







ARTICLE




<https://doi.org/10.1057/s41599-021-00799-6>

OPEN

# Occupational gender segregation and gendered language in a language without gender: trends, variations, implications for social development in China

Qi Su<sup>1</sup>, Pengyuan Liu<sup>2</sup>, Wei Wei<sup>1</sup>, Shucheng Zhu<sup>2</sup> & Chu-Ren Huang<sup>3</sup>

This paper proposes a textual analytics approach to the discovery of trends and variations in social development. Specifically, we have designed a linguistic index that measures the marked usage of gendered modifiers in the Chinese language; this predicts the degree of occupational gender segregation by identifying the unbalanced distribution of males and females across occupations. The effectiveness of the linguistic index in modelling occupational gender segregation was confirmed through survey responses from 244 participants, covering 63 occupations listed in the Holland Occupational Codes. The index was then applied to explore the trends and variations of gender equality in occupation, drawing on an extensive digital collection of materials published by the largest newspaper group in China for both longitudinal (from 1946 to 2018) and synchronic (from 31 provincial-level administrative divisions) data. This quantitative study shows that (1) the use of gendered language has weakened over time, indicating a decline in occupational gender stereotyping; (2) conservative genres have shown higher degrees of gendered language use; (3) culturally conservative, demographically stable, or geographically remote regions have higher degrees of gendered language use. These findings are discussed with consideration of historical, cultural, social, psychological, and geographical factors. While the existing literature on gendered language has been an important and useful tool for reading a text in the context of digital humanities, an innovative textual analytics approach, as shown in this paper, can prove to be a crucial indicator of historical trends and variations in social development.

<sup>1</sup>Peking University, Beijing, China. <sup>2</sup>Beijing Language and Culture University, Beijing, China. <sup>3</sup>The Hong Kong Polytechnic University, Hongkong, China.  
email: [sukia@pku.edu.cn](mailto:sukia@pku.edu.cn); [liupengyuan@pku.edu.cn](mailto:liupengyuan@pku.edu.cn); [churen.huang@polyu.edu.hk](mailto:churen.huang@polyu.edu.hk)

## Introduction

Occupational gender segregation refers to the tendency of men and women to work in different professions (Blackburn et al., 2002). It is an important demographic issue for industrialised societies (Watts, 1998) and a key source of social inequality (Charles and Grusky, 2004). Previous studies have attempted to attribute this inequality to aptitudes differences between the sexes. This can be seen in (Baker and Cornelson, 2018) which claims that males' higher tolerance for noise is the reason for their disproportionately higher levels of employment in noisy occupations. However, such simplistic rationale have been viewed as inadequate by many scholars. Recent studies have focused on relevant socioeconomic, institutional, educational, and cultural barriers to employment for females, such as gender-specific socialisation, the delegation of family responsibilities, and discrimination acts in regulations or at the workplace (Cuttillo and Centra, 2017; Sanchez et al., 2017; Stoet and Geary, 2018; Verniers and Vala, 2018). Furthermore, as researched by Garnham et al. (2015), there is compelling evidence to indicate that a strong correlation between perceived occupational gender distributions and the actual ratios might exist. The perception, or assumption, of occupational gender distributions leads to biased associations, or in other words, an 'automatic stereotyping' (Banaji and Hardin, 1996). This stereotyping can be operated through a spontaneous reflection of cognitive expectancy (Kunda and Oleson, 1997; Yzerbyt et al., 1999), the adaptation to motivational and interpersonal communication goals, or the consideration of interpersonal context and interaction between individuals (Beukeboom, 2014). In other words, the 'women's work-men's work' phenomenon is perpetuated by gender typing in language, which consequently affects the occupational aspirations of both men and women, their entry into an occupation, and their treatment and advancement within an organisation (Reskin and Hartmann, 1986; Ruble, 1983).

The reliance on linguistic gender for the study of gender typing faces two potential challenges. First, as linguistic genders are grammaticalized, it is difficult to ascertain if and to what extent the established convention leads to conceptual gender typing or vice versa. In other words, this research topic faces the challenge of Galton's problem (e.g. Naroll, 1961; Roberts and Winters, 2013) in not being able to rule out the conventionalised dependencies from the environmental dependencies that the study aims to prove. Second, it is not clear what linguistic features are effective cues for a language, such as Chinese, without grammatical genders. To solve Galton's problem, we follow the approach of (Huang et al., 2021). That is, we recognise that each important issue in the humanities and social sciences must be viewed in, and cannot be disassociated from, its sociocultural context. Hence, instead of trying, in vain, to disassociate the sociocultural contexts and render the issues and answers irrelevant, we conduct several studies in different sociocultural contexts. When the same dependency is attested in different sociocultural contexts, we can then prove the dependency confidently, regardless of the context.

To address linguistic cues in a language without linguistic gender, we rely on the observation of (Blum, 1997) that the naming has a very special social function in Chinese. Blum argued that naming terms are used to establish and maintain social hierarchy as well as to 'give face'. As a consequence, pronouns are not typically used to refer to people in Chinese. Instead, people are usually (and often socially required to be) referred to by a combination of their occupation or social role and their proper name. Given that Chinese is a genderless language, this sociolinguistic construct of naming offers a set of direct, and potentially richer, data for gendered language use. Based on the concept of markedness, gender typing can be attested when the use of gender modifiers are not balanced between two genders.

The overuse of overt marking of a particular gender before an occupational term suggests that this gender is marked and that the occupation is gender-typed to be the opposite sex. Namely, an occupation stereotyped to be male-dominant will require an explicit and marked feminine modifier, such as the character 女 'nǚ' (female) (see Farris, 1988). For instance, the expression 女总统 'nǚ zǒngtǒng' (female president) is common in Chinese, while 男总统 'nán zǒngtǒng' (male president) was not attested. Conversely, 男保姆 'nán mǎomǔ' (male nanny) is well attested but 女保姆 'nǚ mǎomǔ' (female nanny) is rare, indicating the gender typing of nannies as female. Gender marking may also imply specific cultural biases, such as 女司机 'nǚ sījī' (female driver) which concretises the perception of a female driver as less competent than a male driver. Another example is 女博士 'nǚ bóshì' (female doctor), which connotes an image of leftover women with marriage anxiety under the traditional discourse of marriage as a life goal for women. In Chinese, these marked uses of gender modifiers (especially for females) can be considered as another instantiation of 'masculine generics' (Stahlberg et al., 2007).

It should be noted that the usage of such marked gender expressions is so pervasive that it has led to an ongoing debate among scholars. Some scholars argue that the usage of feminine forms makes women more visible, while others argue that it goes against the advocacy of gender-fair language (GFL) and that the use of such terms 'undervalues female versions' by implying that terms without gender marking are male (Sendén et al., 2016).

Lastly, regional variations are another important factor underlining the issue of occupational gender segregation and gendered language. Eckert and McConnell-Ginet (1992) showed that language and gender practices are community-based, and hence community-to-community variations are to be expected. Although such variations are commonly observed in demographic studies and can be clearly seen when comparing national data, little systematic text-based analysis has been conducted to show the regional variations of gender typing.

In this paper, we adopt a textual data-driven approach to study gender typing in a language without gender. Grounded in the theory of markedness in linguistics, we interpret our data as follows: if a feminine modifier is used before an occupational term, the occupation is stereotyped to be more appropriate for male (or male-oriented); if a masculine modifier is used, the occupation is stereotyped to be more appropriate for female (or female-oriented). The level of gender markedness of occupational terms will be calculated with a proposed linguistic index (see Eq. (3)). A quantitative analysis was conducted to measure the gender markedness of 63 occupations listed in the Holland Code, using the linguistic index on a multi-domain corpora. To overcome the Galton problem, the results of the quantitative analysis have been confirmed through a correlation analysis with the results of a questionnaire, administered to 244 participants, measuring gender stereotyping of the listed occupations. Then, using Eq. (3), we have explored the synchronic (regional and generic) variations of occupational gender markedness and trends of social change, drawing from an extensive digital collection of national-level and provincial-level newspapers published in China from 1946 to 2018. The coverage of a wide range of diachronic data as well as geographical and cultural variations is another design feature to overcome the Galton problem and also to tackle the unaddressed issue of regional variations.

