Developing Simplified Chinese Psychological Linguistic Analysis Dictionary for Microblog

Rui Gao[1], Bibo Hao[1], He Li[2], Yusong Gao[1], and Tingshao Zhu[1]

[1] Institute of Psychology, University of Chinese Academy of Sciences
Chinese Academy of Sciences, Beijing 100190, P.R. China
[2] National Computer System Engineering Research Institute of China
Beijing, 100083, P.R. China
tszhu@psych.ac.cn,
{gaorui11,haobibo12}@mails.ucas.ac.cn

Abstract. The words that people use could reveal their emotional states, intentions, thinking styles, individual differences, etc. LIWC (Linguistic Inquiry and Word Count) has been widely used for psychological text analysis, and its dictionary is the core. The Traditional Chinese version of LIWC dictionary has been released, which is a translation of LIWC English dictionary. However, Simplified Chinese which is the world's most widely used language has subtle differences with Traditional Chinese. Furthermore, both English LIWC dictionary and Traditional Chinese version dictionary were both developed for relatively formal text. Microblog has become more and more popular in China nowadays. Original LIWC dictionaries take less consideration on microblog popular words, which makes it less applicable for text analysis on microblog. In this study, a Simplified Chinese LIWC dictionary is established according to LIWC categories. After translating Traditional Chinese dictionary into Simplified Chinese, five thousand words most frequently used in microblog are added into the dictionary. Four graduate students of psychology rated whether each word belonged in a category. The reliability and validity of Simplified Chinese LIWC dictionary were tested by these four judges. This new dictionary could contribute to all the text analysis on microblog in future.

Keywords: LIWC, Traditional Chinese, Simplified Chinese, microblog, text analysis.

1 Introduction

The rapid developing social media--microblog has had a significant impact on society, politics, economy, culture and people's daily life [1, 2]. Researchers have carried out a number of studies on microblog [3-7]. Computerized text analysis methods like LIWC (Linguistic Inquiry and Word Count) [8, 9] have been widely used for social media researches [2, 10-12]. LIWC dictionary is the core of LIWC text analysis method [8, 9, 13].

K. Imamura et al. (Eds.): BHI 2013, LNAI 8211, pp. 359–368, 2013.

Simplified Chinese now is the world's most widely used language, but it cannot be analyzed with LIWC because of the vacancy of Simplified Chinese version of dictionary. The Traditional Chinese version of LIWC dictionary — CLIWC(Chinese Linguistic Inquiry and Word Count) [14] dictionary has been released, which makes it possible to analyze Traditional Chinese text with LIWC software. But, Simplified Chinese has subtle differences with Traditional Chinese. Furthermore, both English LIWC dictionary and CLIWC dictionary were both developed for relatively formal text.

In this study, specific exclusive Simplified Chinese LIWC dictionary (SCLIWC) was established according to LIWC dictionary and CLIWC dictionary, and then microblog high frequency words were added into SCLIWC. This dictionary, SCMBWC (Simplified Chinese Microblog Word Count) is a promising approach for both psychological and other kinds of researches based on Microblog.

The rest of this paper is organized as follows. In Section 2, we overview some related work. Section 3 describes how to build the dictionary. The experimental results and discussion are presented in Section 4, followed by the conclusion and future work in Section 5.

2 Related Work

LIWC with its English dictionary is one of the most prestigious tools of content analysis [15]. First significant version of LIWC was released in 1997, after continuing optimizing for decade the latest version of LIWC software and English dictionary is LIWC2007 [9]. LIWC is a milestone in the history of computerized text analysis, and plenty of researches are based on LIWC [16-20].

Establishment of CLIWC made it possible to use computerized text analysis methods in Traditional Chinese text analysis related researches. CLIWC has made an outstanding contribution to Traditional Chinese content analysis area [14].

Traditional Chinese and Chinese Simplified share the same origin; however, along with the development of the times, diversity has been evolved between them [21]. Many Traditional Chinese words, cannot find a unique identifying Chinese Simplified word correspond with it. Figure 1 shows some examples of this kind of words. Furthermore, words spelled the same in these two languages might express dissimilar meanings [22, 23]. More crucial is, compared to differences of the two languages itself, linguistic using differences in their populations merited to be taken into serious consideration [13, 21, 24].

