基于计量的百年中国人名用字性别特征研究

杜冰洁　　刘鹏远*　　田永胜
北京语言大学　信息科学学院
国家语言资源监测与研究平面媒体中心
北京市海淀区学院路15号, 100083
blcudbj@gmail.com　liupengyuan@pku.edu.cn　blcutys@gmail.com

摘要

本文构建了一个包含11万以上条目规模的中国名人人名数据库，每条数据含有人名、性别、出生地等社会文化标签，同时含有拼音、笔画、偏旁等文字信息标签，这是目前已知最大的可用于研究的汉语真人人名数据库。基于该数据库，本文从中选择1919年至今的人名，用定性与定量结合的方法探究人名中汉字的特征和其性别差异以及历时变化。从人名长度来看，男性人名比女性人名长；从人名用字的难易度来看，女性用字比男性更复杂；从用字丰富度来看，人名用字越来越单一和集中化，男性人名的用字丰富度大于女性人名。计算人名用字的性别偏度后发现女性人名的专用自更多。两性用字意象有明显的不同，用字的意象随着时间发生改变，但改变最明显的时间节点是改革开放前后，其中女性的变化比男性显著。除此之外，我们还得出人名中的性别极性字表、各个阶段的高频字表、用字变化趋势表等。

关键词： 中国人名数据库；汉字；性别差异；人名历时变化；定量分析

A Quantified Research on Gender Characteristics of Chinese Names in A Century

Bingjie Du　　Pengyuan Liu*　　**Yongsheng Tian**
Beijing Language and Culture University, School of Information Science
Language Resources Monitoring and Reserch Center
15 Xueyuan Road, Haidian District, Beijing, 100083, China
blcudbj@gmail.com　liupengyuan@pku.edu.cn　blcutys@gmail.com

Abstract

In this paper, a database of Chinese celebrities' names with a scale of more than 110,000 entries is constructed. Each data contains social and cultural labels such as names, gender, and birthplace, as well as Chinese character information labels such as Pinyin, strokes and character components, which is the largest known database of Chinese real people's names that can be used for research. Based on this database, this paper selects names from 1919 to the present and uses a combination of qualitative and quantitative methods to explore Chinese names in character characteristics, gender differences, and diachronic changes. Through research, it is found that the average number of Chinese characters in women's names is higher than that of men. The Chinese characters in female names are more complicated but lower richness than that of males. The use of personal names has become more monotonous and centralized with time. The imagery of Chinese characters used in the names of the two sexes is obviously different. The imagery of the characters changes over time, but the most obvious time point for the

* 通讯作者　Corresponding Author

第十九届中国计算语言学大会论文集，第21页-第31页，海南，中国，2020年10月30日至11月1日。

change is around the reform and opening up, in which the change of women is more significant than that of men. Besides, we also obtained the gender polarity characters table of the names, high-frequency characters table of each stage, characters change trend table, etc.

Keywords: Chinese name database , Chinese Characters , Gender differences , Diachronic change of names , Quantitative analysis

1 引言

人名是不同个体为区分彼此而创造出的指称符号。人名既特殊又普遍，其特殊性表现在，人名属于词汇系统中专有名词的一种，具有指称的唯一性和确定性；其普遍性表现在人名在社会生活中的出现频率极高，在社会系统的正常运作中扮演着十分重要的角色，我们需要“说”名字，也需要“写”名字。与字母文字不同的是汉字具有表意的功能。因此中国人名不仅具有读音上的特殊性，在字形、字义上也具有特殊性，对于人名用字的研究也就显得十分重要。

本文建立了一个大规模中国名人人名数据库，从汉字本体的角度做了跨度长达百年的人名用字分析，同时从性别的角度展开，探究人名中汉字的性别差异。本文发现两性人名在长度、难易度、丰富度、变化趋势等方面都存在显著差异。本文贡献在于：1）建立了目前已知最大规模的真人人名数据库；2）分别从汉字本体及计量语言学两种研究视角进行了人名用字研究，这些研究方法应用到人名中被证明具有一定价值；3）得到了百年两性人名用字特征的差异与演变规律。

2 相关工作

对人名的语言学研究侧重语音、语义等。部分文献对于人名中的汉字，有所体提及。一些文献中(苏培成, 2001; 吴继章, 2001; 邱莉芹and 鞠泓, 2002; 张书岩, 1999; 张书岩, 2004)探讨了人名中出现的的生僻字、多音字、异体字等问题，提倡入名汉字应该规范化。赵越 (2006)、何晓明 (2001)提出中国人取名对合体字独体字等不同字形的讲究，遗憾的是作者并没有针对这一问题进行深入阐述。谢玉娥 (2000)、Jia and Zhao (2019)认为人名中的部分汉字具有性别偏向，但其讨论的主要是汉字的意义上的偏向问题，且没有做详细的定量统计和分析，也未从汉字本体的角度进行考察。关于汉字与性别之间的关系，韩燕et al. (2008)采用事件相关电位(ERP) 技术证明汉语人名具有性别刻板印象。王玉新 (2000)和潘世松 (2004)从汉字的偏旁结构和发展规律论述了汉字结构本身的性别歧视现象。