## Related work

At the lexical level, gender typing can be reflected in the marked use of occupational terms, as the grammatical gender assigned to

the nouns. Prewitt-Freilino et al. (2012), among others, investigated the relationship between the gendering of languages and gender qualities by looking into the linguistic genders of occupational terms across 111 countries. Liben et al. (2002) showed that children interpret occupational terms as gender-embedded. In addition to grammatical gender, gender typing can also be achieved through the marked use of gender terms combined with occupational titles.

From a cross-linguistic perspective, researchers have looked into the correlation between occupational gender markedness in language and stereotyping in psychology through sentence evaluation paradigms (Gygax et al., 2008), eye-tracking studies (Irmen and Rofsberg, 2004), and role noun priming (Cacciari and Padovani, 2007). The level of gender inequality is revealed to be closely linked to the gender markedness of that language (Garnham et al., 2016; Pacheco, 2018; Prewitt-Freilino et al., 2012). Horvath et al. (2016) examined the implication of the masculine form, or feminine–masculine pairs, in the gender-typing of occupations. They found that word pairs are conducive to avoiding a male bias and therefore make women more visible but also lower salary estimates of typically feminine professions. Gaucher et al. (2011) showed that gendered wording is found in job advertisements, especially in male-dominated areas, which contributes to the perpetuation of gender divisions.

Due to the close connection between language markedness, bias, and actual distribution, corpus data proves to be an important source for investigating the changes and tensions regarding gender stereotyping and social change. This can be seen by the growing presentation of females and the efforts being made to promote gender-fair language (Macalister, 2011; Norberg, 2016), while at the same time there is still the existence of gender bias with male terms as the pseudo-generics and norms (notably higher frequencies) (Holmes, 2001; Moser and Masterson, 2014). Baker's English-based collection presents the diverse corpus research on analysing gender-related linguistic issues such as the comparison of male and female speeches in the British National Corpus (BNC), the particular form of expressing disagreement for female scholars, the frequencies of gendered pairs such as man/woman and wife/husband, and the collocation patterns for 'boy' and 'girl' (Baker, 2014).

In sociology, there are a number of indices for the measurement of gender segregation. One of the most frequently used measures is the 'Index of Dissimilarity' ( $D$ ) proposed by Duncan and Duncan (1955). A possible formula for calculating  $D$  is presented in Eq. (1).  $M$  represents the total number of males in employment;  $M_i$  is the number of males in occupation  $i$ ;  $F$  is the total number of females in employment;  $F_i$  is the number of females in occupation (Blackburn et al., 1993; Emerek et al., 2002). This index has been widely used in the analysis of occupational gender segregation (Gross, 1968). However, the Index of Dissimilarity is not without problems. As Preston (1999) highlighted, the change in the size of an occupation with gender imbalances will affect the index even when the gender composition of all occupations remains the same. Other measurements have also been proposed such as the Moir and Selby–Smith segregation indicator (MSS), also called WE Index (Moir and Smith, 1979), Karmel and MacLachlan Index (IP) (Watts, 1995), and the Index of Segregation that is calculated according to the method of marginal matching (Blackburn et al., 1993).

$$D = \frac{1}{2} \sum_{i=1}^n |M_i/M - F_i/F| \times 100 \quad (1)$$

Sociological methods capture the statistical distribution of occupational segregation across nations and time. However, one important element that the above indices do not incorporate is the differentiation between *gender* and *sex*, which this study

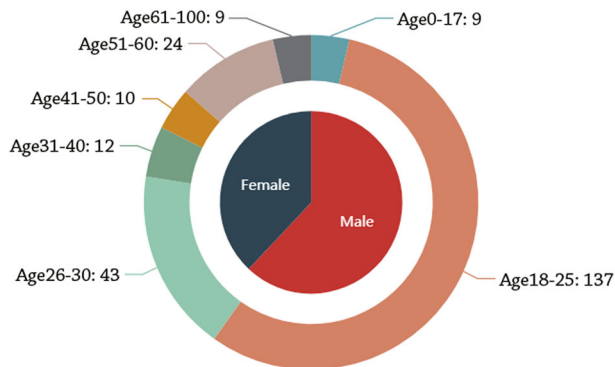
intends to address. Sociological measurements evaluate *gender* segregation by calculating the disproportionate distribution of males and females, and the demographic, physical, and binary divisions that are in fact more related to *sex*, a term that emphasises the biological dimensions of being man or woman. The notion of *gender* is related to the social construction of 'what it means to behave like a man or woman' (Parrillo, 2008), which is closely connected to the cultural beliefs, assumptions, and expectations concerning the social roles of men and women. Such conceptual considerations are another important aspect of occupational gender segregation, which concerns gender typing and gendered language use. While there are both qualitative and quantitative studies employing questionnaires or experiments in researching this issue, a data analytics model, or measure, is still needed; this might enable researchers to investigate the issue on a larger scale and facilitate a compare and contrast of results from different studies in revealing the potential historical, regional, cultural, and linguistic differences.

Studies in digital humanities, especially those embracing a data analytics approach, are also highly relevant for the current research. Berry's two waves of digital humanities describe the advancement as transforming from providing 'a technical support' to conventional humanities research to interacting with texts that are 'born digital' and their associated digital cultures (Berry, 2012). Digital techniques such as stylometry, topic modelling, literary mapping, historical bibliometrics, corpus linguistic techniques, and sequence alignment have facilitated analyses of huge quantities of texts and even multimodal data. Yet, debates on the role and purpose of digital humanities persist. We believe that the answer can be found following Wilhelm Dilthey's 'human sciences' (Makkreel, 2016; Makkreel and Rodi, 1990). That is, as a discipline, the humanities is defined by issues that it is concerned with rather than by certain (discipline-specific) research tools. Given the complexity of all the issues the humanities is concerned with, it is necessary to integrate all available tools and knowledge from various disciplines and to form synergetic views, with a humanistic perspective and some speculations if needed. This is precisely the approach taken in a series of studies researching the interaction of meteorological events, people, and their languages (Huang et al., 2021). Similar approaches have been taken by various recent studies, such as the use of Google N-gram to study social developments and cultural/linguistic variations by Juola (2013) and Li et al. (2020); the application of a linked data approach to cultural heritage material for the discovery of gender properties of (in particular, female) jazz artists by Pattuelli et al. (2017); and the machine learning-based analysis of large-scale newspaper data to uncover inconsistent coverage of female and male politicians as evidence of social gender bias (Leavy, 2018). In line with this vision, we developed an innovative approach that is both longitudinal and cross-sectional, covering features of historical trends, and generic and regional variations based on large-scale diversified corpora.

## Methodology

We adopted the Holland Codes, with a total of 63 categories of occupations (Holland, 1959), in the questionnaire and the later quantitative analysis<sup>1</sup>. In Holland's theory of career interests, six personality types were identified, usually referred to by their initials as RIASEC, the descriptions of which are shown below:

- Realistic (R)—practical, physical, hands-on, mechanically inclined
- Investigative (I)—analytical, intellectual, scientific, explorative
- Artistic (A)—creative, original, independent, expressive
- Social (S)—cooperative, supporting, helping, healing/nurturing



**Fig. 1 The age and gender distributions of the participants.** The centre circle indicates the distribution of male and female participants. The outer ring indicates the distribution of different age groups.

