International Journal of Psychology

IUPsyS

International Journal of Psychology, 2016
DOI: 10.1002/ijop.12259

Personality expression in Chinese language use

Lin Qiu[1], Jiahui Lu[1], Jonathan Ramsay[2], Shanshan Yang[1], Weina Qu[3], and Tingshao Zhu[3]

[1]Division of Psychology, Nanyang Technological University, Singapore
[2]SIM University, Singapore
[3]Institute of Psychology, Chinese Academy of Sciences, Beijing, China

To date, little research has investigated personality expressions in languages other than English. Given that the Chinese language has the largest number of native speakers in the world, it is vitally important to examine the associations between personality and Chinese language use. In this research, we analysed Chinese microblogs and identified word categories and factorial structures associated with personality traits. We also compared our results with previous findings in English and showed that linguistic expression of personality has both universal- and language-specific aspects. Expression of personality via content words is more likely to be consistent across languages than expression via function words. This makes an important step towards uncovering universal patterns of personality expression in language.

Keywords: Personality; Chinese; Language use; Linguistic analysis; Microblog; Social media.

Language use reflects important social and psychological processes (Pennebaker, Mehl, & Niederhoffer, 2003). Studies have shown ample evidence of the association between personality traits and word use in a variety of English writing samples (Pennebaker & King, 1999; Pennebaker et al., 2003). Nonetheless, little is known about the extent to which personality is reflected in the verbal content of other languages (Pennebaker et al., 2003). This issue is made more pressing by the fact that about one-fifth of the world population lives in China, and almost 1.2 billion people speak Chinese as their native language (Paul, Simons, & Fennig, 2015). Compared with English (which has around 335 million native speakers), Chinese is much more widely spoken. Given these disparities and the sheer number of Chinese speakers, understanding how Chinese language use is associated with personality will improve our knowledge about the psychological processes of the world's largest language community and make an important step towards uncovering universal patterns of personality expression in language.

Past research on personality and the English language has often utilised Linguistic Inquiry and Word Count (LIWC; Tausczik & Pennebaker, 2010): a software program that analyses writing samples to assess the frequency of certain words. Developed by combining grammatical rules with the content of various psychological measurement scales, LIWC can reliably measure many psychological attributes including emotions, personality, thinking styles and social relationships (Tausczik & Pennebaker, 2010). LIWC counts the frequency of words in pre-defined categories, each of which belongs to one of two broad groups: content words and function words (Tausczik & Pennebaker, 2010). Content words express the semantic meaning of language and include categories such as positive emotion and social processes. Functions words indicate the grammatical relationships between content words. LIWC contains functional word categories such as pronouns and articles. Both content and function word categories have been found to predict Big Five personality traits in English samples (Pennebaker & King, 1999; Pennebaker et al., 2003).

Despite the lack of cross-cultural research on the associations between language and personality, considerable evidence suggests that the structure of personality—particularly the Big Five—model is stable across cultures (McCrae & Costa, 1997). The model has been validated not only in Western samples but also in Asian samples including Chinese (e.g., McCrae, Costa, del Pilar, Rolland, & Parker, 1998). Therefore, it is logical to expect that many of the linguistic markers of personality traits should be similar between Chinese and English. This relationship is likely to be particularly strong in the case of content words, as such words typically take the form of nouns, regular verbs and adjectives (Tausczik

Correspondence should be addressed to Dr Lin Qiu, Division of Psychology, Nanyang Technological University, HSS-04-15, 14 Nanyang Drive, Singapore 637332. (E-mail: linqiu@ntu.edu.sg).

& Pennebaker, 2010) and refer to objects, activities and concepts with a defined meaning that the majority of them are likely to possess equivalents across languages (Brown, 1991).

However, Chinese and English have very different grammatical rules. For example, a number of Chinese function word categories such as second person plural pronoun and preposition phrase endings do not have equivalents in the English language. In Chinese, verbs do not have tenses. The concept of time is expressed through the use of adverbs as tense markers. In addition, *pronoun drop*, a linguistic phenomenon where the subject pronoun of a sentence such as "I" or "You" is dropped, is common in Asian languages such as Chinese, Japanese or Korean, but is not permitted in Germanic languages such as English, German and Swedish (Kashima & Kashima, 1998). It is not grammatically correct to omit the subject "I" in such a phrase as "I had dinner with my family yesterday" in English, but such a practice is acceptable in Chinese or Japanese. Given these substantial grammatical differences, it is likely that the Chinese language features associations between grammar and personality that do not exist in English and vice versa. Consequently, we expected that the manifestation of personality in function words would exhibit less correspondence with English than manifestation in content words, as the use of function words is determined by highly language-specific grammatical rules.