以上关于人名的研究多是基于几百上千条人名，样本数量较小。从研究方法来看，多是共时研究，或者两个时间段的对比研究。关于人名的历时研究时间跨度较小，难以宏观反映人名在一段时间范围内的变化。人名中的汉字研究较为单一，多是从汉字意象的角度进行解释，人名中汉字本身的特征研究几乎没有。而有关认知科学的实验又证明人名是具有一定刻板印象存在的，因此关于人名汉字的性别倾向研究可进一步探讨汉字性别歧视的现象。同时，目前缺乏一个公开的、可支持人名共时历时研究的中国人名数据库。

3 数据

3.1 人名数据库构建

本文基于知识图谱中的信息，来建立中国名人人名数据库。选择名人来建设人名数据库的原因有二：1.可最大程度上保证人名及相关信息的真实性；2.本文假设是否能成为名人与其姓名不相关，因此该语料库也可以支持对中国人名的其他研究。构建过程主要如下：

1）抽取。从百科人物知识图谱[0] 抽取了名人相关信息（这样可以较好的保证人名信息的真实准确性），抽取的条目具体为：姓，名，性别，出生日期，出生地；

2）筛选。原百科人物知识图谱中的每个名人信息分布混杂，并且有大量国外名人信息。为了最大化获得人名数据并且保证数据包含所需的几个维度，在抽取过程中我们主要通过姓名长

[0]http://openbase.openkg.cn/

度、出生地等筛选中国名人姓名。中国人姓名的主流格式是两个字的姓名，即姓+单名、三个字的姓名，即姓+双名，因此我们将姓名长度限定为两个字和三个字。在出生地方面，我们将关键词限定为中国所有省市，并将添加了“中国“村”等不同粒度的关键词。考虑到有些名人信息不包含出生地信息，我们又添加了“民族”这一判断规则；

3）信息补充。为丰富对中国人名的研究，我们又为每个名人条目补充了拼音、笔画、偏旁等信息。该信息来源于一个开源中华新华字典数据库[1]。对于部分不在字典数据库的汉字，我们利用人工的方法补充字典信息再进行匹配。

最终建成的中国名人人名数据库[2]共有111564个条目，每个条目包含姓、名、性别、出生地以及人名用字的拼音、笔画、偏旁等信息，其中有男性人名条目83706条，女性人名条目27858条。在这些条目中有54264条包含出生日期，时间跨度从古代至今，主要以近现代人名为主。该语料库可为中国人名多维度研究提供数据支持。

3.2 研究对象

从上述的人名数据库中抽取1919年至今等性别比例的人名作为研究对象[3]，从汉字本体的角度探究人名的长度、人名汉字难易度、人名用字变化在两性中的差异及其历时发展变化规律，从侧面了解近一百年中国社会政治、经济、文化的发展变化。

所选时间段中的人名中共出现了2342个字种，男性人名中的字种有1800个，女性人名中的字种有1807个。之所以选择这一时段是因为本文希望对近现代近100年的人名从汉字的角度做定量分析，而1919年作为近代史开端，自然成为本次研究的时间起点。本文对数据做两种划分，详见表1：

按自然年份划分	1919-1938	1939-1958	1959-1978	1979-1998	1999至今
男名	1168	2637	5275	3742	106
女名	1202	2670	5208	3624	224
按重要历史事件划分	1919-1949	1950-1965	1966-1978	1979至今	-
男名	2279	3689	3112	3848	-
女名	2279	3689	3112	3848	-

Table 1: 研究对象的历时划分

1）按照自然年份划分。本文希望能以时间均匀的角度观察人名变化的规律。每二十年一个阶段划分，共五个阶段；

2）按照重要历史事件为时间点进行划分。本文假设重大政治经济事件对人们起名的影响较大。这个方式主要用于对自然年份划分的对比和说明。

基于中国百年人名数据库及表1的划分，文本希望能揭开百年来中国人名变化趋势的一角，并试图回答以下问题：1）人名在长度上是否有性别差异？百年来人名长度变化情况与原因？2）在用字难易度上是否有性别差异？百年来人名用字难度变化情况与原因？3）在用字丰富度上是否有性别差异？百年来人名用字丰富度的变化情况及原因？4）具体用字是否在时间维度上有显著差异？百年来人名具体用字变化情况及原因？