- Enterprising (E)—competitive environments, leadership, persuading, administer over
- Conventional (C)—detail-oriented, organising, clerical, structured

Grounded in the RIASEC model, the Holland Codes, or the Holland Occupational Themes, were developed for the classification of occupations. Holland’s typology and the subsequently developed career interest assessments and inventories have dominated the field of career planning and counselling throughout the past decades. Notably, Holland’s Self-Directed Search (SDS) has been translated into more than 30 languages (Bullock et al., 2009). The Chinese version of SDS has been widely used across higher education institutions in mainland China as the major tool for examination of students’ occupational choices since it was first translated and introduced in the 1990s (Long and Peng, 2000; Long et al., 1996). Its dominant status in occupation planning provides the rationale for choosing the Holland Codes for the classification of occupations in the present study.

The possible gender stereotyping of the 63 occupations was investigated through a questionnaire. As the questionnaire was primarily intended to measure the matching between the gendered perception of occupations and the gendered use of occupational terms, rather than functioning a wide-ranging social survey, the convenience sampling approach was used for the selection of the participants. The participants consisted of 244 people (151 males, 93 females), with a balance of different age and gender groups. The age and gender distributions of the participants are presented in Fig. 1.

The questionnaire was designed based on Likert scale 5-point response anchors. The participants’ task was to decide to what extent each occupation is more appropriate for males (1 point) or females (5 point), with 3 in the middle. The scores of Cronbach’s alpha tests were >0.9, indicating the high reliability of the questionnaire. Table 1 presents the results of different participant groups<sup>2</sup>. Overall, no significant difference was shown among the participant groups in their gender associations for the different occupations ( $M = 3.15$ ,  $SD = 0.89$ ). Nevertheless, the number of occupations that are considered to be ‘male jobs’ are slightly more than ‘female jobs’. The only major difference between male and female participants exists in the four artistic occupations, namely, *artist*, *cartoonist*, *painter*, and *writer*. Specifically, the male group identified men as more suitable for these four occupations, whereas the female group considered women to be more suitable. Within the different age groups, the gendered orientation is slightly higher among participants over 30 years of age, compared to those aged under 30.

Grounded in the aforementioned gender markedness theory, with the use of a masculine/feminine modifier before each

**Table 1 Comparisons of different participant groups in gender-occupation perception.**

	<i>M</i>	<i>SD</i>	<i>SEM</i>	<i>t</i> value	<i>p</i> (two tails)
Male	3.17	1.00	0.04	0.96	0.34
Female	3.14	0.82	0.02		
Age > 30	3.20	0.96	0.03	-1.41	0.16
Age ≤ 30	3.14	0.87	0.02		
Total	3.15	0.89	0.02		

occupational term as the key indicator, we proposed a novel equation to measure the degree of gender segregation (DGS).  $TF_m$  in Eq. (2) is the term frequency of an occupation with a masculine mark (e.g. 男医生 ‘nán yīshēng’ male doctor) and  $TF_f$  is the term frequency of an occupation with a feminine mark (e.g. 女医生 ‘nǚ yīshēng’ female doctor).

$$DGS = k \times \frac{\alpha TF_f^{\nu}}{TF_m + TF_f} + b \tag{2}$$

In this generalised form of DGS,  $k$  and  $b$  are parameters for scaling. To enable a comparison between the questionnaire results and the quantitative analysis on gendered occupation, i.e., results calculated with DGS,  $k$  and  $b$  were set to 4 and 1, respectively, conforming to the 5-point Likert scale in the current study.  $\alpha$  and  $\nu$  represent the external scale factors that contribute to the gendered use of occupational terms, such as the overall gender distribution in a society and the vertical segregation within occupations or sectors (Blackburn et al., 2001); these should be separately identified and measured, and were set as 1 in this study. With this parameter setting, it was anticipated that the index would serve as a measure of the horizontal dimensions of occupational segregation by identifying the extent of gender-dominance of occupations (GO). Therefore, the index can be further represented as

$$GO = 4 \times \frac{TF_f}{TF_m + TF_f} + 1 \tag{3}$$

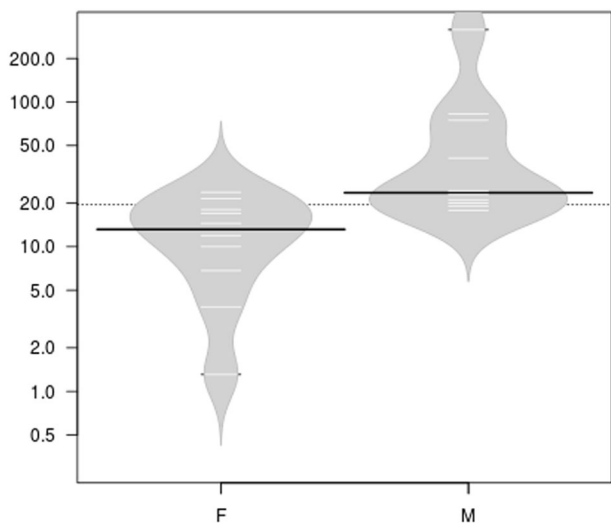
We calculated the GO value of the multi-domain subcorpora of the BCC Corpus<sup>3</sup>. A Pearson correlation analysis (see Table 2) showed a significant positive correlation between the questionnaire results (cognitive gender segregation) and the quantitative gender segregation calculated from Eq. (3), indicating that the markedness calculated by the GO value can act as an effective quantitative measure of conceptual occupation segregation by gender reflected in language use.

To understand the distribution of GO values upon different occupations, the GO index for each occupation was further evaluated. The index shows the proportion of feminine markedness to the total gender markedness. Therefore, if the frequencies of masculine and feminine markedness were identical, the GO index would equal 3 (i.e., in the middle of [1, 5]). If an occupation was either male-dominant or female-dominant, the index would be greater or smaller than 3. However, it should be noted that almost all the GOs are larger than 3. This can be explained in two ways. As Greenberg (1966) pointed out, markedness is not an absolute binary but a scalar concept from most marked to least marked. Alternatively, when considering a language moulded by its long history of patriarchal hegemony, the over-representation of the feminine-marked ‘female job’ can also serve to stereotype the vertical segregation. For example, ‘female nurse’ stereotypes the position of female workers as a position of lesser value within the medical profession. This ideological connotation is shaped by the external factors denoted by the parameter  $\alpha$  and  $\nu$  in Eq. (2). Nevertheless, GO can still show variation tendency of gender segregation with the simplified  $\alpha$  and  $\nu$ .

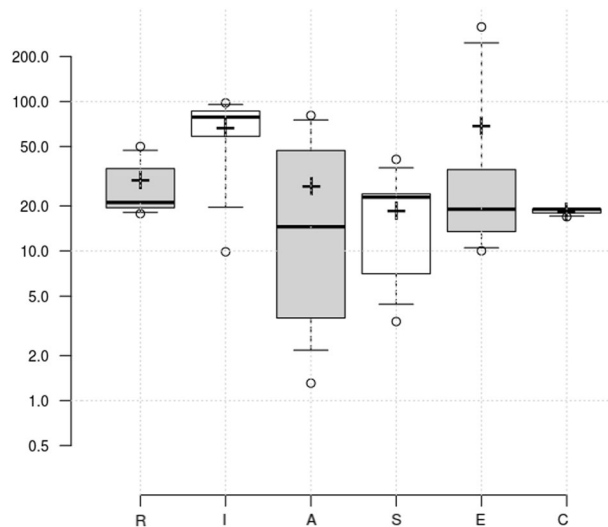
**Table 2 Pearson correlations between cognitive gender segregation and quantitative gender segregation.**

		Cognitive gender segregation	Quantitative gender segregation
Cognitive gender segregation	Pearson Correlation	1	0.312*
	Sig. (2-tailed)		0.013
	N	63	63
Quantitative gender segregation	Pearson Correlation	0.312*	1
	Sig. (2-tailed)	0.013	
	N	63	63

\*p < 0.05.



