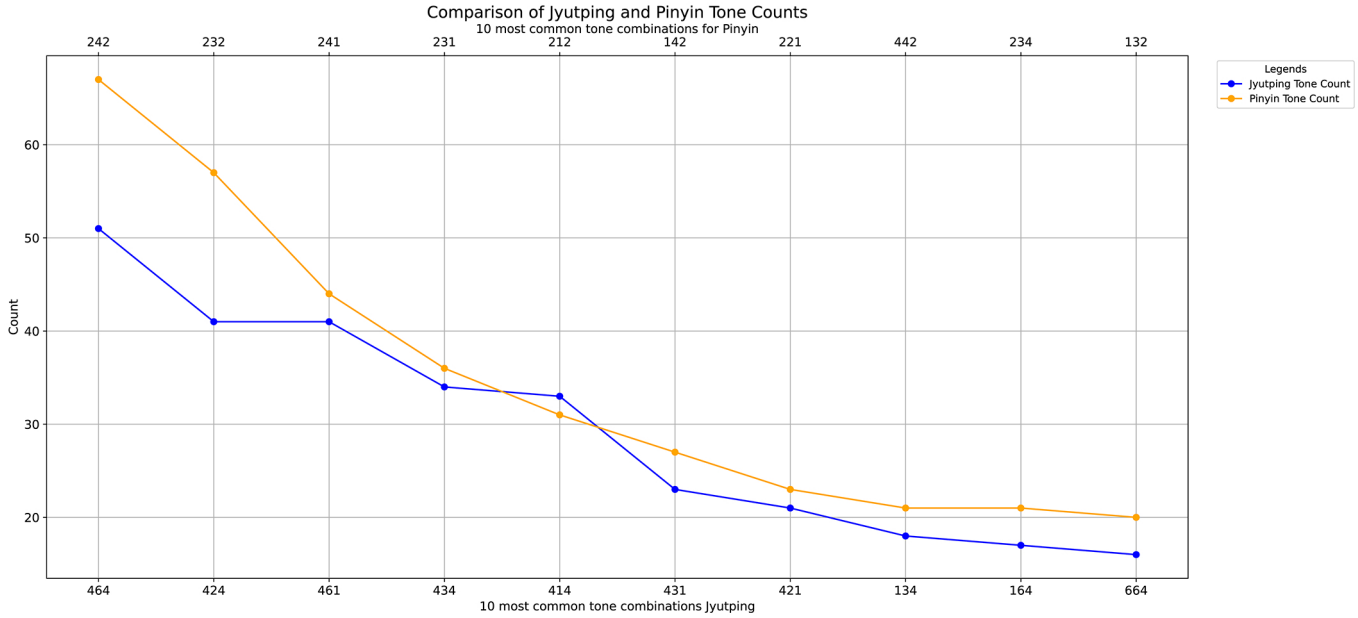
We demonstrate that both Jyutping and Pinyin are promising alternatives to representing Chinese characters, accounting for pronunciation-based errors, rare names and tonal information, compared to HKG-romanisation or non-tonal Pinyin.

For data linkage researchers with access to Chinese names along with HKG-romanised named, using alternative encodings such as Jyutping and Pinyin would substantially improve blocking performance. Contextual knowledge on data sources can help in choosing appropriate blocking and linkage strategies. For example, in the 1880-1990 US linkage, Chinese names in the US Census are more likely to be based on Cantonese than Mandarin. Blocking strategies based on Jyutping surnames may therefore increase efficiency of linkage by identifying more true matches, but will increase the number of missed matches, compared to Pinyin or other romanisations. As romanised Chinese names are ascribed by immigration officers in this particular setting, Jyutping will represent forenames more precisely and sensitively than Pinyin for this linkage.

Phonetic encodings can be used independently to deal with missing data or used to calculate probabilistic weights. We recommend including each term of the character and phonetic encoding separately to provide more flexibility in weight adjustment. The above recommendations are limited to datasets not already encoded. We are not aware of any approaches that could consistently re-encode tonal languages, especially when tonality of characters is not captured in romanised encoding. Using frequency-based methods to probabilistically assign re-encoded characters could be possible, by incorporating character position, vowels, and tonality. However, this approach may require strong assumptions on the distribution of the tones in any established dictionary to be the same as the sample. Future work should explore the feasibility and contexts in which re-encoding is suitable.

Developing a language-specific pre-processing and romanisation approach requires leadership and skills from people who speak these languages. We ask database owners in developing countries or former colonial regions (such as

Figure 3: Top 10 most common tonal combinations for 3-character Chinese names, in Jyutping (bottom x-axis) and Pinyin (top x-axis). Number on y-axis corresponds to count of names with those tonal combinations



Hong Kong), where English remains the dominant language in which these databases are operating, to consider how names can be best represented and how tonal information can be retained in their databases. As for developed countries, tone-sensitive romanisation systems provide more flexibility in developing linkage strategies and could improve linkage quality for minoritised ethnic populations and migrants. Operationally, with vast advancement in voice-to-text transcription, asking individuals to say their name in their mother tongue may be a simple way to record extra information that is conducive to data linkage. This is relevant for other tonal languages, as well as character-based non-tonal languages.

Collecting, preserving and utilising people's names in their original languages, or alternatively standardised romanisation systems, is ethically and socially pertinent and may support the development of language-specific pre-processing and linkage strategies that result in more inclusive research data that better represents the targeted populations.