4 人名用字性别差异及历时分析

4.1 名字长度

文字同语言一样是一种信息交流的工具，人名中的汉字是记录人名内涵的书写符号。很多汉字都能独立表达一定含义，人名的内涵可以通过每个汉字的排序、人名汉字的多少（即名字长度）等来传达。本文将只有一个字的名字称为单名（如：杜甫），两个字的名字称为双名（如，周树人）。

[1]https://github.com/HelloDreamen/chinese-xinhua。共收录了16142个汉字的相关信息。

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

[3]本文仅考虑常规两个字和三个字的姓名。对于传统的复姓，因样本相对较少，故不在研究范围内。近些年出现的类似“王张”的双姓，本文当做单姓处理。

我们计算了数据中两性人名的平均长度，其中女性人名长度均值2.88，男性人名长度均值2.91。随后做了皮尔逊卡方检验[4]，结果表明，男女人名中的单名和双名分布存在统计学意义上的差异。具体统计结果见表2:

	值	自由度	渐进显著性（双侧）
皮尔逊卡方	75.287[a]	1	.000
连续性修正[b]	74.937	1	.000
似然比	75.565	1	.000
有效个案数	25856	-	-
a. 0 个单元的期望计数小于5,最小期望计数为1374.00;b. 仅针对2x2 表进行计算			

Table 2: 性别与人名长度卡方检验

为了解各个阶段单双名的具体分布情况，下图从自然年份的划分观察近百年来中国人名长度的变化:

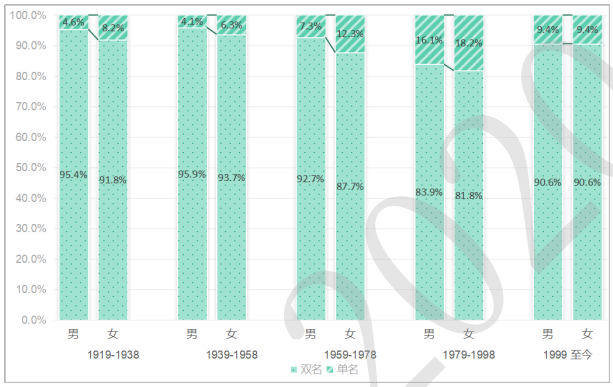


Figure 1: 自然年份中两性人名单双名的变化趋势

由图1可知：1）女性人名中的单名比例一直高于男性，但是两性人名中的单双名差异随着时间的发展不断缩小；2）总体而言，人名中单名双名的比例呈现一定波动，但是双名占据绝对优势；3）1979-1998区间单名比例是一个峰值，但是进入21世纪，单名比例又有所降低，几乎与1919-1938区间单名比例持平。

双名是魏晋门阀制度盛行，强调宗族家谱以后才逐渐占据主流的。因为族谱的存在，在取名时中间一个字需要固定，因此中国人名大多数是双名，而家谱和宗法制度有着密切的联系，宗法制度强调与家族中男性长辈的血缘亲疏，其主要精神为“嫡长继承制”。这就会对家族中男性晚辈身份的约束。所以男性的双名比例始终高于女性，反之女性单名比例高于男性。即使除去按字辈取名的习俗，双名有两个汉字承载的信息量也大于单名，增加名字内涵的同时避免了重名的概率，所以双名一直占据主流地位。但是随着近现代中国各种思想解放运动展开，一定程度上打破了传统的宗法制度，按照字辈取名的习俗也逐渐减少。原本按照字辈取名的双名，实际上只添加了名字末尾的一个新信息，如今中间的字没有了，依然只需要添加一个新信息，所以单名的比例就呈现增高的趋势。在改革开放初期，随着思想观念的进一步解放以及追求个性的心理特征，单名比例不断增长，在1979-1998年到达高峰。但是进入21世纪，随着人口的增多、姓名规范意识的增强，单名比例又开始下降，双名逐渐增多。

4.2 用字难易度

本文的难易度仅指书写难易程度或认读的难易程度。我们采用汉字常用等级这一指标衡量人名汉字的难易度，并以国家语言文字工作委员会1988年1月发布《现代汉语3500常用字表》作为标准。该字表中的字主要满足基础教育和文化普及的基本用字需要，因此人名中的汉字若是来自于该字表，则用字在认读或书写上相对简单。该表中的汉字有一级常用字和二级常用字之分，对于不在字表中的字统称为用非常用字，即用字相对复杂。

[4]本文所有统计检验均采用SPSS Statistics 25.0.0软件计算得出。

按照性别分组，采用卡方检验验证用字难度是否存在统计学意义上的差异，结果表明，男女人名用字的难度具有统计学意义上的差异。具体统计结果见表3:

	值	自由度	渐进显著性（双侧）
皮尔逊卡方	899.676[a]	2	.000
似然比	908.557	2	.000
有效个案数	48964	-	-
a. 0 个单元格的期望计数小于5;最小期望计数为2204.47。			

Table 3: 性别与用字难度卡方检验

由于单名在编码长度上比双名短一位，大大增加了重名的可能。取名人为了降低重名的概率，会在选字入名时有意选择一些复杂的、不常用的字。由于女性的单名比例高于男性，为排除名字长度的影响，本文将单名和双名分离，分别研究两性人名用字的难易程度。下图2是按自然年份划分的单名用字难易程度的分布。

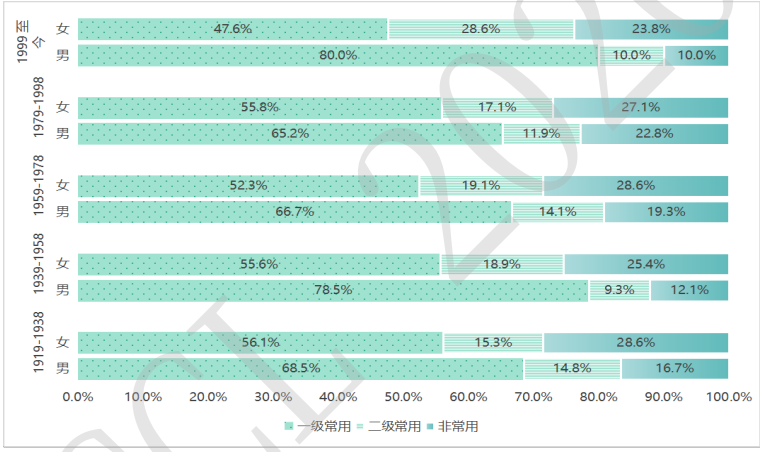


Figure 2: 自然年份中用字难易程度的性别差异（单名）

由图2可知：1）单名用字以一级常用字为主，其次是非常用字，最后才是二级常用字。二级常用字比较稳定，非常用字最不稳定；2）女性单名用字较之男性更复杂，但是两性人名的非常用字总体呈现出下降的趋势。男性用字难易度波幅较大，较女性更加不稳定；3）对比图3后发现，人名的非常用字出现过两次低谷阶段，第一次低谷时期是1939-1958年附近，第二次则是1999年至今。在1966-1979年附近，非常用字出现过一次激增阶段。

人名主要的功能是区分彼此并且满足社会成员之间的互动，所以名字要便于辨认，因此常用字总体占比会更多。但同时人名也是个人身份的标签，为了体现个性或者表达某些特殊含意，也会出现特殊的字，因此非常用字也会占据一定比例。而正是出于这些特殊的情感表达，会在某些特定时间段内出现较大变化，于是就会出现上图2,3中所看到的较大起伏。值得注意的是，男性人名的非常用字波幅大于女性，男性非常用字的激增阶段对应的是1966-1978这一时期，其名字更容易与特定事件和时段挂钩。人名非常用字的两次低谷对应的原因可能有所不同。第一次非常用字的低谷伴随的是一级常用字的增长，而第二次则伴随的是二级常用字的增长。我们的猜想是第一次低谷相应时期的新生儿父母多经历了多年战争，受教育机会少，文化程度低，在取名的时候受到文化程度等因素的限制，用字相对其他时期更加简单。而第二次低谷则更可能是新时期语言文字工作者和有关政府对于姓名规范的呼吁。

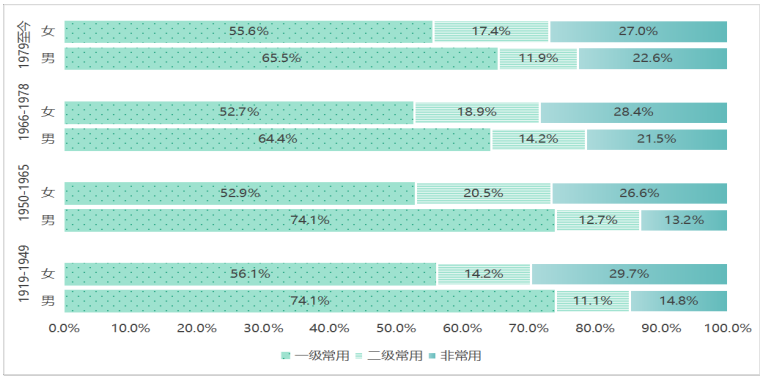


Figure 3: 重大历史事件中用字难易程度的性别差异(单名)