PRESENT STUDY

In this study, we used LIWC to analyse writing samples from two groups of users on Sina Weibo, one of the largest microblogging platforms in China. Social media was chosen as the preferred medium of analysis for two reasons: Ecological validity and accessibility. Traditional research has primarily relied on language samples collected from controlled diary studies or decontextualized laboratory environments (Tausczik & Pennebaker, 2010), compromising the ecological validity of the resulting findings. More research is needed to examine language use in naturalistic settings.

Furthermore, the use of online data allows access to a wealth of content generated by a large number of people in geographically remote places. This greatly increases the scope and convenience of research while also enhancing ecological validity. Scholars have argued that psychological research should harness the benefits of online data to complement traditional methods (Gosling & Mason, 2015). Therefore, we opted to use online microblogs as our writing samples.

Two separate samples were recruited. Sample 1 were online participants from Mainland China, and Sample 2 were overseas Chinese students. The consistent word-personality associations found between them allowed us to identify stable patterns of personality expression. In addition, we used factor analysis to identify language structures associated with personality traits. We used the large sample (Sample 1) to establish the basic factor structure, before using data from the small sample (Sample 2) to investigate the its replicability.

METHOD

Participants

Participants in Study 1 were recruited online from mainland China. We developed a software tool and sent participation requests to 50,000 Sina Weibo users who (a) posted more than two and less than 50 microblogs every day and (b) had been using Sina Weibo for more than 1 month. The selection criteria were designed to identify active users and avoid accounts created for spamming purposes. A total of 470 Sina Weibo users (females = 292, males = 178) responded and participated in our study for payment of RMB30 (US$4.8) per person. This low response rate was likely due to the huge amount of spam on Sina Weibo, as the frequency of solicited messages may desensitise people to such requests. The sample featured a diverse range of ages but was unsurprisingly skewed towards younger individuals. Many (41.5%) of the participants were younger than 20 years old, while a further 38.7% of participants were between the ages of 21 and 25. Participants between the ages of 26 and 30 comprised 14.3% of the sample, while 3.2% participants were between the ages of 31 and 35. The remainder of the sample (2.3%) were aged 36 and above.

Participants in Sample 2 were 90 Chinese students at a large university in Singapore (females = 67, mean age = 22.4 years, SD = 2.52). All participants were Chinese nationals currently pursuing their undergraduate degree in Singapore. They had been in Singapore for 2–4 years. They participated in our study for payment of S$5 (US$4.03) and completed the same survey as those who participated in Study 1.

Measures and procedure

Participants in both Study 1 and Study 2 provided informed consent before commencing participation. They agreed to complete a personality survey and allow us to collect their public profile information and microblogs on Sina Weibo. Each participant completed the Big Five Personality Inventory (BFI; John, Donahue, & Kentle, 1991) and demographic questions (i.e., age, gender and ethnicity). The BFI is a well-established and widely used measure of personality. It exhibited acceptable reliability in our study (see Table 1).

We developed a software tool and used it to download participants' microblogs through Sina Weibo API. We

TABLE 1
Descriptive statistics for self-ratings of personality traits in S1 and S2

	S1 (n = 470)			*S2 (n = 90)*		
	Mean	*SD*	*Cronbach's α*	*Mean*	*SD*	*Cronbach's α*
Extraversion	3.20	.62	.69	3.11	.57	.70
Agreeableness	3.63	.57	.63	3.72	.48	.63
Conscientiousness	3.15	.56	.69	3.22	.58	.77
Neuroticism	3.06	.63	.68	3.00	.70	.82
Openness	3.61	.56	.71	3.63	.50	.76

then removed information such as reposts written by others, timestamps, geo-locations and embedded URLs, to obtain original texts written by the participants. All participants had written more than 50 microblogs. On average, each participant in Sample 1 wrote 1237.7 microblogs (SD = 841.50) within the time period of 445.61 days (SD = 161.36). Each participant in Sample 2 wrote 276.9 microblogs (SD = 224.77) in a period of 565.78 days (SD = 189.46). The large difference between the average number of microblogs per person is mainly due to the selection criteria in Sample 1 that required users to post more than two microblogs every day.

We replaced commonly used emoticons with corresponding phrases in everyday language and used a widely used Chinese lexical analyser ICTCLAS (Zhang, Liu, Cheng, Zhang, & Yu, 2003) to segment the writing samples into words, because Chinese texts do not contain word delimiters such as whitespaces. ICTCLAS uses a unified framework that includes part-of-speech tagging, disambiguation and unknown words recognition, to identify individual words in a piece of Chinese text. We identified 7,180,608 words in Study 1 and 392,322 words in Study 2.