**Fig. 2 Relative segregation of female- and male-dominant occupations.** The left plot shows the  $TF_f/TF_m$  of the top 10 female-dominant occupations. The right plot shows the  $TF_f/TF_m$  of the top 10 male-dominant occupations.



**Fig. 3 Segregation on the six Holland categories.** The plots indicate the  $TF_f/TF_m$  calculated for each of the six Holland categories.

Figure 2 presents the relative segregation of occupations, calculated by  $TF_f/TF_m$  (the main part in Eq. (3) without scaling parameters). The left plot in the figure shows the  $TF_f/TF_m$  of the top 10 female-dominant occupations from our questionnaire. The right plot contains the results of the top 10 male-dominant occupations. Figure 3 describes the combined results for all six Holland categories. Researchers have claimed that gender differences in Holland personality types and vocational interests are visible (Su et al., 2009). For example, there are more males in Realistic and Investigative categories, and more women in Artistic, Social, and Conventional types. Evidenced in Figs. 2 and 3, the ratios of  $TF_f$  to  $TF_m$  generated by male-dominant occupations are higher than those of female-dominant occupations<sup>4</sup>. This result confirms the validity of the proposed equation.

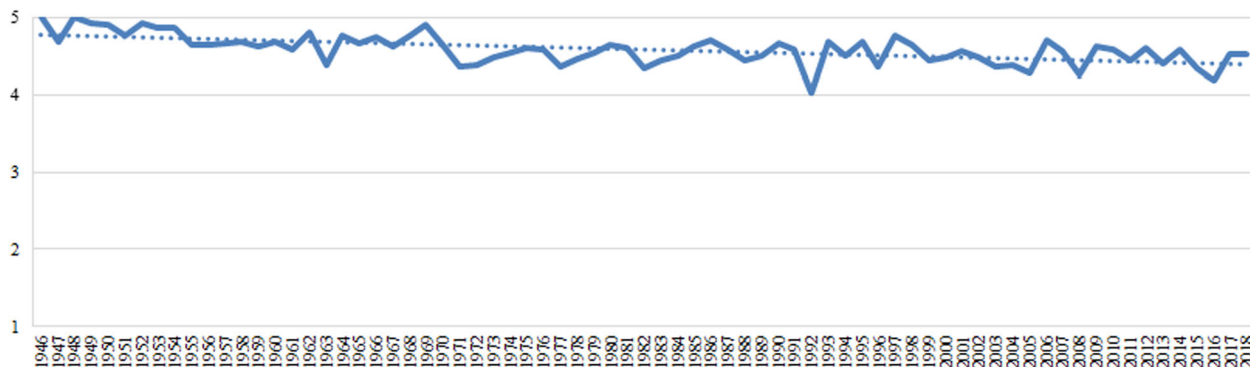
In what follows, we investigated the diachronic changes and generic as well as regional variations of occupational gender segregation based on the linguistic index. Our diachronic study covers 72 years (1946–2018), based on a corpus of *People’s Daily*, the largest and most read newspaper in China. For genre differences, we examined the four sub-corpora of the BCC Corpus: *literature*, *newspaper*, *microblog*, and *technology*. Our study of regional variations is based on newspapers published in 2018 from 31 provincial-level administrative divisions in China<sup>5</sup>. It should be noted that newspapers from Hong Kong, Macau, and Taiwan are not included because of their orthographic and lexical differences, as well as their attested variations in grammatical

features (Xu et al., 2020) and textual features (Hou and Huang, 2020), which hinders a direct comparison of degrees of gender segregation.

**Results**

**Historical trends.** The GO value for each year in the designated period was calculated based on Eq. (3). Figure 4 shows the diachronic distribution of occupational gender markedness over the past 70 years. From Fig. 4 it can be seen that the overall tendency of gender marking has slightly declined over time with an overall level of stability, though also exhibiting some year-to-year fluctuations. This result provides corpus evidence to support the finding of (Popp et al., 2003) that ‘gender stereotypes have weakened over time’. However, the fluctuations also reflect the instability and tensions of the pursuit of gender quality in profiling occupations. The changes seen in gender markedness reflect and might also be influenced by the social, economic, and discursive transformations that have taken place during the examined time period.

From the establishment of the People’s Republic of China in 1949 to near the beginning of the Cultural Revolution in 1966, as presented in Fig. 4, the level of markedness slightly decreased, which reflects the entrance of the females into the workplace. During that period, the state’s strong control on labour allocation drove women into the workforce. Most urban women had wage labour organised by the state, in factories or businesses, while rural women worked in the labour service of people’s communes (Wang, 1999). At that time, the ‘liberation’ discourse was

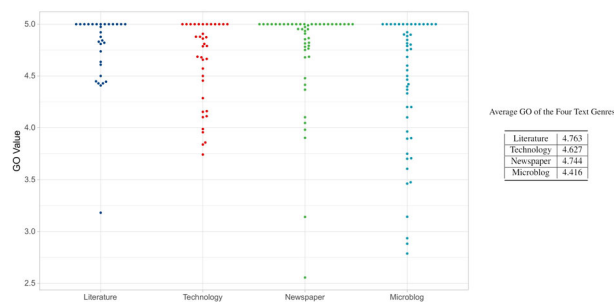


**Fig. 4 Diachronic distribution of occupational gender markedness in the corpus of *People's Daily* (1946–2018).** The vertical axis indicates the range of the GO values (between 1 and 5). The horizontal axis indicates the period of time (by year).

constructed to say that women's socialist liberation could only be achieved through participation in production, with the term 'housewife' becoming 'a scorned urban social category and increasingly a historic relic' (Zheng, 2003). The motto '妇女能顶半边天' (Women hold up half the sky) became widely popular. A significant rise of gender marking can be observed around the beginning of the Culture Evolution in 1966, which might be explained by the breakdown of state labour organisations and the economy.

Between the 1970s and 1980s, minor and slow upturns of gender marking were accompanied by relatively steep drops, especially after the reform and opening-up in 1978. The slim growth in gender marking might be related to the changes in economic organisation as well as social conceptions of the time. With the new labour market developing in China, and as the state was weakening its control on organising jobs, the gender bias in education and recruitment became more evident, widening the income gap and increasing gender segregation of occupations (Cooke, 2004). Moreover, passing through great social and economic disturbances, people overthrew the liberalisation discourse, considering the previous gender employment policies to be as 'a relic of Maoist egalitarianism' and an impediment to economic growth. Consequently, the discourse of and campaigns for 'women return home' and gender differentiation gained public support. However, in the early 1990s, feminist voices revitalised the ideology of female liberation and employment in the proposal raised by the Women's Federation (Zheng, 2003). Since then, policies have also been implemented to counter gender discrimination in the workforce. Notably, in 1990, *Regulations of Prohibited Types of Occupational Posts for Female Employees* was put forward, and in 1992, *The PRC Law on Protecting Women's Rights and Interests* was created. This time period shows the lowest point of gender markedness, as can be seen in Fig. 4, reflecting the drastic social transformations of occupations and gender and the strong influence of state policy at the time.

The sharp growth of the GO values after 1992 might be indicative of another crucial social change that occurred during the period with the restructuring of state-owned and collective-owned enterprises. As the state withdrew its power in labour allocation, massive layoffs occurred due to redundancies that were identified as a result of restructuring; gender inequality surfaced. After the mid-1990s, 62.8% of those laid-off workers were women (Zheng, 2003). Moreover, as revealed by a survey of over 200 women cadres in governmental organisations in Zhejiang Province in 1997, men and women differed considerably in their weighing of career and family life (Wang, 1999). Women at that time tended to prioritise family interests over their careers, demonstrating the persistence of the sociocultural expectations for gendered social roles.