图4是按自然年份划分的双名用字难易程度的分布。

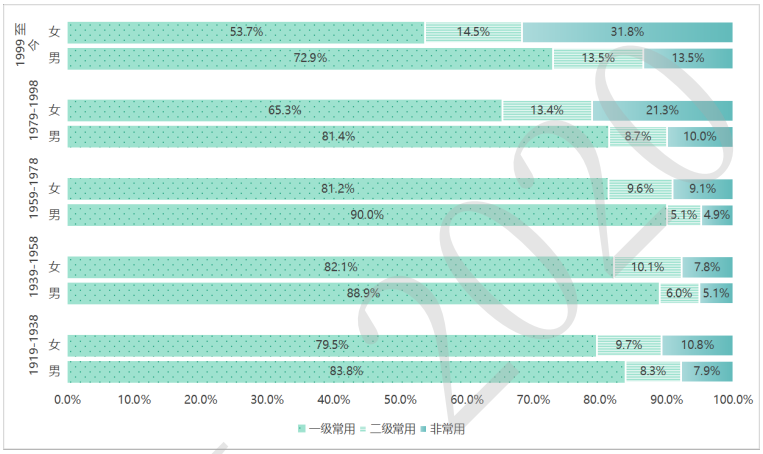


Figure 4: 自然年份中用字难易程度的性别差异（双名）

由图4可以知:

1）与单名类似，双名的一级常用字占据绝对优势，二级常用字最稳定；2）女性双名的整体用字仍然难于男性，两性人名用字在1978年后都发生了显著变化，一级常用字减少，非常用字呈现上升趋势；3）改革开放后人名用字与前三个阶段有明显区别，而这期间又分为两个时期，21世纪后的两性用字的难易程度差异变大，女性非常用字增幅比男性更显著。

虽然双名难易度的三个等级分布于单名类似，但是整体上双名用字比单名简单，这印证了本节开头的假设即用字的难易度与名字难易度有一定关联。与单名逐渐变得更简单所不同的是，双名逐渐变得复杂，这与二者的基数有关。

4.3 人名用字丰富度

在计量语言学中测量词汇丰富度常用的指标是型例比（TTR，即type-token ratio），该指标计算的是文本中不同的词语在所有词语中所占的比例，TTR值越大,说明文本使用的词汇越丰富。但是TTR的值受到语料库大小的影响，因此我们在本文中使用的“吉罗指数”(Teich, 2012; Yu, 2010)，它是TTR的变体，可减少文本大小对于丰富度的影响，其公式为:

$$Index \quad of \quad Guiraud = Types/\sqrt{Tokens}$$

其中，types是型符，tokens是类符。我们将词汇丰富度这一指标应用到人名用字中，计算人名用字的丰富度，计算得到历时百年的人名用字总体丰富度为11.45。以施建刚and 邵斌 (2016)计算的“兰卡斯特现代汉语语料库”传记和散文子库中的吉罗指数参照，该语料库的吉罗指数为43.93，远高于人名用字的吉罗指数，这说明相对于普通汉语书面语，人名用字的丰富度较低，用字比较集中。

为探究不同年代用字丰富度的变化以及两性人名用字丰富度的差异，我们分别按照自然年代的划分计算了男性和女性人名用字的吉罗指数：

	1919-1938	1939-1958	1959-1978	1979-1998	1999至今
男	15.03	12.98	11.81	14.58	10.34
女	14.95	12.82	11.73	13.60	12.00

Table 4: 两性人名用字丰富度的吉罗指数

从表4中可以看到，从1919年至今，人名用字的丰富度总体呈现出降低的趋势，也就是人名用字逐渐单一化，进入人名的汉字越来越集中。人名作为专有名词的一种，其专有性和独特性越来越突出。同时对比两性人名用字的吉罗指数发现，1998年以前，男性人名用字比女性人名用字丰富。

4.4 百年人名高频字及用字性别偏度

人名与所处的时代、相关政治历时事件、社会经济等有密切关系，不同年代的人取名有一定的特点。我们对近百年来两性人名中出现的汉字进行统计，取高频字的前15个字。两性人名用字的总体差异，见表5。在这些高频字中只有“华”字是重叠的，其在现代汉语词典 (2016)中的解释主要有“中国、繁华、精英、美丽”等，这些代表了中国人名最常包含的意义。