Subsequently, we used the Simplified Chinese version of LIWC (Huang et al., 2012) to generate word frequencies in the two samples. Out of the 72 LIWC categories, 26 categories had frequencies lower than 1% in both samples and were removed from further analysis. All the remaining categories had frequencies larger than 1% in both samples, except one category (i.e., second singular pronoun) had a frequency of 1.45% in Study 1 and 0.94% in Study 2. To further ensure that the word categories that we focused on were relatively stable, we followed the procedure in Qiu, Lin, Ramsay, and Yang (2012) and split each participant's writing sample into two halves by randomly selecting half of the microblogs from the whole sample and applied LIWC analysis to each half. We correlated the word frequencies and found that all the remaining categories had an above moderate correlation coefficient ($r > .3$; Cohen, 1988) in Study 1, while two categories (i.e., *certainty* and *perceptual process*) had a correlation coefficient lower than 0.3 in Study 2. We removed the two categories from further analysis.

RESULTS

Correlations with personality

We correlated LIWC word frequencies with participants' Big Five personality traits. A total of 49 (22.27%) out of 220 correlations in Sample 1 and 19 (8.64%) out of 220 correlations in Sample 2, were significant at $p < .05$, exceeding chance. Table 2 shows the word categories that had significant correlations with at least one personality trait in one sample. We added a column in the table to show how these categories were associated with personality traits in previous English samples for comparison (see the column *Se* in Table 2). There were 43 associations between personality traits and word categories (21 with function word categories and 22 with content word categories) that were significant in at least one sample and had the same direction in both samples. These associations suggest relatively stable personality expressions in two samples.

Seven (33.3%) out of the 21 associations between personality traits and function word categories replicated findings in English. For example, extraversion was positively correlated with personal pronouns (including first person and second person singular pronouns), suggesting that extraverts tended to be concerned about people. This is consistent with the theoretical definition of extraversion (Costa & McCrae, 1996) and has been found in English samples (Hirsh & Peterson, 2009; Oberlander & Gill, 2006; Pennebaker & King, 1999; Yarkoni, 2010). Extraversion was negatively correlated with impersonal pronouns, suggesting that extraverts were less likely to use language of an impersonal nature. This association had been found in English text messages (Holtgraves, 2011) and tweets (Qiu et al., 2012). Extraversion was negatively associated with negations, supporting past findings (Tausczik & Pennebaker, 2010) and reflecting extraversion's connection with reduced cognitive complexity (Costa & McCrae, 1996). Neuroticism negatively correlated with prepositions, replicating past findings in English essays (Mairesse, Walker, Mehl, & Moore, 2007).

Fourteen (66.7%) out of the 21 associations between personality traits and function word categories had not

TABLE 2
Correlations between personality and LIWC word categories

		Extraversion			*Agreeableness*			*Conscientiousness*			*Neuroticism*			*Openness*			*Age*		*Gender*	
Word categories	*Examples*	*S1*	*S2*	*Se*	*S1*	*S2*	*Se*	*S1*	*S2*	*Se*	*S1*	*S2*	*Se*	*S1*	*S2*	*Se*	*S1*	*S2*	*S1*	*S2*
Function words																				
Total pronouns	你(you), 他们(they), 它(it)	.09	.12	[+][5]	−.02	.00	[+][12]	−.07	−.05		**.10***	.03	[+][5]	.04	.01	[−][12]	−.31**	−.38**	.21**	.40**
Personal pronouns	他(he), 在下(I), 你们(you)	***.15****	.17	[+][4]	−.03	−.04		**−.11***	−.05		.09	.08		.01	.03		−.38**	−.41**	.24**	.38**
First person singular	本人(I), 我(I), 自己(myself)	***.14****	.10	[+][4,5,9,12,13]	−.04	−.02	[+][6,10,12]	**−.14****	−.06		.08	.12	[+][1,10,11,12,13]	.03	.00	[−][10,12]	−.30**	−.34**	.25**	.41**
Second person singular	你(you), 您(you)	**.11***	***.24****	[+][12]	−.05	−.08		−.05	.01		.08	.06	[+][13],[−][12]	.01	.08	[−][11,12]	−.36**	−.32**	.14**	.13
Second person plural	你们(you), 您们(you), 汝等(you)	.06	.12		−.07	.03		***−.15****	−.04		.08	*−.20*		−.06	.12		−.22**	−.14	.08	.12
Impersonal pronouns	它(it), 那些(those)	***−.11****	−.06	[−][4,11]	.02	.14		.07	−.05		.08	−.12		.06	−.04		.05	−.15	.00	.27**
Special articles	本, 该, 每	−.06	−.03		.08	.17		.04	−.03		−.02	−.05		***−.10****	−.03		.18**	.07	−.11*	−.02
Auxiliary verbs	可能 (may), 应该(should), 不必(shouldn't)	−.08	−.09	[−][11]	.01	.01		−.03	.11		**.12****	.07		−.02	−.01		−.11*	−.22*	.27**	.31**
Multifunction words	的, 有, 是	***−.11****	−.02		.07	.17		**.11***	.11		.02	*−.16*		.07	.11		.20**	.01	.06	.25*
Tense markers	已经(already), 之前(before), 日后(later)	−.02	.09		−.04	.22*		−.07	***.26****		.05	−.11		−.01	−.06		−.09	.10	−.01	.11
Progress markers	了(already), 至今(until now), 近期(recently)	−.03	.09		−.08	.19		**−.10***	***.23****		.09	−.09		.00	−.03		−.11*	.10	.00	.03
Prepositions	到(to), 与(with), 关于(about)	−.02	***.23****		***.18****	.16		***.15****	.09		−.08	*−.22**	[−][5]	.08	.12	[+][8,11,12]	.19**	.03	−.11*	.22*
Preposition end	之中(in), 以上(above), 为止(until)	−.01	.13		.08	***.30****		.08	.14		−.06	−.20		−.08	.04		.24**	.04	−.16**	−.07
Conjunctions	和(and), 但是(but), 然而(whereas)	***−.10****	.03	[+][9]	***.11****	−.01		.07	.14		.04	−.02	[+][9,11]	.09	.05		.07	.01	.10*	.31**
Negations	不(no), 未必(not), 绝不(never)	***−.13****	−.18	[−][5,7,10]	−.02	.00	[−][8,11]	.02	−.06	[−][10,12]	.03	−.02	[+][5,11,12]	−.03	.01	[−][12]	−.05	−.02	−.07	.13
Quantifiers	一些(some), 所有(all), 众多(many)	−.05	−.09		**.11***	.17		.08	−.06		−.03	−.07		.01	.04		.31**	.04	−.09*	.15
Quantity unit	条, 头, 枝	.07	−.05		***.13****	.20		.07	.08		**−.12***	−.15		**−.09***	−.07		.15**	.43**	−.24**	−.18
Content words																				
Common verbs	走(walk), 去(go), 看(see)	.00	.14		.01	.18		.02	***.23****		.09	−.12	[+][9]	−.02	.12	[−][11]	−.04	−.02	.16**	.29**