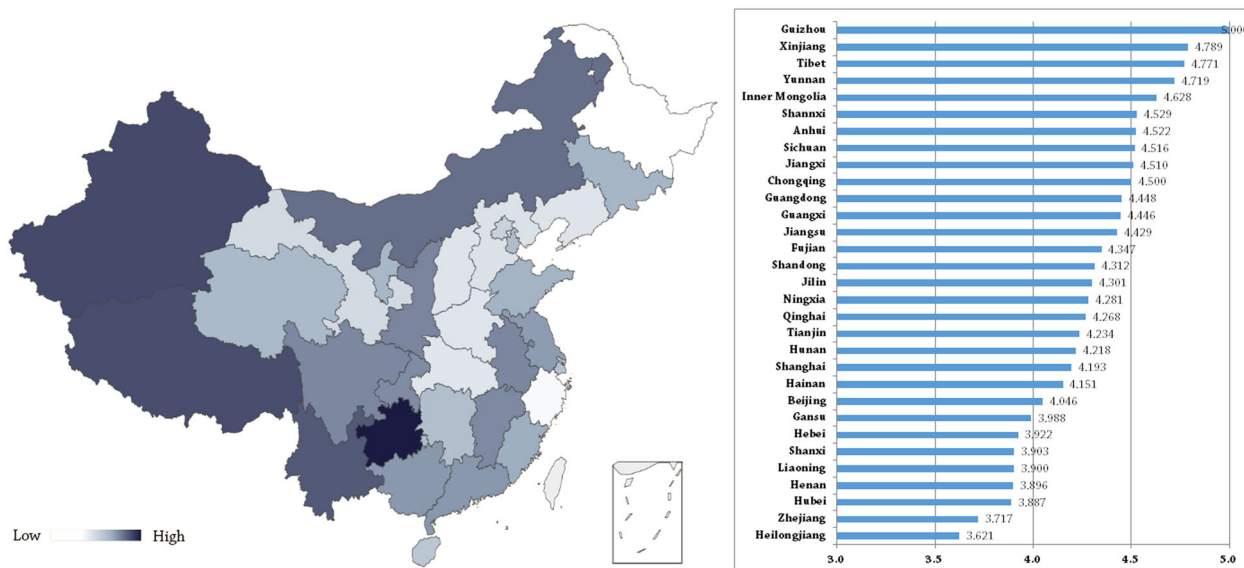


**Fig. 5 Distributions and means of occupational gender markedness for the four text genres.** The dots on the left display the distributions of GO values for the four genres. The table on the right displays the means of GO values.

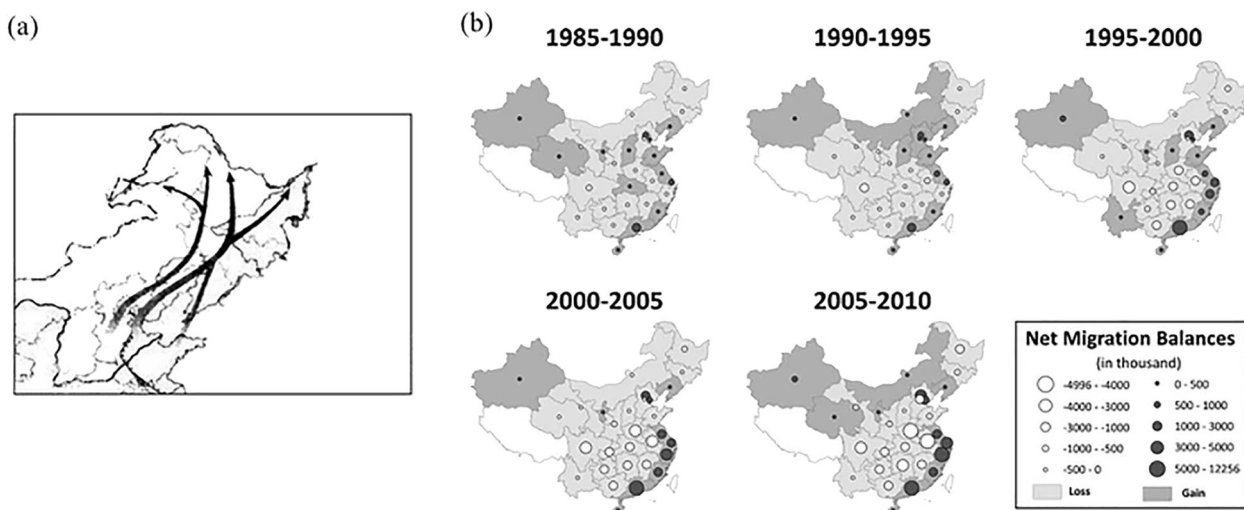
As the above analysis shows, the pursuit for gender equality in occupations was full of tensions. The process of social change in China is wrought with different push and pull forces through policy support, persisting gender discrimination, economic and political disturbances and development, the perceived social roles and family values, and changes in public discourses. When entering into language use, those changes might reflect on the profiling of occupations, and at the lexical level, the markedness of different professions.

**Genres.** Figure 5 shows both the distributions and means of occupational gender marking encoded in different genres. The genre of *microblog* has the lowest value of gender marking, *literature* the highest, and *technology* and *newspaper* in the middle. The differences in occupational gender markedness among the genres match the varied speed of change across genres from the perspective of linguistic conservatism. In linguistics, a conservative form or variety is one that has changed relatively little over its history, or which is relatively resistant to change. It has been noticed that written varieties change more slowly than spoken varieties (Fromkin et al., 2010). In other words, writing tends to be more conservative and stylistically fixed than speech.

Putting the four text genres in the 'oral-literate' continuum (Biber and Finegan, 2001), *microblog*, as the most colloquial among the four, shows the lowest level of markedness, most strongly reflecting the overall weakening trend of gender segregation. In contrast, *literature*, as one of most typically written forms, lags behind spoken genres in language development and shows the highest level of gender markedness. It is easy to understand that the language used in technical articles tends to be more innovative, when compared to other types of writing, as one linguistic feature of scientific writing is its technicality (Fang, 2005) which urges rapid language change in alignment



**Fig. 6 Geographic distribution of occupational gender markedness in China.** The figure shows the GO values for 31 regions, with Hong Kong, Taiwan, and Macao excluded.



**Fig. 7 Interprovincial migration flows in China.** **a** The ‘Crushing into Guangdong’ migration routines. **b** The interprovincial migration flows in China in five time periods between 1985 and 2010 (Liu et al., 2014). Reproduced with permission of Springer Science Business Media Dordrecht; copyright © Springer Science Business Media Dordrecht, all rights reserved.

with the fast development of technology. Therefore, these findings might offer support for a connection to be built between gender markedness and linguistic conservatism.

**Regions.** The GO values for 31 regions in China were calculated. Figure 6 shows the geographic distribution of gender segregation (the GO value), with darker colours showing a higher degree of segregation. It is noted that the provinces with relatively low gender segregation usually have massive flows of inter-provincial migration and rich cultural integration. Yet the provinces with higher gender segregation are usually located in remote regions and have their own ethnic cultures, which might suggest a link between the degree of gender segregation and the scale of immigration and language contact.