入學考　阿妈　米田共
阿公　俗辣　娘卡好

Fig. 1. Examples of Word could not find unique corresponding Chinese Simplified word

Chinese Simplified population and Users of Traditional Chinese share the same origin. But, because the diverse social ideologies and distinct living environments, two populations have gradually produced a lot of differences on the language usage in the past over 60 years. Language usage differences is a major challenge to building intercultural LIWC dictionaries [13], which represent word count based computerized text analysis research method.

Therefore, It is imperative that Simplified Chinese LIWC dictionary (SCLIWC) should to be established. It is the basic requirements to apply word count based text analysis method into Chinese Simplified.

3 Method

3.1 Development of Simplified Chinese LIWC (SCLIWC)

There are computer programs which could try to translate Traditional Chinese into Chinese Simplified [25, 26], and vice versa. But SCLIWC dictionary as a promising Chinese Simplified text analysis approach, subtle translation deflection introduced by programs might lead to extra unessential deviations which cannot be ignored in further researches.

In order to best guarantee the efficiency of SCLIWC dictionary, each lexical item were checked and validated manually. Twenty-one graduate students from University of Chinese Academy of Science were recruited to develop SCLIWC dictionary. They are all native speaker of Simplified Chinese.

Firstly, 21 judges were divided into three groups averagely. Each group independently processed CLIWC [14] lexical items one by one, and generate response Simplified Chinese lexical items. These generated items have the closest meaning with Traditional Chinese lexical items and conform to the language usage habits of Chinese Mainland population. Subjecting to majority rule, for group disagreements with lexical items, all members discussed and voted to make the final decision. Eventually each group delivered their version of SCLIWC.

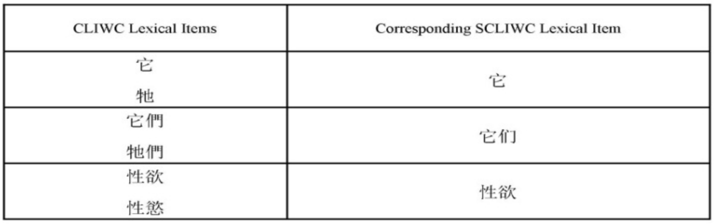
Secondly, another three judges (also native speaker of Simplified Chinese) who are familiar with the LIWC dictionary framework (including authors of this article) validated these three versions of SCLIWC. If the three versions differed on specific lexical items, judges discussed and voted according to majority rule.

Finally, there are some different Traditional Chinese words correspond with the same Chinese Simplified word. Some lexical items in SCLIWC were merged. Instances of more than one lexical item in CLIWC share the same word (the same Chinese characters) in SCLIWC were shown in Table 1.

3.2 Sina Microblog High Frequency Words Selection

Based on Sina micro-blog platform, we have developed an application--mental map. By calling the Sina microblog API, through this application basic information (exclusive microblog statuses) of 99,925,821 users were collected. We adopted the following rules to filter 99,925,821 users:

Table 1. Examples of merged Lexical Items.



CLIWC Lexical Items	Corresponding SCLIWC Lexical Item
它 牠	它
它們 牠們	它们
性欲 性慾	性欲

1. Users who published no status in recent three months or posted less than 512 statuses in total were excluded.
2. Users who publish more than 40 statuses every day are much likely to be advertisement users or entertainment star users. They were excluded, too.

After filtering, An ID list was generated including 1,953,485 microblog active users whose microblog statuses texts are appropriate for scientific research. By calling the Sina microblog API, these users' statuses texts were completely downloaded. From these 1,953,485 users, two groups of samples were randomly selected. Each group consists of 10,000 users, 20,000 in total. NLPIR2013 (ICTCLAS2013) system[27, 28] is one of the most widely used word parser in studies about Chinese language. NLPIR2013 was used for Chinese word segmentation in this study. Microblog statuses of users in both groups were parsed and stop words were filtered. Main stop words which are related to linguistic psychological characteristics had been included in SCLIWC dictionary, so stop words were excluded when selecting microblog high frequency words. High frequency words were selected according to the following steps:

Firstly, both groups' user statuses texts were separately calculated to get two sets of top 5,000 high frequency words in each group. We name these two word sets S1 and S2. Then, we merged the two groups' user statuses text, and calculated the set of top 5,000 high frequency words in this merged group. We name this word set S3. Table 2 shows the overlap of this three word sets. S1 and S2 have more than 84% high frequency words in common. S1 and S2 respectively have 91.62% and 93.04% the same words with S3. The overlaps indicated that both sample groups we randomly picked could represent high frequency words used in Sina microblog environment.

Finally, excluding stop words and words already in SCLIWC dictionary, top 5000 high frequency words of the merged group were selected as candidates for SCMBWC dictionary. In Figure 2, word frequency rates of top five thousand high frequency words were shown. The total word count of twenty thousand Sina microblog users is 832737854. Word frequency rate of a specific word equals the times this word appears in this whole texts materials dived 832737854, then plus 10000. The word frequency rates subject to long tail distribution. Therefore, top five thousand high frequency words could cover the major part of words which frequently appears in Sina microblog statuses. Figure 3 gives the list top one hundred words of the most high frequency words in Sina microblog.

Table 2. High frequency word sets overlap counts

	S1	S2	S3
S1		4204	4581
S2	4204		4652
S3	4581	4652	

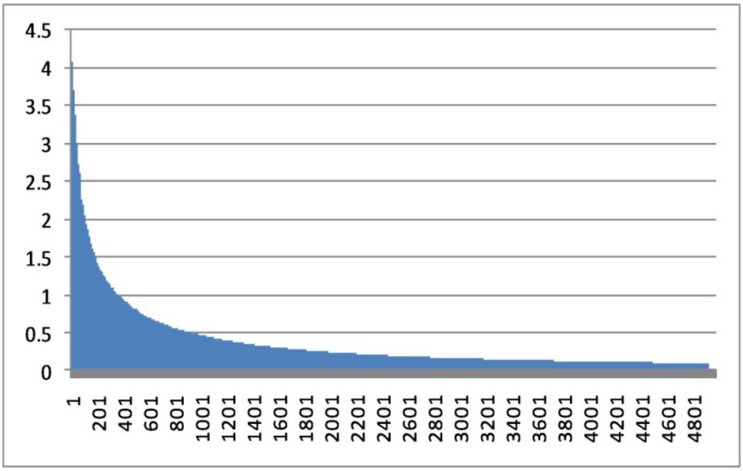


Fig. 2. Word Frequency Rate Distribution

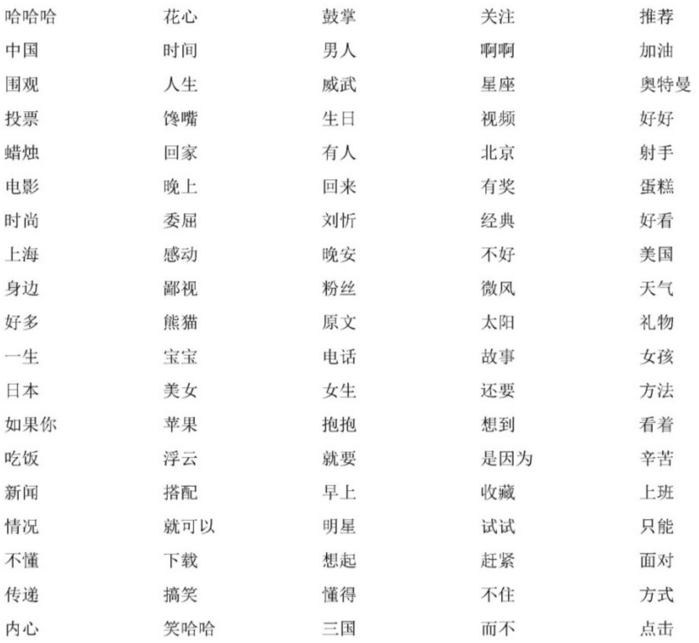
3.3 SCMBWC Dictionary Development, Internal Reliability and External Validity

The development of SCMBWC dictionary can be divided into the following three steps.