TABLE 2
continued

		Extraversion			*Agreeableness*			*Conscientiousness*			*Neuroticism*			*Openness*			*Age*		*Gender*	
Word categories	*Examples*	*S1*	*S2*	*Se*	*S1*	*S2*	*Se*	*S1*	*S2*	*Se*	*S1*	*S2*	*Se*	*S1*	*S2*	*Se*	*S1*	*S2*	*S1*	*S2*
Adverbs	曾经(once), 非常(very), 渐渐(gradually)	−***.10****	.00		.02	.02		−.03	.05		***.15*****	.08	[−]^{9}	.03	.05	[−]^{11}	−.09*	−.17	.22**	.39**
Numbers	一 (one), 百(hundred), 千(thousand)	.04	−.04	[+]^{1,7,12}	.04	.07	[+]^{12}	***.10****	.09		−.05	−.15	[−]^{5}	−.08	−.08	[−]^{12}	.16**	.20	−.18**	−.21
Social processes	谈话(talk), 接纳(accept), 打招呼(greeting)	.06	***.24****	[+]^{3,8,9,10,11,12}	.03	.07	[+]^{12}	.04	.00	[+]^{2}	.06	−.17	[−]^{1}	−.01	.11	[−]^{6,12}	−.14**	−.29**	.14**	.20
Humans	成人(adult), 宝宝(baby), 男孩(boy)	−.07	−.08	[+]^{3,8,12}	.08	.16		***.15*****	−.03	[+]^{2},[−]^{12}	.00	−.12	[+]^{5},[−]^{8}	−.08	−.01	[−]^{12}	.16**	−.20	.00	.20
Affective processes	气愤(angry), 感恩(gratitude), 失望(disappointed)	.08	−.01	[+]^{1,11}	−.05	.02	[+]^{2}	−.02	.00		.02	.05		−.03	−.12	[−]^{11,12}	−.23**	−.25*	.15**	.16
Positive emotion	高兴(happy), 满足(satisfied), 甜蜜(sweet)	.08	−.02	[+]^{5,10,11,12}	−.03	.14	[+]^{2,10,12}	.01	.07	[+]^{10}	−.03	−.11	[−]^{5,10}	−.01	−.09	[+]^{8},[−]^{11,12}	−.18**	−.23*	.14**	.19
Negative emotion	担忧(worried), 丑恶(ugly), 糟糕(terrible)	.04	−.01	[−]^{10}	−.06	−.10	[−]^{4,10,12}	−.09	−.05	[−]^{6,10,12}	***.11****	.14	[+]^{3,4,5,10,12}	−.07	−.15		−.23**	−.20	.03	.06
Cognitive processes	理解(understand), 选择(choose), 质疑(question)	−***.14*****	.05		.05	.11		.03	*.19*	[−]^{12}	***.10****	−.03	[+]^{12}	−.02	.06	[−]^{12}	.00	−.09	.10*	.30**
Insight	了解(understand), 恍然大悟(know), 体会(realise)	−.08	.18		−.04	.20		−.04	***.28*****		***.10****	−.12		.00	.05	[+]^{10}	−.08	.04	.00	.10
Discrepancy	欠缺(lack), 必须(must), 期待(expect)	*−.07*	−.08	[−]^{8}	.01	.00	[−]^{8}	.01	.08	[−]^{10,12}	***.11****	.10	[+]^{8,12,13}	−.03	−.02	[−]^{12}	−.10*	−.26*	.23**	.35**
Tentative	大约(about), 未定(unsure), 差不多(almost)	−***.14*****	.02	[−]^{9,10,12}	.01	.04		.00	−.02	[−]^{12,13}	***.12*****	.10	[+]^{12}	.00	.05	[+]^{10}	−.07	−.18	.16**	.38**
Inclusive	包括(include), 附近(near), 添加(add)	−.03	.07	[+]^{5,9,12}	***.12****	.18	[+]^{3,12}	.00	.14		.00	−.04	[+]^{1,9,13}	−.06	.02	[+]^{8}	.03	.04	.03	.31**
Exclusive	取消(cancel), 但是(but), 除外(exclude)	−***.11****	.03	[−]^{9,10}, [+]^{13}	.01	*−.19*	[−]^{11}	−.02	.01	[−]^{3,10,12}	***.12*****	.08	[+]^{9,12}	.01	.09	[+]^{3,10}	−.03	−.20	.11*	.23*