Figure 7 shows two main inter-provincial migration flows in different historical periods in China. Heilongjiang province, the northernmost province in China, has experienced at least two

great waves of immigration in its history. Since the second half of the 19th century, during the hundred-year period, mass Han<sup>6</sup> migration started to flow into Manchuria<sup>7</sup>. This event is known as ‘Chuang Guangdong’ (literally ‘crashing into Guangdong’); Guangdong is a traditional name of Manchuria. It is the largest inter-province migration activity since the Qing Dynasty, and a well-known historical event (Lu, 1987). Another influential migration flow to Heilongjiang occurred in 1955, with the availability of land and the support of the state policy of ‘develop the Great Northern Wilderness’, which attracted settlers from other provinces. Between the years 1955 and 1974, Heilongjiang received the highest proportion of inter-provincial migrants among all of the Chinese provinces (Liang and White, 1996). After China’s reform and opening-up in 1979, a tremendous amount of internal migration from inland to the coastal provinces (e.g. Guangdong and Zhejiang) was observed, mostly in pursuit of better job opportunities. These migration flows caused dramatic changes in the culture and language of the receiving regions. As a

result, cultural integration and language contact might have accelerated the developmental process of the regional varieties of a language; a correlation between higher degrees of linguistic innovation and lower GO values can be seen, such as in the cases of Heilongjiang and Zhejiang (the lowest compared to the other provinces).

It is believed by some scholars that the contact between varieties of language caused by migration leads to innovations (Manning, 2005). For example, Bowern (2010) noticed that highly mobile societies, such as pastoralist or nomadic ones, exhibit different patterns of variant propagation than dense, sedentary societies. Observation suggests that higher rates of exogamy may lead to greater community-internal linguistic diversity and therefore greater variation and change. A theory seeking to explain this phenomenon was developed by Trudgill (1986). His account of dialect-contact driven language change pairs accommodation with social network-based reasoning regarding frequency. Granovetter (1973, 1983) further explained that the diffusion of linguistic innovations is mediated by weak social ties and resisted in networks with strong social ties.

As for the provinces that show the highest gender segregation, we can see from Fig. 6 that they are mainly concentrated in the remote and minority regions of China. The provinces of Xinjiang, Tibet, and Inner Mongolia are listed as minority autonomous regions in China. Guizhou and Yunnan are also the provinces with a large population of minorities, while all other provinces in China have a Han majority. These provinces have also had influxes of Han populations across the years, notably, Xinjiang; however, with their geographically remote locations and culturally distinct features, these minority regions might show stronger levels of conservatism in their linguistic varieties, resulting in high GO values. Some linguists propose that the languages in a peripheral position keep most of the archaisms (Manczak, 1988), while others propose that languages in the central position have more archaic features. If the link between gender segregation and linguistic conservatism stands, the current finding would support the view that languages in remote areas are more conservative. Geographical isolation and cultural seclusion might impact regional locations causing change at a relatively slower rate.

## Discussion

In this paper, we designed an innovative method of digital humanities research to discover social development trends. We proposed and implemented a data analytics-based approach that measures gender segregation in occupations in a language without gender; specifically, we developed a new linguistic index that is grounded in gender markedness theory and has been validated using a questionnaire survey. By leveraging the relations establishing function of naming in Chinese, we developed a gender modification-based markedness scale to uncover gender typing in a genderless language. Since these terms are conceptual and not conventionalised by grammar, we expect that our approach will be just as effective when applied to languages with grammatical gender. Furthermore, by sampling newspaper data from the past 70 years and from a wide range of locations in China, we were able to conduct both longitudinal and cross-sectional studies with the same data analytics methodology. Consequently, we have been able to reach a comprehensive overview of occupational gender segregation and gendered language use covering generic differences, regional variations, and historical changes. By considering both changes and variations, we gained a deeper understanding of how different factors interact and contribute to our current state.

In addition to validating the versatility and power of our innovative data analytic approach based on diachronic and synchronic corpus data, we also uncovered the complex and competing forces

behind gendered language and occupational gender segregation. As can be seen with the general trend, we showed that China has steadily achieved a higher level of gender equality over the past 70 years, and that different geographic regions progressed at different paces. It can be reasonably concluded that the decrease in the use of gendered language reflects social change in the Chinese context, which has been guided by explicit policy as well as economic development. However, we also observed that the shift over the past 70 years did not form a single smooth curve. Instead, the line is jagged with several high and low points. With this observation, we believe that our study has revealed the tension of competing agents for change (and for inertia). By studying chronological as well as regional details of the jagged portions of the development lines, we present a more in-depth understanding of the competing forces behind the development of gender equality and the breaking down of occupational gender segregation in China.

We also established the genre-dependency of gendered language. We demonstrated that new media and new genres of web-based language contributed greatly to the reduction of gendered language, whereas formal genres are more conservative and resistant to changes for gender fairness. We suspect that different professions (i.e., language for specific purposes) will also show different trends. This is another area and set of data worthy of future research.

One of the most interesting findings of this paper is the correlation between gender equality and sociocultural changes. We established that there is a range of differences among the varieties of the same language. Trends of gendered use do not dovetail with the distribution of different language varieties. Instead, what we found was that the regions with high population mobility and heterogeneous cultural backgrounds, such as Zhejiang and Heilongjiang, showed the lowest level of gendered language usage. Moreover, we also discovered that the most homogeneous region, in terms of language and culture, is not necessarily the most conservative. The most conservative regions are in fact those with two strong competing cultures or communities. This provides a possible rebuttal of the hypothesis that immigrant populations are always more conservative. In fact, they are more conservative when they are challenged by another strong culture and have to self-preserve. In other words, new environments do lead to innovations, unless the new environment is too unaccommodating such that it forces the speakers to become defensive and ultimately more conservative.

In conclusion, our research has shown that a data analytics-based digital humanities approach can be applied to the study of historical trends and variations in social development. Using the linguistic index, longitudinal analysis reveals the social changes and trends in gender equality in occupational settings in China over the past 70 years. The variations in gender markedness in a genderless language, measured by our approach, reflect the complex interaction between cognitive, linguistic, and social changes of occupational gender segregation. It is our hope that this approach can be extended to research on other languages and that the results of this study can contribute to the broader field of occupational gender segregation, complementing the demographic-based measurements in sociology and reflecting the changes occurring in language use.

## Data availability

The questionnaire and corpus data are not publicly available due to copyright protection and confidential reasons. The questionnaire data can be made available by the corresponding author in anonymised form on reasonable request.



Received: 13 June 2020; Accepted: 12 April 2021;

Published online: 01 June 2021

## Notes

- R: masseur, tailor, cook, worker, cosmetician, machine repairman, barber, cleaner, racer, killer, driver, bartender; I: forensic, pilot, engineer, captain, scientist, veterinarian, doctor, horticulturist; A: translator, writer, singer, gardener, painter, journalist, architect, agent, cartoonist, model, magician, designer, dancer1(), dancer2(), artist, musician, presenter, actor/actress; S: security guard, nurse, coach, teacher, police officer, soldier, SWAT, firefighter, athlete; E: CEO, purchasing agent, ambassador, tour guide, shop assistant, judge, lawyer, businessman, sales clerk, diplomat; C: bank teller, attendant, administrator, accountant, salesman, postman.
- The quantitative results of this study were calculated using R.
- <http://bcc.blcu.edu.cn/>
- The Y-axis has been log scaled since the number of gendered markedness with occupations vary widely.
- Chongqing Daily (Chongqing) 64.4MB; Qianjiang Evening News (Zhejiang) 67.2MB; Yunnan Daily (Yunnan) 59.4MB; Xinjiang Daily (Xinjiang) 44.7MB; Sichuan Daily (Sichuan) 68.3MB; Tibet Daily (Tibet/Xizang) 56.5MB; BinHai Times (Tianjin) 49.5MB; PuDong Times (Shanghai) 30.2MB; Shaanxi Daily (Shaanxi) 76.2MB; Shanxi Evening News (Shanxi) 71.9MB; QILu Evening News (Shandong) 23.2MB; Qianghai Daily (Qianghai) 57.1MB; Ningxia Daily (Ningxia) 52.7MB; Inner Mongolia Daily (Inner Mongolia/Neimenggu) 62.5MB; Shenyang Evening News (Liaoning) 30.8MB; Anhui Daily (Anhui) 67.4MB; Jiangxi Daily (Jiangxi) 46.8MB; Nanjing Daily (Jiangsu) 65.5MB; Jilin Daily (Jilin) 61.6MB; Hunan Daily (Hunan) 72.3MB; Hubei Daily (Hubei) 76.0MB; Harbin Daily (Heilongjiang) 46.6MB; Zhengzhou Evening News (Henan) 61.5MB; Yanzhao Metropolis Daily (Hebei) 64.3MB; Hainan Daily (Hainan) 78.3MB; Guizhou Daily (Guizhou) 29.3MB; Guangxi Daily (Guangxi) 67.1MB; Yangcheng Evening News (Guangdong) 57.9MB; Lanzhou Daily (Gansu) 56.1MB; Fujian Daily (Fujian) 83.1MB; Beijing Daily (Beijing) 88.4MB
- Han is the majority ethnic group in China.
- Manchuria is the historical region of northeastern China. It consists of three Northeast provinces of China, i.e. Heilongjiang, Jilin, and Liaoning.