男高频字	明、文、国、华、建、志、军、伟、德、平、海、林、东、成、永
女高频字	丽、英、红、晓、华、芳、玲、玉、兰、梅、小、文、秀、萍、慧

Table 5: 百年来两性人名高频字

在我们感性认知中，男女人名在用字上应该有所不同，有些字在男性名字中常用，有些则在女性名字中常用。但是哪些汉字具有明显的倾向，以及汉字本身具有怎样的特征呢？本文设计了人名用字的性别偏度这一指标，从定量上对人名与性别的关联进行评估考察，其公式为：

$$V = \frac{P_{male} - P_{female}}{P_{total}}$$

对于V大于0的汉字我们认为其偏向于男性用字，小于0的汉字我们认为其偏向于女性用字，最终得到偏男汉字652个，偏女汉字432个。当V=0时，代表该汉字在男女中的分布均衡，我们称之为中性字。当V=±1.00时，则代表该汉字仅出现在一种性别之中，我们称之为极男/女字。取总字频为前1000的所有汉字中的性别偏度，最终得出百年人名性别极性字表(表5)。

极男（29）	栋、彪、乾、腾、庚、敦、涌、朋、营、干、创、钊、挺、甲、关、典、财、录、仰、猛、巨、罡、谋、专、炯、熠、纲、炯、熠
极女（61）	琴、婉、蕾、娣、妹、茜、婧、瑛、娅、筠、妙、莺、婕、苹、媚、蕙、妤、姿、姗、姝、荷、婵、女、黛、玛、翎、嫣、苑、涓、妃、甜、蔓、笛、珈、鸽、菱、璧、颐、舞、瑰、函、姐、鹃、纳、蜜、蜀、箐、霭、瑗、玟、拉、涟、漪、欧、嫦、绣、飘、俏、菡、蘅、薰
中性（28）	乐、庭、桐、淼、又、祯、阿、李、陆、汶、地、季、闻、贻、翼、铮、尘、薪、至、层、呈、古、隽、临、珑、施、晏、尹

Table 6: 百年高频人名用字性别极性字表

可以发现在高频字中，极女字较极男字更多，也就是在取名系统中女性人名的专用字较多。我们对表5中比较集中的几个意象作了简要归纳，得到表6：

类别	字义/意象	字例
偏男	具有较强动作性 凶猛、巨大等意	腾、涌、营、干、创、挺、仰 彪、猛、巨
偏女	描写女性姿态气质 女性称谓 含有娇小、美丽等意象	婉、媚、姿 娣、妹、姐 蕾、莺、茜、婧
中性	姓氏用字	李、陆、季、古、施、尹

Table 7: 两性用字意象归纳

4.5 人名用字历时变化趋势

我们分性别将所有年份的用字进行统计，得出不同年份的高频字表，表8、9:

	男高频字	女高频字
1919-1938	德、华、良、文、明、振、民、元 成、家、国、昌、一、兴、荣	英、华、兰、玉、珍、芳、芬、淑 梅、琴、文、丽、凤、秀、桂
1939-1958	国、明、文、德、生、林、华、光 建、正、平、学、新、海、祥	英、兰、华、玉、玲、淑、芳、丽 秀、小、美、萍、敏、凤、珍
1959-1978	明、文、军、国、华、建、平、志 东、永、伟、海、林、春、新	红、丽、英、晓、梅、玲、华、萍 芳、秀、霞、玉、慧、燕、兰
1978-1998	龙、俊、伟、文、子、鹏、明、杰 晓、东、小、宇、志、海、天	佳、婷、晓、丽、子、雨、文、娜 小、思、琳、雅、嘉、梦、一
1999至今	嘉、俊、宇、泽、子、博、涵、浩 杰、天、轩、艺、柏、晨、海	佳、子、儿、思、怡、雅、一、涵 琪、倩、诗、彤、雯、馨、钰

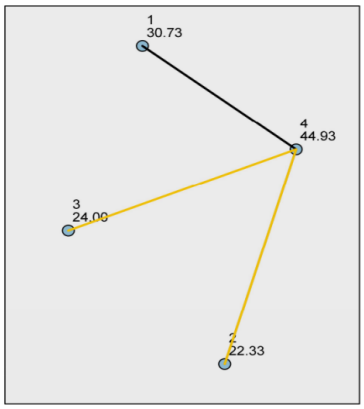
Table 8: 自然年份中两性人名高频字分布

	男高频字	女高频字
1919-1949	德、文、华、国、明、昌、成、林 良、正、家、元、生、民、学	英、华、兰、玉、芳、淑、珍、丽 秀、梅、芬、凤、桂、美、文
1950-1965	明、国、建、华、文、平、林、德 生、志、光、军、新、荣、伟	英、丽、玲、华、萍、兰、秀、晓 芳、梅、小、红、玉、霞、凤
1966-1978	文、军、明、国、东、华、志、海 建、伟、平、峰、春、永、成	红、丽、晓、英、梅、玲、芳、慧 霞、华、玉、艳、燕、小、秀
1979至今	俊、龙、伟、文、子、鹏、明、杰 宇、晓、小、东、海、天、志	佳、婷、子、晓、丽、雨、文、思 娜、小、雅、琳、一、嘉、怡