TABLE 2
continued

		Extraversion			Agreeableness			Conscientiousness			Neuroticism			Openness			Age		Gender	
Word categories	*Examples*	*S1*	*S2*	*Se*	*S1*	*S2*	*Se*	*S1*	*S2*	*Se*	*S1*	*S2*	*Se*	*S1*	*S2*	*Se*	*S1*	*S2*	*S1*	*S2*
Biological processes	头晕(dizzy), 流汗(sweat), 拥抱(hug)	***.11****	.09		−.01	***.27****	[+][2]	−.09	.13		.09	−.13		−.03	.10		−.05	.05	.23**	.13
Body	脖子(neck), 皮肤(skin), 睡(sleep)	.05	.09	[−][5]	−.07	.18	[−][3]	***−.15****	.10	[−][3]	.08	−.13	[+][3]	−.06	.14		−.10*	−.02	.12**	.10
Relativity	以前(past), 相比(comparably), 达到(reach)	.07	.06		***.10****	***.34****	[+][12]	.08	*.19*		−.09	−.09		***−.10****	−.06	[−][12]	.15**	.17	−.13**	.22*
Motion	通过(through), 靠近(approach), 参加(participate)	.05	.10		.00	.22*	[+][12]	.03	.17		−.06	−.01		***−.11****	.04		.09*	.09	−.08	.18
Space	里面(inside), 街道(street), 台上(on stage)	.02	.02		***.12****	***.23****	[+][12]	.08	.03		***−.11****	−.09	[−][12]	.01	−.07	[−][12]	.29**	.07	−.17**	.09
Time	期间(period), 过去(past), 秋天(autumn)	.06	.07		.07	***.31****	[+][12]	.06	***.24****	[+][12]	−.04	−.07		***−.12****	−.04	[−][12]	−.02	.16	−.03	.22*
Work	工厂(factory), 面试(interview), 薪水(salary)	−.02	.10	[+][2], [−][8,12]	.07	.10		***.13****	.21*	[+][3]	***−.15****	***−.28****	[+][8],[−][3,5]	***−.13****	.08	[−][8]	.12*	.16	−.34**	−.20
Achievement	擅长(skilled), 赢得(win), 高手(master)	−.10*	.07	[−][8,12]	.07	.20		***.22****	***.30****	[+][3,12]	***−.12****	−.19		−.04	.03		.24**	.21*	−.23**	−.19

Note: Gender: 1= Male, 2= Female. Se indicates findings in past English samples. Only categories that correlate significantly with at least one trait are shown. Italicised values indicate correlations that remained significant after controlling for age and gender. Bold values indicate correlations that remained significant after controlling for number of postings.
LIWC = Linguistic Inquiry and Word Count.
* $p<.05$;** $p<.01$, two tailed.

been reported before. For example, conscientiousness was negatively associated with personal pronouns (including first person singular and second person plural), suggesting that conscientious individuals focused less on interpersonal issues, because personal pronouns indicate attentional focus on people (Tausczik & Pennebaker, 2010). Conscientiousness was also associated with prepositions, a linguistic feature indicating concern with precision (Tausczik & Pennebaker, 2010), reflecting the trait definition (e.g., Costa & McCrae, 1996). Openness was negatively related to the use of quantity unit and special articles. As these words are used to quantify persons or objects, this suggests that individuals with higher degree of openness are less likely to make quantified statements.