## References

- Baker M, Cornelison K (2018) Gender-based occupational segregation and sex differences in sensory, motor, and spatial aptitudes. *Demography* 55:1749–1775. <https://doi.org/10.1007/s13524-018-0706-3>
- Baker P (2014) Using corpora to analyze gender. Bloomsbury Academic, London
- Banaji MR, Hardin CD (1996) Automatic stereotyping. *Psychol Sci* 7:136–141. <https://doi.org/10.1111/j.1467-9280.1996.tb00346.x>
- Berry DM (2012) Introduction: understanding digital humanities. In: Berry DM (ed) *Understanding digital humanities*. Palgrave Macmillan, London, pp. 1–20
- Beukeboom CJ (2014) Mechanisms of linguistic bias: How words reflect and maintain stereotypical expectancies. In: Forgas JP, Vincze O, Laszlo J (eds) *Social cognition and communication*. Psychology Press, London, pp. 313–330
- Biber D, Finegan E (2001) Diachronic relations among speech-based and written registers in English. In: Biber D, Conrad S (eds) *Variation in English: multi-dimensional studies*, 1st edn. Routledge, London, pp. 66–83
- Blackburn RM et al. (2002) Explaining gender segregation. *Br J Sociol* 53:513–36. <https://doi.org/10.1080/0007131022000021461>
- Blackburn RM, Brooks B, Jarman J (2001) The vertical dimension of occupational segregation. *Work Employ Soc* 15:511–538. <https://doi.org/10.1017/S0950017001000332>
- Blackburn RM, Jarman J, Siltanen J (1993) The analysis of occupational gender segregation over time and place: considerations of measurement and some new evidence. *Work Employ Soc* 7:335–362. <https://doi.org/10.1177/095001709373001>
- Blum SD (1997) Naming practices and the power of words in China. *Lang Soc* 26:357–379. <https://doi.org/10.1017/S0047404500019503>
- Bowern C (2010) Correlates of language change in Hunter-Gatherer and other ‘small’ languages. *Lang Linguist Compass* 4:665–679. <https://doi.org/10.1111/j.1749-818X.2010.00220.x>
- Bullock EE et al. (2009) Hollanda’s theory in an international context: applicability of RIASEC structure and assessments. *Career Plan Adult Dev J* 25:29–58
- Cacciari C, Padovani R (2007) Further evidence of gender stereotype priming in language: semantic facilitation and inhibition in Italian role noun. *Appl Psychol* 28:277–293. <https://doi.org/10.1017/S0142716407070142>
- Charles M, Grusky DB (2004) *Occupational ghettos: the worldwide segregation of women and men* (Studies in social inequality). Stanford University Press, Stanford
- Cooke FL (2004) *HRM, work and employment in China*. Routledge, London
- Duncan OD, Duncan B (1955) Residential distribution and occupational stratification. *Am J Sociol* 60:493–503. <https://doi.org/10.1086/221609>
- Eckert P, McConnell-Ginet S (1992) Think practically and look locally: language and gender as community-based practice. *Annu Rev Anthropol* 21:461–490
- Emerck R et al. (2002) Indicators on gender segregation. In: Rubery J, Fagan D, Grimshaw D et al (eds) *Indicators on gender equality in the European employment strategy*. Report prepared for the European Commission, Brussels
- Fang Z (2005) Scientific literacy: a systemic functional linguistics perspective. *Sci Educ* 89:335–397. <https://doi.org/10.1002/sce.20050>
- Farris CS (1988) Gender and grammar in Chinese: with implications for language universals. *Mod China* 14:277–308
- Fromkin V, Rodman R, Hyams N (2010) *An introduction to language*, 9th edn. Cengage Learning
- Garnham A, Doehren S, Gyax P (2015) True gender ratios and stereotype rating norms. *Front Psychol* 6:1023. <https://doi.org/10.3389/fpsyg.2015.01023>
- Garnham A et al. (2016) Editorial: language, cognition and gender. *Front Psychol* 7:772. <https://doi.org/10.3389/fpsyg.2016.00772>
- Gaucher D, Friesen J, Kay AC (2011) Evidence that gendered wording in job advertisements exists and sustains gender inequality. *J Pers Soc Psychol* 101:109–128. <https://doi.org/10.1037/a0022530>
- Granovetter MS (1973) The strength of weak ties: a network theory revisited. *Social Theory* 1:201–233. <https://doi.org/10.2307/202051>
- Cutillo A, Centra M (2017) Gender-based occupational choices and family responsibilities: the gender wage gap in Italy. *Fem Econ* 23:1–31. <https://doi.org/10.1080/13545701.2017.1285041>
- Granovetter MS (1983) The strength of weak ties: a network theory revisited. *Social Theory* 1:201–233. <https://doi.org/10.2307/202051>
- Greenberg JH (1966) *Language universals: with special reference to feature hierarchies*. De Gruyter, Mouton, Berlin
- Gross E (1968) Plus ça change...? The sexual structure of occupations over time. *Soc Probl* 16:198–208. <https://doi.org/10.2307/800005>
- Gyax P et al. (2008) Generically intended, but specifically interpreted: when beauticians, musicians, and mechanics are all men. *Language Cogn Process* 23:464–485. <https://doi.org/10.1080/01690960701702035>
- Holland JL (1959) A theory of vocational choice. *J Couns Psychol* 6:35–45. <https://doi.org/10.1037/h0040767>
- Holmes J. (2001) A corpus-based view of gender in New Zealand English. In: Hellinger M, Bußmann H (eds) *Gender across languages: the linguistic representation of women and men*, vol 1. John Benjamins, Amsterdam, pp. 115–136. <https://doi.org/10.1075/impact.9.10hol>
- Horvath LK et al. (2016) Does gender-fair language pay off? The social perception of professions from a cross-linguistic perspective. *Front Psychol* 6:1–12. <https://doi.org/10.3389/fpsyg.2015.02018>
- Hou R, Huang CR (2020) Classification of regional and genre varieties of Chinese: a correspondence analysis approach based on comparable balanced corpora. *Nat Lang Eng* 26:613–640. <https://doi.org/10.1017/S1351324920000121>
- Huang CR et al. (2021) From language to meteorology: kinesis in weather events and weather verbs across Sinitic languages. *Humanit Soc Sci Commun* 8:1–13. <https://doi.org/10.1057/s41599-020-00682-w>
- Irmen L, Roßberg N (2004) Gender markedness of language: the impact of grammatical and nonlinguistic information on the mental representation of person information. *J Lang Soc Psychol* 23:272–307. <https://doi.org/10.1177/0261927X04266810>
- Juola P (2013) Using the Google N-Gram corpus to measure cultural complexity. *Lit Ling Comput* 28:668–675. <https://doi.org/10.1093/lc/fqt017>
- Kunda Z, Oleson KC (1997) When exceptions prove the rule: How extremity of deviance determines the impact of deviant examples on stereotypes. *J Pers Soc Psychol* 72:965–79. <https://doi.org/10.1037/0022-3514.72.5.965>
- Leavy S (2018) Uncovering gender bias in newspaper coverage of Irish politicians using machine learning. *Digit Scholarsh Humanit* 34:48–63. <https://doi.org/10.1093/lc/fqy005>
- Li L, Huang CR, Wang VX (2020) Lexical competition and change: a corpus-assisted investigation of gambling and gaming in the past centuries. *SAGE Open* 10:1–14. <https://doi.org/10.1177/2158244020951272>
- Liang Z, White MJ (1996) Internal migration in China 1950–1988. *Demography* 33:375–384. <https://doi.org/10.2307/2061768>
- Liben LS, Bigler RS, Krogh HR (2002) Language at work: children’s gendered interpretations of occupational titles. *Child Dev* 73:810–828. <https://doi.org/10.1111/1467-8624.00440>
- Liu Y et al. (2014) Interprovincial migration, regional development and state policy in China, 1985–2010. *Appl Spat Anal* 7:47–70. <https://doi.org/10.1007/s12061-014-9102-6>
- Long L, Peng P (2000) The development of college speciality finder for self-directed search in China. *Acta Psychol Sin* 32:453–457
- Long L, Peng P, Zheng B (1996) A report on the use of Self-Directed Search [SDS] Inventory. *Appl Psychol* 2:44–51
- Lu Y (1987) A brief history of the immigration from Shandong to the northeast in the Qing and Republican periods. *Shanghai Academy of Social Sciences, Shanghai Macalister J* (2011) Flower-girl and bugler-boy no more: changing gender representation in writing for children. *Corpora* 6:25–44. <https://doi.org/10.3366/cor.2011.0003>