Step One: assigning high frequency words into SCLIWC categories. Four Psychology PhD candidates from Institute of Psychology Chinese Academy of Science were recruited as judges. First of all, they independently assign Sina microblog high frequency words into SCLIWC categories.

Step Two: judges' rating phase. After four version of category word lists were amassed, SCMBWC dictionary category scales were established subject to following set of rules:

1. If more than two judges' version of category word lists support a word to fall into this category, the word fall into this category.
2. If two judges' version of category word lists support a word to fall into this category, but another two were against. Four judges discussed this word, and then voted again. Only if new polls indicated that more than 2 judges considered that the word belongs to this category, then the word fall into this category. Otherwise, this word was abandoned.



哈哈哈	花心	鼓掌	关注	推荐
中国	时间	男人	啊啊	加油
围观	人生	威武	星座	奥特曼
投票	馋嘴	生日	视频	好好
蜡烛	回家	有人	北京	射手
电影	晚上	回来	有奖	蛋糕
时尚	委屈	刘忻	经典	好看
上海	感动	晚安	不好	美国
身边	鄙视	粉丝	微风	天气
好多	熊猫	原文	太阳	礼物
一生	宝宝	电话	故事	女孩
日本	美女	女生	还要	方法
如果你	苹果	抱抱	想到	看着
吃饭	浮云	就要	是因为	辛苦
新闻	搭配	早上	收藏	上班
情况	就可以	明星	试试	只能
不懂	下载	想起	赶紧	面对
传递	搞笑	懂得	不住	方式
内心	笑哈哈	三国	而不	点击

Fig. 3. Top 100 High Frequency Words in Sina Microblog

Step Three: another three judges who are familiar with the SCMBWC dictionary framework (including authors of this article) rating SCMBWC dictionary categories focus on inclusion and exclusion. Internal reliability and external validity were rated according to following steps. Sub step one, five categories word lists were randomly picked: Ingest, Certain, Space, Leisure, religion. Sub step two, for each word in this five categories, judges rated whether this word belong to current category or not. Only if two or more judges agreed to keep the word in current category, the word remained. Otherwise, the word was removed from the scale list. Sub step three. Judges rated the discrimination of SCMBWC dictionary category lexical items. They voted whether words in a high level category belong to sub level categories.

In process of developing SCLIWC, three judges' agreement is about 94%. The percentages of three judges' agreement for the sub step two and three in SCMBWC development were over 95%.

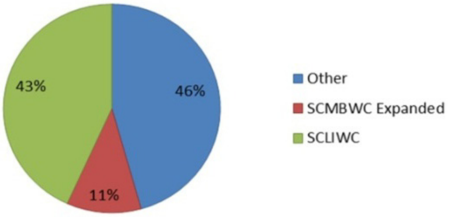


Fig. 4. The percentages of words captured by the dictionary

4 Result

Two thousand users were randomly picked from 1,953,485 microblog active users. We respectively process their status texts via LIWC2007 software with SCLIWC and SCMBWC dictionary. Figure 4 shows the percentages of words captured by SCLIWC and SCMBWC dictionary in total word counts. SCMBWC dictionary improve the words captured by dictionary by about eleven percent. In average of each user words captured by SCLIWC and SCMBWC dictionary are 43.56% for SCLIWC and 54.68% for SCMBWC. The improvements of each specific user's status texts are shown in Figure 5. For every single user, apparently, many more words he or she used in microblog statuses were recognized by SCMBWC dictionary. In table 3, psychological and personal concern categories features average and standard deviation are listed. It's obviously that SCMBWC dictionary covers higher proportion of psychological and personal concern related words. Therefore, more information could be able to extract from microblog text content for each user. That might possibly contributes to further knowledge discovery in social media web sites.