Seventeen (77.2%) out of the 22 associations between personality traits and content categories were consistent with past findings in English. Extraversion was related to social processes, a relationship that has been consistently found in many English samples (Hirsh & Peterson, 2009; Nowson, 2006; Oberlander & Gill, 2006; Pennebaker & King, 1999; Qiu et al., 2012; Yarkoni, 2010). Agreeableness was positively associated with inclusive and relativity (including space and time) words, replicating the associations found in blogs (Yarkoni, 2010). Conscientiousness was associated with time, work and achievement, supporting past findings (Hirsh & Peterson, 2009; Yarkoni, 2010) and reflecting the fact that conscientious individuals are achievement-oriented and hardworking (Barrick & Mount, 1991). Neuroticism correlated with negative emotion words, consistent with past findings (Hirsh & Peterson, 2009; Mairesse et al., 2007; Pennebaker & King, 1999) and reflecting the fact that individuals with higher degree of neuroticism experience more negative emotions (McCrae & Costa, 1987). Neuroticism also correlated with discrepancy, tentative and exclusive words; findings that were consistent with the results of previous research (e.g., Nowson, 2006; Oberlander & Gill, 2006; Yarkoni, 2010; Yee, Harris, Jabon, & Bailenson, 2011). These words indicate making distinctions (Nowson, 2006; Pennebaker & King, 1999) and suggest that emotionally unstable individuals tend to make distinctions in their writings. Neuroticism negatively correlated with work related words, supporting the negative relationship between neuroticism and work performance (Barrick & Mount, 1991; Nowson, 2006). Openness was negatively correlated with relativity and time, replicating relations found in personal essays (Mairesse et al., 2007).

Only five (22.7%) out of the 22 associations between personality traits and content categories were new. Extraversion was correlated with biological processes, suggesting that extraverts mentioned more about topics related to activities such as eating and sex than their introverted counterparts. Neuroticism was also negatively correlated with achievement words, supporting the negative relationship between neuroticism and work performance (Barrick & Mount, 1991).

The above results showed that while 66.7% of personality expressions associated with function word categories had not been found in English samples before, only 22.7% of those associated with content categories were new. This supported our hypothesis that personality expressions associated with function word categories are more likely to be language specific than those associated with content categories.

Gender, age and number of postings

In our samples, females used more personal pronouns, discrepancy, tentative and filler words, consistent with previous findings in English (Mehl & Pennebaker, 2003). In addition, older individuals used fewer social words and personal pronouns (including first and second singular pronouns), supporting previous results and indicating older individuals' more infrequent engagement in social activities (Pennebaker & Stone, 2003). However, we found that older individuals used fewer positive and negative emotion words. This is inconsistent with previous English findings about older individuals used more positive but fewer negative emotion words (Pennebaker & Stone, 2003).

To examine if the observed word-personality correlations were contingent on age and gender, we calculated partial correlations by controlling gender and age. Among the previously found correlations, 31 out of 49 (63.27%) in Sample 1, and 15 out of 19 (78.95%) in Sample 2, remained significant. This indicates that while some linguistic cues might reflect characteristics related to age and gender, those remained significant were likely to be directly related to personality traits (see italicised correlations in Table 2).

We also examined if the observed word-personality correlations were contingent on the number of postings. We calculated partial correlations controlling for the number of postings. Among the previously found correlations, 48 out of 49 in Sample 1, and 15 out of 19 in Sample 2, remained significant (see Table 2). This indicates that the majority of the observed word-personality correlations were independent from the number of postings.

Factor analysis

Factor analysis clusters words based on their natural co-occurrence to identify their common discourse functions. It has been used to reveal psychological constructs underlying language patterns (Pennebaker et al., 2003). We conducted an exploratory factor analysis on the Sample 1 data following the procedure used by Pennebaker and King (1999). First, we included word categories that had a high correlation ($r > .7$) in the previous split-half

TABLE 3
Factor loading

Categories	*Factor 1 Making distinction*	*Factor 2 Reflection*	*Factor 3 Objective description*	*Factor 4 Socialisation*
Exclusive	.87			
Conjunction	.83			
Tentative	.78			
Adverbs	.76	.44		
Discrepancy	.73			
Impersonal pronouns	.72			
Tense markers		.93		
Insight		.83		
Interjunctions		.68		
Assent		.67	−.45	
Space			.65	
Numbers			.56	
Non-fluencies			−.50	
Quantity unit			.48	
Positive emotion			−.47	
Time			.45	
Motion			.43	
Social processes				.83
Second person singular pronouns				.80
First person singular pronouns				.45

Note: Only loadings of .40 and above are shown. *Negative emotion*, *negations* and *body* were removed from display because they have loadings below .40.

analysis. Second, we excluded categories that were largely included in other categories. For example, the majority of prepositions appear in inclusive and exclusive categories. Therefore, we removed the preposition category. Third, we removed categories that did not include specific words (i.e., total word count). Fourth, categories related to specific topics were excluded because they reflected personal interests rather than psychological processes. This resulted in a total of 23 word categories in the factor analysis. The KMO test and Bartlett's Test of Sphericity indicated that the data were appropriate for factor analysis (KMO = .791, Bartlett's Test of Sphericity $p < .001$). A scree analysis suggested that a 4-factor solution would best fit the data. Table 3 shows the results of factor analysis by forcing four factors with maximum-likelihood extracted method and varimax rotation.