Recommendations

For data linkage, where possible, use systematic romanisation systems (Jyutping/Pinyin) instead of non-systematic romanised systems (HKG-romanisation) to identify unique records. Phonetic encoding can potentially add benefits over non-tonal representations for blocking and linkage. Where names are only available in romanised formats, character re-encoding is possible. While such an approach has potential, it comes with strong assumptions on tone distribution that rely on established name-tonal databases for each language. We recommend analysts to develop language-specific pre-processing algorithms to enhance linkage rates for known under-represented groups.

For database design and management, we recommend switching to Unicode or other encoding standards to capture non-alphabetical characters. Database managers should also aim to explore database design to additionally capture tonal information and people's names in their original languages. This enables data linkers to apply the name romanisation method that is most appropriate for the intended linkage.

Acknowledgement

We would like to sincerely thank the reviewers for their valuable comments and suggestions which helped improve and clarify this manuscript.

Ethics statement

No ethics approval is required as only openly accessible data are used.

Contributions

JL conceptualised the project, designed, analysed and wrote up the first draft. All authors contributed to critical reviewing and revising the manuscript. All authors read and approved the final manuscript before submission and agreed with the decision to submit the manuscript.

Funding

This work was supported by the Wellcome Trust [212953/Z/18/Z].

Conflict of interests

All authors declare no competing interests.

Availability of data and materials

Raw name list cleaned data and codes are available on UCL Research Data Repository.

References

1. Harron K. Data linkage in medical research. BMJ Medicine. 2022 Mar 1;1(1). Available from: https://bmjmedicine.bmj.com/content/1/1/e000087. https://doi.org/10.1136/bmjmed-2021-000087

2. Doidge JC, Harron KL. Reflections on modern methods: linkage error bias. International Journal of Epidemiology. 2019 Dec 1;48(6):2050–60. https://doi.org/10.1093/ije/dyz203

3. UNGEGN Working Group. Arabic REPORT ON THE CURRENT STATUS OF UNITED NATIONS ROMANIZATION SYSTEMS FOR GEOGRAPHICAL NAMES [Internet]. United Nations; 2018. Available from: https://arhiiv.eki.ee/wgrs/rom1_ar.pdf.

4. McGrath-Lone LM, Libuy N, Etoori D, Blackburn R, Gilbert R, Harron K. Ethnic bias in data linkage. The Lancet Digital Health. 2021 Jun 1;3(6):e339. https://doi.org/10.1016/S2589-7500(21)00081-9

5. Lam J. Terminating bias: How Arnold Schwarzenegger showed us the importance of spelling names correctly. Significance. 2024 Nov 1;21(5):36–41. https://doi.org/10.1093/jrssig/qmae078

6. Abramitzky R, Boustan LP, Eriksson K. Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration. Am Econ Rev. 2012 Aug;102(5):1832–56. https://doi.org/10.1257/aer.102.5.1832

7. ONS. ONS website, methodology article, Refugee integration outcomes data-linkage pilot: Census 2021 linkage methodology [Internet]. 2023 [Accessed 11 Jul 2024]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/internationalmigration/articles/refugeeintegrationoutcomesdatalinkagepilot/census2021linkagemethodologyupdate#linkage-methods.

8. Office for National Statistics (ONS), released 17 August 2023, ONS website, statistical bulletin, Births by parents' country of birth, England and Wales; 2022 [Internet]. 2023 [Accessed 19 Aug 2024]. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/bulletins/parentscountryofbirthenglandandwales/2022.

9. Postel HM. Record Linkage For Character-Based Surnames: Evidence from Chinese Exclusion. Explor Econ Hist. 2023 Jan;87:101493. https://doi.org/10.1016/j.eeh.2022.101493

10. Jyutping Cantonese Romanization Scheme - The Linguistic Society of Hong Kong [Internet]. 1993 [Accessed 19 Aug 2024]. Available from: https://lshk.org/jyutping-scheme/

11. Chen LL. Hanyu Pinyin. In: The Routledge Encyclopedia of the Chinese Language. Routledge; 2016.

12. 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.

13. Myers MD, Klein HK. A Set of Principles for Conducting Critical Research in Information Systems. MIS Quarterly. 2011;35(1):17–36. https://doi.org/10.2307/23043487

14. Stuart-Smith J, Cortina-Borja M. A law unto themselves? An acoustic phonetic study of tonal consonants in Panjabi. In: Willi A, Probert R, editors. Oxford: Oxford University Press; 2012. p. 61–82. Available from: https://eprints.gla.ac.uk/70581/.

15. Christen P, Schnell R. Thirty-three myths and misconceptions about population data: from data capture and processing to linkage. International Journal of Population Data Science. 2023 Jan 31;8(1). Available from: https://ijpds.org/article/view/2115. https://doi.org/10.23889/ijpds.v8i1.2115

16. Luc W. pinyin-jyutping: Convert a Chinese sentence to Pinyin or Jyutping. 2023.

17. Python.org [Internet]. 2024 [Accessed 9 Dec 2024]. Welcome to Python.org. Available from: https://www.python.org/.

18. CC-Canto – A Cantonese dictionary for everyone [Internet]. [Accessed 20 Aug 2024]. Available from: https://cantonese.org/.

19. Lam J. Jyutping Project - Raw Data and Clean Data [Internet]. University College London; 2024 [Accessed Aug 19 2024]. Available from: https://rdr.ucl.ac.uk/articles/dataset/Jyutping_Project_-_Raw_Data_and_Clean_Data/26504347/1.

20. Newman M. Power laws, Pareto distributions and Zipf's law. Contemporary Physics. 2005 Sep;46(5):323–51. https://doi.org/10.1080/00107510500052444

Abbreviations

ABE:	Abramitzky, Boustan, Eriksson method
HKG-romanisation:	Hong Kong Government Cantonese Romanisation system
US:	United States

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