Table 9: 重大历史事件中两性人名高频字分布

按照总体高频字表中每个字对应的排序，对历时层面的前15个高频字进行编码。前人的研究认为，人名用字与历史事件有关，因此在划分人名阶段的时候采用了重大历史事件作为节点。所以我们首先选择表9的内容，采用独立样本Kruskal-Wallis检验。检验结果表明男性用字的分布在不同阶段上不具有统计学意义上的差别，而女性用字具有统计学意义上的差异。具体统计检验结果见图5。

因为女性用字分布存在差异，所以进一步对女性用字做了成对比较，探究不同阶段之间的差异。如图6所示，图中不同时段用节点表示（每个节点显示样本的平均秩），彼此之前的关系



样本1-样本2	检验统计	标准错误	标准 检验统计	显著性	调整显著性
2-3	-1.667	6.373	-.262	.794	1.000
2-1	8.400	6.373	1.318	.187	1.000
2-4	-22.600	6.373	-3.546	.000	.002
3-1	6.733	6.373	1.057	.291	1.000
3-4	-20.933	6.373	-3.285	.001	.006
1-4	-14.200	6.373	-2.228	.026	.155

每行会检验零假设：样本 1 和样本 2 分布相同。
显示渐进显著性（双侧检验）。显著性水平为 .05。
Bonferroni 校正已针对多个检验调整显著性值。

Figure 6: 女性用字分布与年代的成对比较

用实线连接，其中黄色的实线代表具有显著差异性。可以看到，女性用字差异主要来源于第四个阶段，即改革开放以后。这说明女性用字在改革开放后发生了显著变化。

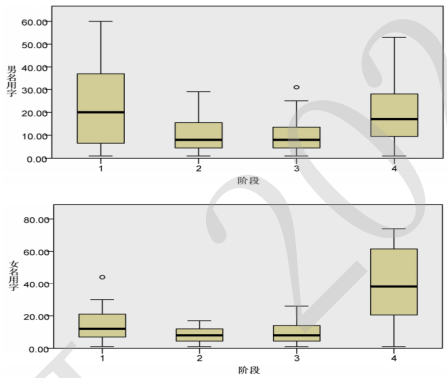


Figure 5: 人名用字的Kruskal-Wallis检验，检验总计量为60。男性人名均值6.704，双侧渐进显著性为0.082，女性人名均值15.627，双侧渐进显著性为0.001。

鉴于差异主要分为两个大的阶段，即1979年前后（改革开放前后），因此我们对这一时期前后两性人名高频字变化的主要特点做出归纳，见表10:

性别	时期	字义/意象	例字
男	1979年以前	品德与志向；	德、正、伟、荣、志；
		代表时政；	国、军、红、建；
	1979年以后	表外貌；	俊；
		取动物中宏大高远的有关意象；	鹏、龙；
女	1979年以前	品德与美貌；	丽、美、淑；
		取植物、美玉中美好的有关意象；	兰、玉、梅；
	1979年以后	译名常用字；	娜、琳、菲；
		表梦幻、优雅等意象。	雨、诗、怡、雅；

Table 10: 改革开放前后高频字主要变化

从表中可以看到，1979年以前男性人名部分与时政有关联，两性人名对于政治历史事件的敏感度是不一样的，相比之下男性人名与政治历史事件联系更加紧密，从不同阶段的男性高频字可以看出时代特点。在新中国成立前强调传统“修身治国平天下”的理想，男性高频用字跟多

反映个人品德修养。表8中可以看到在第一阶段“德”“良”等字都在前15个高频字中。在第二阶段男性人名所反映出的伟大志向、建设祖国的心愿，如“建”“志”等。第三阶段则反映出革命与建设的潮流，如“军”“伟”。男性人名用字分布虽然总体没有显著差异，但每个阶段都随着时代特征而体现出不同的侧重，高频字的排序随之发生变化。

不过变化是缓慢且滞后的，可以通过两种时间划分方式的重叠部分来推测不同时间段内部的变化，以“德”“建”二字为例，从“德”字在表8,9中的排序变化可以看出该字在新中国成立前使用非常频繁，新中国成立后起使用频率并没有迅速降低，甚至在早期使用依然较多，随着时间的发展逐渐变得不那么常用。类似地，1950-1958年段中“建”的使用不如1959-1965年多。与其他历时阶段比，“建”字在建国初期呈现小高峰，反映了特定的时代特征。而在这一阶段内部再比较，该字的使用又呈现出了逐渐上升的趋势。因此人名的变化虽然受政治事件的影响但是其变化是具有渐进性和滞后性。