The first factor (eigenvalue = 4.39) included six word categories with a loading of .4 or greater. They were three cognitive process categories including tentative, exclusive and discrepancy words, and three functional categories including impersonal pronouns, adverbs and conjunctions. This factor was similar to the making distinction factor in English which included tentative, exclusive and discrepancy words (Pennebaker & King, 1999). Therefore, this factor was termed Making Distinction.

The second factor (eigenvalue = 3.26) had five word categories, including tense marker, insight, interjunction, assent and a secondary loading of adverbs. This factor suggested thoughtful discussion of the past, present and future. Therefore, it was termed Reflection.

The third factor (eigenvalue = 2.44) included eight word categories. They were time, space, numbers, quantity unit, fewer non-fluencies, fewer positive emotions and a secondary loading of fewer assent. This factor indicated description of time, space and objects. It was termed Objective Description.

The fourth factor (eigenvalue = 1.95) contained social words, and first and second person singular pronouns. It indicated social interaction and personal attention and was labelled Socialisation.

Factor scores were calculated using weighted sum scores of items loaded on each factor. We correlated the factor scores with participants' Big Five personality dimensions (see Table 4). Extraversion positively correlated with Socialisation and negatively correlated with Making Distinction, suggesting that extraverts were more likely to mention social activities but less likely to indicate distinctions than introverts. Agreeableness correlated with Objective Description, showing that more agreeable individuals were more likely to mention about objective matters in their writings. Conscientiousness positively correlated with Objective Description and negatively correlated with Reflection, indicating that conscientious individuals tended to talk about factual matters rather than their own thoughts. Neuroticism was positively associated with Making Distinction, indicating that higher neuroticism individuals had more attentional focus on differences. The negative association between extraversion and Making Distinction has been found in English before (Pennebaker & King, 1999).

To examine the replicability of the above factor structure, we conducted a maximum-likelihood factor analysis with varimax rotation on Sample 2. Diagnostic tests indicated that a factor model was appropriate for the data (KMO = .763, Bartlett's Test of Sphericity $p < .001$). A four factor solution was indicated by the scree plot. We performed Pearson correlations between the columns of responding factors in the two samples. The correlations for the four factors were: Factor 1, .83; Factor 2, .83; Factor 3, .85; Factor 4, .65 (all $dfs = 23$, $ps < .001$). Coefficients of congruence for the four factors ranged from positive .48 to .88, and off-factor coefficients ranged from −.30 to .45, with a mean of .082. Following procrustes rotation macro in SPSS (McCrae, Zonderman, Costa, Bond, & Paunonen, 1996), the four coefficients of congruency were: .88 (factor 1), 0.87 (factor 2), 0.83 (factor 3) and 0.74 (factor 4). These results showed that sample 1 and sample 2 had an acceptable degree of factor congruence.

TABLE 4
Correlations between language factors and personality traits

	Extraversion	*Agreeableness*	*Conscientiousness*	*Neuroticism*	*Openness*
Making distinction	−.15**	.06	.04	.11*	.04
Reflection	−.01	−.06	−.09*	.07	−.00
Objective description	−.00	.08†	.10*	−.08	−.06
Socialisation	.12*	−.03	−.03	.07	−.01

†$p < .10$; *$p < .05$; **$p < .01$.

GENERAL DISCUSSION

The present study identified associations between Chinese language and personality and suggests that linguistic expression of personality has both universal- and language-specific aspects. Most fundamentally, expression of personality via content words is more likely to be consistent across languages than expression via function words. These findings have important implications.

Firstly, our results show that many word-personality associations are language independent. For example, the correlations between extraversion and social processes, conscientiousness and work, and neuroticism and negative emotion remain consistent between Chinese and English. Furthermore, the correlation coefficients all range between 0.1 and 0.3 in both Chinese and English (e.g., Pennebaker & King, 1999), suggesting modest but consistent effects of personality on word choice across languages. In addition, our factor analysis shows that the psychological meaning of word categories can exhibit some stability across languages. Although Chinese and English have different grammars and vocabularies, they both have an underlying factor that indicates Making Distinction.

Secondly, our study shows that there are personality expressions that are specific to Chinese. In particular, the majority of personality expressions associated with function words in Chinese have not been documented in English. This is likely because function words do not carry real, independent meaning and their use is strongly influenced by the language's grammatical rules. Due to the large linguistic difference between Chinese and English, it is unsurprising that personality expressions in function words would vary considerably across different languages. In contrast, content words are imbued with real meaning, indicating the things people think and do. As universal concepts that are highly likely to be cross-culturally applicable (Brown, 1991), the associations between content words and personality are less likely to manifest in a language-specific manner.