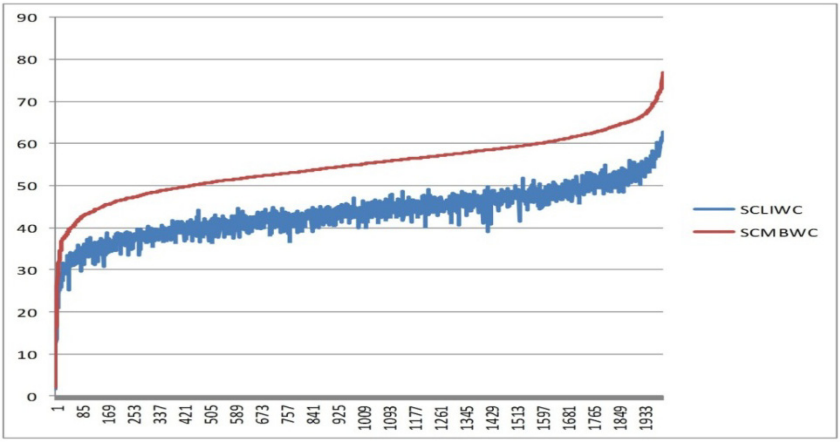


Fig. 5. The percentages of words captured by SCLIWC and SCMBWC dictionary for each user

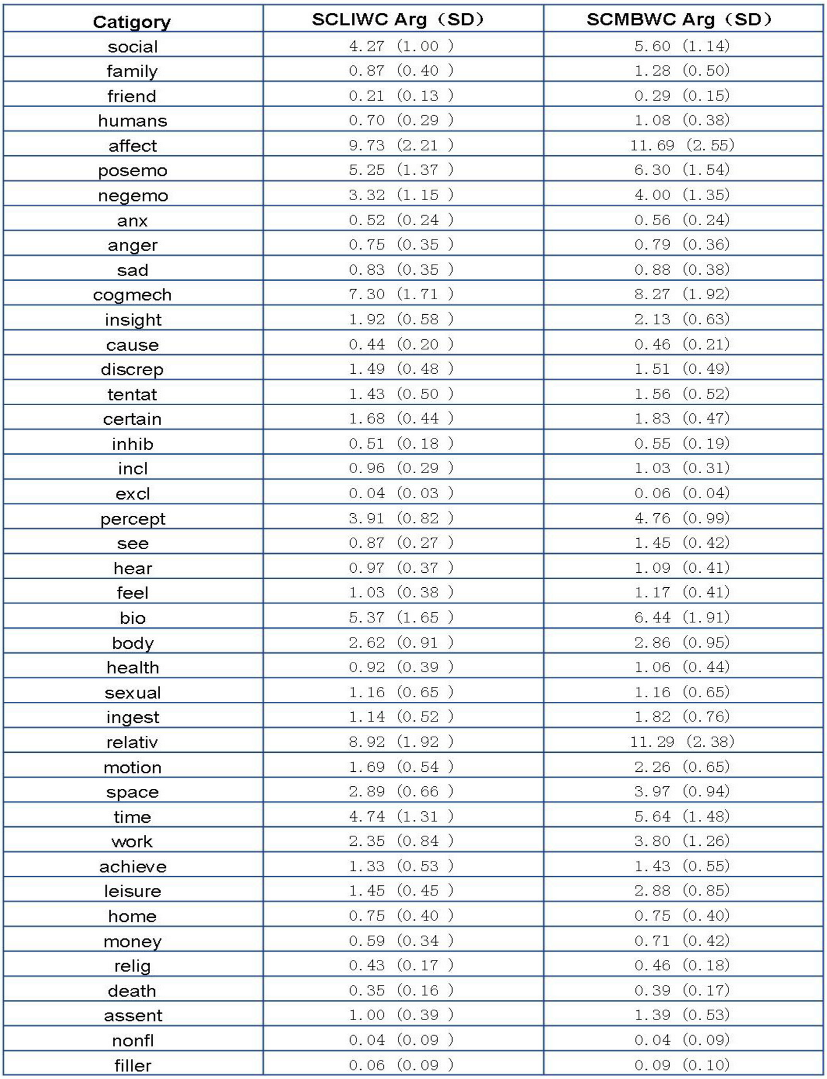
While using LIWC2007 to process Chinese content, we found that it was designed for western language, and cannot process Chinese content appropriately sometimes. We have implemented a prototype system TextMind that is optimum for processing Simplified Chinese. Using SCLIWC and SCWBWC, TextMind works effectively with high performance. TextMind provides an all-in-one solution for Simplified Chinese analysis, and we intend to release it after thoroughly testing.