为了更清晰地显示人名高频用字的变化，我们分性别总结了各个阶段排名均在50的字，取前10个，归稳定且常用；取第四个阶段在前50，并且呈现上升趋势的前20个字归为上升快且当前常用(Max(Rank1,Rank2,Rank3)-Rank4)；取前三个阶段的任意阶段曾在前50，但在第四个阶段下降的前20个字归为下降快且曾经常(Rank4-Min(Rank1,Rank2,Rank3))。最终得到表11。

	男	女
稳定且常用	成、德、国、海、华、林、明、平文、祥、永、志	芳、慧、娟、君、丽、玲、美、敏文、小、晓、雪、燕、玉
上升快且当前常用	磊、博、锋、洋、豪、佳、君、军超、嘉、波、勇、宇、强、飞、小阳、涛、峰、江	儿、涵、雨、婷、菲、诗、妍、艺萱、宇、梦、妮、倩、心、欣、怡嘉、丹、颖、佳
下降快且曾经常用	武、贵、孝、全、万、山、学、民克、刚、昌、绍、仁、荣、宗、胜景、鸿、良、世	生、荣、桂、素、菊、建、德、莲平、芬、珍、瑞、淑、芝、琼、秀世、利、珠、爱

Table 11: 用字变化趋势表

5 结语

本文构建了一个中国名人人名数据库，条目共11万+，每个条目含有人名、性别、出生地等社会文化标签，同时含有拼音、笔画、偏旁等文字信息标签。该语料库可以支持对人名的地域、历时、性别等多个维度的研究。

在人名数据库的支持下，本文选取1919至今的人名作为研究对象，从人名长度、用字难易度、丰富度等角度进行探究。研究发现男性人名比女性长，但两性人名长度的差异随着时间而不断缩小。建国以来单名比例不断增加，但是进入21世纪又逐渐减小。单名用字比双名难，女性人名用字比男性难，男性用字难易程度波动较大。人名中的二级常用字最为稳定，其次是一级常用字。在用字的丰富度上，随着时间的发展人名用字越来越体现出其专有性的特征，丰富度逐渐降低。男性人名的用字总体上比女性用字丰富。通过计算人名的性别偏度指标后发现女性人名专用字更多。改革开放对人名用字格局产生了重要影响，女性用字的变化显著。男性人名与时政联系更加紧密，用字的变化虽然受时政的影响，但其变化具有渐进性和滞后性。

这些发现可以帮助我们进一步了解人名的发展变化规律，探究汉字中的的性别差异。当然本文还存在一些缺陷，例如在自然年份中1919年至今这一时段的人名较少，对数据的分析产生一定影响。下一步，我们将持续补充新增名人人名数据及相关信息，并从偏旁、地域等维度进行深入研究。

致谢

感谢论文辅导老师对本论文的帮助，感谢各位匿名评审老师的修改建议。本论文受教育部人文社会科学研究规划基金资助项目（18YJA740030）和北京语言大学研究生创新基金项目（20YCX155）资助。

参考文献

Jizheng Jia and Qiyang Zhao. 2019. Gender prediction based on chinese name. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 676–683. Springer.

Elke Teich. 2012. *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*, volume 5. Walter de Gruyter.

Guoxing Yu. 2010. Lexical diversity in writing and speaking task performances. *Applied linguistics*, 31(2):236–259.

何晓明. 2001. 姓名与中国文化/中国文化新论丛书. 人民出版社.

吴继章. 2001. 也谈人名中的异体字. 语文建设, 8.

张书岩. 1999. 从人名看50 年的变迁. 语文建设, 4.

张书岩. 2004. 姓名· 汉字· 规范. 北京广播学院出版社.

施建刚and 邵斌. 2016. 基于语料库的汉语译文翻译共性研究——以《苏东坡传》汉译本为例. 外国语言文学, 33(2):97r104.

潘世松. 2004. 汉字结构的性别歧视倾向论析. 求索, 12:212–214.

王玉新. 2000. 汉字认知究. 山东大学出版社.

现代汉语词典. 2016. 第七版. 中国社会科学院语言研究所词典编辑室编. 北京: 商务印书馆.

苏培成. 2001. 谈人名中的异体字. 语文建设, 5.

谢玉娥. 2000. 人名、性别、文化——对男人名,女人名文化现象的考察. 中国文化研究, (1):103–108.

赵越. 2006. 汉人韵律情结, 命名文化与aba (a') 命名方式. 语文学刊, (13):26.

邱莉芹and 鞠泓. 2002. 人名用字中使用生僻字情况的调查与分析. Ph.D. thesis.

韩燕, 邱江, and 张庆林. 2008. 性别刻板化人名推测判断中的冲突效应. Ph.D. thesis.