LIMITATIONS

One limitation of the present study is that these two samples were drawn from different cultural contexts: Chinese speakers in mainland China and Chinese-speaking students studying overseas. It is possible that language usage of non-resident Chinese nationals may be influenced by their acculturation into the host nation, and therefore may differ from that of Chinese individuals residing in their home country. However, the consistent personality-language associations and factor structure we found across the two samples suggest that there are personality expressions in Chinese language that are stable without being influenced by acculturation.

Another limitation of our study pertains to the usage of online microblogs to study personality expression. While online data provide us access to large and diverse samples from different locations, they do have the limitation of being biased towards younger, more technically savvy users, as well as possibly being affected by the unique norms and affordances of online communication (Gosling & Mason, 2015). In addition, the low response rate in Sample 1 may indicate that our sample is not representative of the entire online population. Future research is needed to examine if our results can be generalised to other samples.

ACKNOWLEDGEMENT

This work was supported by Singapore Ministry of Education AcRF Tier 1 Grant RGT 37/13 awarded to the first author.

Manuscript received April 2015
Revised manuscript accepted January 2016

REFERENCES

References marked with an asterisk indicate articles included in Table 2.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1–26.

Brown, D. (1991). *Human universals*. New York, NY: McGraw-Hill.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Costa, P., & McCrae, R. (1996). Toward a new generation of personality theories: Theoretical contexts for the five-factor

model. In J. S. Wiggins (Ed.), *The five factor model of personality: Theoretical perspectives* (pp. 51–87). New York, NY: Guilford Press.

*Gill, A., & Oberlander, J. (2003). Perception of email personality at zero-acquaintance. Extraversion takes care of itself; Neuroticism is a worry. Paper presented at the Proceedings of the 25th Annual Conference of the Cognitive Science Society. [1]

*Golbeck, J., Robles, C., & Turner, K. (2011). Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems* (pp. 253–262). ACM. [2]

Gosling, S., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, *66*, 877–902.

*Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of Research in Personality*, *43*, 524–527. [3].

*Holtgraves, T. (2011). Text messaging, personality, and the social context. *Journal of Research in Personality*, *45*, 92–99. [4].

Huang, C.-L., Chung, C. K., Hui, N. H. H., Lin, Y.-C., Seih, Y., Lam, B. C. P., & Pennebaker, J. W. (2012). The development of the Chinese linguistic inquiry and word count dictionary. *Chinese Journal of Psychology*, *54*(2), 185–201.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five inventory – versions 4a and 54*. Berkeley, CA: University of California.

Kashima, E. S., & Kashima, Y. (1998). Culture and language the case of cultural dimensions and personal pronoun use. *Journal of Cross-Cultural Psychology*, *29*(3), 461–486.

*Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, *30*, 457–50. [5].

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81–90.

McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. *American Psychologist*, *52*, 509–516.

McCrae, R. R., Costa, P. T., Jr., del Pilar, G. H., Rolland, J. P., & Parker, W. D. (1998). Cross-cultural assessment of the five-factor model: The Revised NEO Personality Inventory. *Journal of Cross-Cultural Psychology*, *29*, 171–188.

McCrae, R. R., Zonderman, A. B., Costa, P. T., Jr., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory factor analysis versus procrustes rotation. *Journal of Personality and Social Psychology*, *70*(3), 552–566.

*Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, *90*, 862–877. [6].

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of Personality and Social Psychology*, *84*(4), 857–870.

*Nguyen, T., Phung, D., Adams, B., & Venkatesh, S. (2011). Towards discovery of influence and personality traits through social link prediction. Paper presented at the 5th international AAAI conference on weblog and social media. [7].

*Nowson, S. (2006). The language of weblogs: A study of genre and individual differences. Unpublished doctoral dissertation, University of Edinburgh, Edinburgh, UK. [8].

*Oberlander, J., & Gill, A. J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes*, *42*, 239–27. [9].

Paul, L. M., Simons, G. F., & Fennig, C. D. (2015). *Ethnologue: Languages of the world*. Dallas, TX: SIL International.

*Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language us as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296–1312. [10].

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547–577.

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, *85*, 291–301.

*Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*, *46*, 710–718. [11].

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54.

*Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*, 363–373. [12].

*Yee, N., Harris, H., Jabon, M., & Bailenson, J. N. (2011). The expression of personality in virtual worlds. *Social Psychological and Personality Science*, *2*(1), 5–12. [13].

Zhang, H. P., Liu, Q., Cheng, X. Q., Zhang, H., & Yu, H. K. (2003). Chinese lexical analysis using hierarchical hidden markov model. Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17 (63–70).