5 Conclusion

Percentage of words captured by the SCLIWC dictionary indicates that words usage in internet environment like Sina microblog are much more diverse compared to formal text materials[9, 14]. Percentage of words captured by the SCMBWC dictionary improves above 10 percent, especially captured more words in category of psychological processes and its sub categories, such as social processes, affective processes, cognitive processes and etc. Internal Reliability and External Validity of those two dictionaries are well guaranteed by four groups of judges.

SCLIWC bridges the gap between LIWC software and Simplified Chinese. What is more, SCMBWC suggests a promising approach for further text analysis of Chinese Simplified in various internet environments.

Table 3. Category Features Average and Standard Deviation of 2000 users



Catigory	SCLIWC Arg（SD）	SCMBWC Arg（SD）
social	4.27 (1.00)	5.60 (1.14)
family	0.87 (0.40)	1.28 (0.50)
friend	0.21 (0.13)	0.29 (0.15)
humans	0.70 (0.29)	1.08 (0.38)
affect	9.73 (2.21)	11.69 (2.55)
posemo	5.25 (1.37)	6.30 (1.54)
negemo	3.32 (1.15)	4.00 (1.35)
anx	0.52 (0.24)	0.56 (0.24)
anger	0.75 (0.35)	0.79 (0.36)
sad	0.83 (0.35)	0.88 (0.38)
cogmech	7.30 (1.71)	8.27 (1.92)
insight	1.92 (0.58)	2.13 (0.63)
cause	0.44 (0.20)	0.46 (0.21)
discrep	1.49 (0.48)	1.51 (0.49)
tentat	1.43 (0.50)	1.56 (0.52)
certain	1.68 (0.44)	1.83 (0.47)
inhib	0.51 (0.18)	0.55 (0.19)
incl	0.96 (0.29)	1.03 (0.31)
excl	0.04 (0.03)	0.06 (0.04)
percept	3.91 (0.82)	4.76 (0.99)
see	0.87 (0.27)	1.45 (0.42)
hear	0.97 (0.37)	1.09 (0.41)
feel	1.03 (0.38)	1.17 (0.41)
bio	5.37 (1.65)	6.44 (1.91)
body	2.62 (0.91)	2.86 (0.95)
health	0.92 (0.39)	1.06 (0.44)
sexual	1.16 (0.65)	1.16 (0.65)
ingest	1.14 (0.52)	1.82 (0.76)
relativ	8.92 (1.92)	11.29 (2.38)
motion	1.69 (0.54)	2.26 (0.65)
space	2.89 (0.66)	3.97 (0.94)
time	4.74 (1.31)	5.64 (1.48)
work	2.35 (0.84)	3.80 (1.26)
achieve	1.33 (0.53)	1.43 (0.55)
leisure	1.45 (0.45)	2.88 (0.85)
home	0.75 (0.40)	0.75 (0.40)
money	0.59 (0.34)	0.71 (0.42)
relig	0.43 (0.17)	0.46 (0.18)
death	0.35 (0.16)	0.39 (0.17)
assent	1.00 (0.39)	1.39 (0.53)
nonfl	0.04 (0.09)	0.04 (0.09)
filler	0.06 (0.09)	0.09 (0.10)

Acknowledgements. The authors gratefully acknowledge the generous support from National High-tech R&D Program of China (2013AA01A606), NSFC (61070115), Key Research Program of CAS (KJZD-EW-L04), Strategic Priority Research Program (XDA06030800) and 100-Talent Project (Y2CX093006) from Chinese Academy of Sciences.