- Makkreel RA (2016) Wilhelm Dilthey. In: Zalta EN (ed). The Stanford encyclopedia of philosophy (fall 2016 edition). <https://plato.stanford.edu/archives/fall2016/entries/dilthey/>
- Makkreel RA, Rodi F (eds) (1990) Wilhelm Dilthey: selected works, vol I: introduction to the human sciences. Princeton University Press, Princeton
- Manczak W (1988) Bartoli's second "norm". In: Fisiak J (ed.) Historical dialectology: regional and social. De Gruyter Mouton, Berlin, pp 349–355. <https://doi.org/10.1515/9783110848137.349>
- Manning P (2005) Migration in world history. Routledge, London
- Moir H, Smith JS (1979) Industrial segregation in the Australian labour market. *J Ind Relat* 21:281–291. <https://doi.org/10.1177/002218567902100302>
- Moser F, Masterson J (2014) Are there signs of change in gendered language use in children's early reading material? *Gen Lang* 8:71–89. <https://doi.org/10.1558/genl.v8i1.71>
- Naroll R (1961) Two solutions to Galton's problem. *Philos Sci* 28:15–39
- Norberg C (2016) Naughty boys and sexy girls: the representation of young individuals in a web-based corpus of English. *J Engl Linguist* 44:291–317. <https://doi.org/10.1177/0075424216665672>
- Pacheco LM (2018) Gender asymmetries in news reports. *Miscelánea* 57:121–139
- Parrillo VN (ed) (2008) Encyclopedia of social problems. Sage Publications, New York
- Pattueli MC, Hwang K, Miller M (2017) Accidental discovery, intentional inquiry: leveraging linked data to uncover the women of jazz. *Digit Scholarsh Humanit* 32:918–924. <https://doi.org/10.1093/lc/fqw047>
- Popp D et al. (2003) Gender, race, and speech style stereotypes. *Sex Roles* 48:317–325. <https://doi.org/10.1023/A:1022986429748>
- Preston JA (1999) Occupational gender segregation trends and explanations. *Q Rev Econ Financ* 39:611–624. [https://doi.org/10.1016/S1062-9769\(99\)00029-0](https://doi.org/10.1016/S1062-9769(99)00029-0)
- Prewitt-Freilino JL, Caswell TA, Laakso EK (2012) The gendering of language: a comparison of gender equality in countries with gendered, natural gender, and genderless languages. *Sex Roles* 66:268–281. <https://doi.org/10.1007/s11199-011-0083-5>
- Reskin BF, Hartmann HI (eds) (1986) Women's work, men's work: sex segregation on the job. The National Academies Press, Washington, DC
- Roberts S, Winters J (2013) Linguistic diversity and traffic accidents: lessons from statistical studies of cultural traits. *PLoS ONE* 8:e70902. <https://doi.org/10.1371/journal.pone.0070902>
- Ruble TL (1983) Sex stereotypes: issues of change in the 1970s. *Sex Roles* 9:397–402. <https://doi.org/10.1007/BF00289675>
- Sanchez D et al (2017) Familial ethnic socialization, gender role attitudes, and ethnic identity development in Mexican-origin early adolescents. *Cultur Divers Ethnic Minor Psychol* 23:335–347. <https://doi.org/10.1037/cdp0000142>
- Sendén MG, Bäck EA, Lindqvist A (2016) Introducing a gender-neutral pronoun in a natural gender language: the influence of time on attitudes and behavior. *Front Psychol* 6:893. <https://doi.org/10.3389/fpsyg.2015.00893>
- Stahlberg D et al (2007) Representation of the sexes in language. In: Fiedler k (ed) Social communication. Psychology Press, New York, pp. 163–187
- Stoet G, Geary DC (2018) The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychol Sci* 29:581–593. <https://doi.org/10.1177/0956797617741719>
- Su R, Rounds J, Armstrong PI (2009) Men and things, women and people: a meta-analysis of sex differences in interests. *Psychol Bull* 135:859–884. <https://doi.org/10.1037/a0017364>
- Trudgill P (1986) Dialects in contact (Language in Society 10). Blackwell, Oxford
- Verniers C and Vala J (2018) Justifying gender discrimination in the workplace: the mediating role of motherhood myths. *PLoS ONE* 13:e0201150. <https://doi.org/10.1371/journal.pone.0190657>
- Wang CH (1999) Gender differences in their perceptions and ideologies among political cadres. *Collect Women's Stud* 2:17–20
- Watts M (1995) Divergent trends in gender segregation by occupation in the United States: 1970–92. *J Post Keynes Econ* 17:357–379. <https://doi.org/10.1080/01603477.1995.11490035>
- Watts M (1998) Occupational gender segregation: index measurement and econometric modeling. *Demography* 35:489–500. <https://doi.org/10.2307/3004016>
- Xu H et al. (2020) Light verb variations and varieties of Mandarin Chinese: comparable corpus driven approaches to grammatical variations. *Corpus Linguist Ling Theory*. <https://doi.org/10.1515/clt-2019-0049>
- Yzerbyt VY, Coull A, Rocher SJ (1999) Fencing off the deviant: the role of cognitive resources in the maintenance of stereotypes. *J Pers Soc Psychol* 77:449–462. <https://doi.org/10.1037/0022-3514.77.3.449>
- Zheng W (2003) Gender, employment and women's resistance. In: Perry E, Selden M (eds) Chinese society, change, conflict, and resistance. Routledge, London, pp. 176–204. <https://doi.org/10.4324/9780203302606-15>

### Acknowledgements

This research was funded by the National Key Research and Development Programme of China (2019YFC1521200) and the National Natural Science Foundation of China (72010107003).

### Competing interests

The authors declare no competing interests.

### Additional information

Correspondence and requests for materials should be addressed to Q.S., P.L. or C.-R.H.

Reprints and permission information is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons

Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021