References

1. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences 110(15), 5802–5805 (2013)
2. Tumasjan, A., et al.: Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In: ICWSM, pp. 178–185 (2010)
3. Ding, X., et al.: De-anonymizing Dynamic Social Networks. In: 2011 IEEE Global Telecommunications Conference, Globecom 2011 (2011)
4. Ebner, M., et al.: Microblogs in Higher Education - A chance to facilitate informal and process-oriented learning? Computers & Education 55(1), 92–100 (2010)
5. Eysenbach, G.: Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. Journal of Medical Internet Research 11(1) (2009)
6. Jansen, B.J., et al.: Twitter Power: Tweets as Electronic Word of Mouth. Journal of the American Society for Information Science and Technology 60(11), 2169–2188 (2009)
7. Narayanan, A., Shmatikov, V.: De-anonymizing Social Networks. In: Proceedings of the 2009 30th IEEE Symposium on Security and Privacy, pp. 173–187 (2009)
8. Pennebaker, J.W., et al.: The Development and Psychometric Properties of LIWC2007 (2007)
9. Tausczik, Y.R., Pennebaker, J.W.: The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology 29(1), 24–54 (2010)
10. Choy, M.: Effective Listings of Function Stop words for Twitter (IJACSA) International Journal of Advanced Computer Science and Applications 3(6), 8–11 (2012)
11. Golbeck, J., Robles, C., Turner, K.: Predicting personality with social media. In: CHI 2011 Extended Abstracts on Human Factors in Computing Systems, pp. 253–262. ACM, Vancouver (2011)
12. Golbeck, J., Robler, J., Edmondson, M., Turner, K.: Predicting Personality from Twitter. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), Boston, USA, pp. 149–156 (2011)
13. Piolat, A., et al.: The French dictionary for LIWC: Modalities of construction and examples of use. Psychologie Francaise 56(3), 145–159 (2011)
14. Huang, C.-L., et al.: The Development of the Chinese Linguistic Inquiry and Word Count Dictionary. Chinese Journal of Psychology 55(2), 185–201 (2012)
15. Lowe, W.: Software for content analysis–A review (2013)
16. Borelli, J.L., et al.: Experiential connectedness in children's attachment interviews: An examination of natural word use. Personal Relationships 18(3), 341–351 (2011)
17. Ireland, M.E., Pennebaker, J.W.: Language Style Matching in Writing: Synchrony in Essays, Correspondence, and Poetry. Journal of Personality and Social Psychology 99(3), 549–571 (2010)

18. Ireland, M.E., et al.: Language Style Matching Predicts Relationship Initiation and Stability. Psychological Science 22(1), 39–44 (2011)
19. Tumasjan, A., et al.: Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape. Social Science Computer Review 29(4), 402–418 (2011)
20. Zehrer, A., Crotts, J.C., Magnini, V.P.: The perceived usefulness of blog postings: An extension of the expectancy-disconfirmation paradigm. Tourism Management 32(1), 106–113 (2011)
21. Peng, G., Minett, J.W., Wang, W.S.Y.: Cultural background influences the liminal perception of Chinese characters: An ERP study. Journal of Neurolinguistics 23(4), 416–426 (2010)
22. Chung, F.H.-K., Leung, M.-T.: Data analysis of Chinese characters in primary school corpora of Hong Kong and mainland China: preliminary theoretical interpretations. Clinical Linguistics & Phonetics 22(4-5), 379–389 (2008)
23. Chung, W.Y., et al.: Internet searching and browsing in a multilingual world: An experiment on the Chinese Business Intelligence Portal (CBizPort). Journal of the American Society for Information Science and Technology 55(9), 818–831 (2004)
24. Ramirez-Esparza, N., et al.: The psychology of word use: A computer program that analyzes texts in Spanish. Revista Mexicana De Psicologia 24(1), 85–99 (2007)
25. Akers, G.A.: LogoMedia TRANSLATE (TM), version 2.0. In: Richardson, S.D. (ed.) Machine Translation: From Research to Real Users, pp. 220–223 (2002)
26. Al-Dubaee, S.A., Ahmad, N.: New Direction of Applied Wavelet Transform in Multilingual Web Information Retrieval. In: Fifth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2008 (2008)
27. Zhang, H.-P., et al.: Chinese lexical analysis using hierarchical hidden markov model. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, vol. 17. Association for Computational Linguistics (2003)
28. Zhang, H.-P., et al.: HHMM-based Chinese lexical analyzer ICTCLAS. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, vol. 17. Association for Computational Linguistics (2003)