Exploring Mental Lexicon in an Efficient and Economic Way: Crowdsourcing Method for Linguistic Experiments

Shichang Wang, Chu-Ren Huang, Yao Yao, Angel Chan
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
shi-chang.wang@connect.polyu.hk
{churen.huang, y.yao, angel.ws.chan}@polyu.edu.hk

Abstract

Mental lexicon plays a central role in human language competence and inspires the creation of new lexical resources. The traditional linguistic experiment method which is used to explore mental lexicon has some disadvantages. Crowdsourcing has become a promising method to conduct linguistic experiments which enables us to explore mental lexicon in an efficient and economic way. We focus on the feasibility and quality control issues of conducting Chinese linguistic experiments to collect Chinese word segmentation and semantic transparency data on the international crowdsourcing platforms Amazon Mechanical Turk and Crowdflower. Through this work, a framework for crowdsourcing linguistic experiments is proposed.

1 Introduction

Mental lexicon as a theoretical construct has two important implications. For an individual, it is where all the grammatical and world information is stored and organized to enable speech. For a group of speakers of the same language, however, the mental lexicon is a shared knowledge structure allowing speakers to process and understand what each other said. WordNets, for example the English WordNet (Miller, 1995) and the Chinese WordNet (CWN) (Huang et al., 2003), and ontologies, for example the Suggested Upper Merged Ontology (SUMO) (Niles and Pease, 2001) and the Sinica BOW (Huang et al., 2010), have been proposed as a representational framework for this shared mental lexicon; and psycho- and neuro-linguistic experiments have been designed to explore how individuals access their mental lexicon. However, the question of whether there is a shared principle or strategy of mental lexicon by all speakers of the same language was never seriously studied as the cognitive experimental paradigm does not allow manipulation of a large number of subjects simultaneously. In this paper, we explore the possibility of conducting lexical access related experiments through crowdsourcing. With the crowdsourcing experiments, we intend to ask specific question about the share strategy of determination of lexical units, as well as determination of semantic transparencies, two issues that would have direct implications of how individuals access their mental lexicon.

Many scholars discuss applying crowdsourcing method to language resource construction recent years (Snow et al., 2008; Callison-Burch and Dredze, 2010; Munro et al., 2010; Gurevych and Zesch, 2013). Crowdsourcing has been proved to be an efficient tool to build lexical resources, for example, Wiktionary, whose goal is to become the free online dictionary for all the words in all languages; Biemann (2013) presents another example which creates the Turk Bootstrap Word Sense Inventory for 397 frequent nouns from scratch using Amazon Mechanical Turk. And there is more and more literature focusing on conducting experiments on crowdsourcing platforms (Schnoebelen and Kuperman, 2010; Paolacci et al., 2010; Berinsky et al., 2011; Rand, 2012; Mason and Suri, 2012; Crump et al., 2013). Using crowdsourcing method, it is easier to access highly diverse and huge amount of participants, so it is possible to obtain more representative language behavioral data. The anonymous nature of crowdsourcing makes the participants more open to contribute sensitive data. And Crowdsourcing experiments are usually much faster and cheaper than laboratory experiments which enables " faster iteration between developing theory and

Zock/Rapp/Huang (eds.): Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon, pages 105–113,
Dublin, Ireland, August 23, 2014.

executing experiments" (Mason and Suri, 2012). It can be a promising tool to explore mental lexicon in an efficient and economic way.

MTurk and Crowdflower are perhaps the two most important MTurk-like crowdsourcing platforms. MTurk is the platform appears most frequently in the literature, so popular that it represents a major genre of crowdsourcing and its name has become the name of that genre. Crowdflower is a rapid developing platform and is drawing more and more attention. Although they are both MTurk-like platforms, they differ from each other. On the MTurk platform, invalid responses submitted can be manually rejected which is a very convenient quality control method; however Crowdflower has a much larger worker pool than MTurk. Since MTurk is one channel of Crowdflower and Crowdflower can access the worker pool of MTurk[1], besides MTurk, Crowdflower has several dozens of other channels to which it can distribute tasks. More importantly, Crowdflower is more accessible to requesters outside the U.S. (MTurk does not support requesters outside the U.S. by now). Crowdflower basically doesn't support manual rejection of invalid responses but it integrates an effective quality control method named *Test Questions* which uses predefined gold standard questions to measure the quality of contributions of workers and screens low quality workers automatically in order to produce high quality data. Unfortunately, it is not suitable to our task, for it requires multiple submissions form a worker. Neither MTurk nor Crowdflower is a native Chinese crowdsourcing platform, so we can suppose that native Chinese speakers can only occupy a small proposition in their worker pools, in this case, a larger worker pool means higher possibility of successful data collection.

We have two objectives in this study: (1) to check if it is feasible to conduct Chinese language experiments to collect Chinese word segmentation and semantic transparency (Libben, 1998) data which can be used to explore the mental lexicon of Chinese speakers on international crowdsourcing platforms, such as Amazon Mechanical Turk (MTurk) and Crowdflower; (2) to identify and solve some quality control and experimental design issues in order to obtain high quality data and to establish a preliminary framework for crowdsourcing linguistic experiments.

2 Initial Calibration Tests

Before the experiment, we conducted initial calibration tests. The purpose is to lay a basic foundation (e.g., general experimental parameters, quality control methods, etc.) for the experiment. There are four tests. We employed a problem-driven bootstrapping strategy in the design and conduct of these tests in order to accumulate knowledge effectively. The repeated procedure is like this: one test is started and once a problem has been identified, the test will be paused or stopped; after a proper solution has been found, a modified version of that test will be resumed or a new test will be designed and started.

2.1 Parameters

Both MTurk and Crowdflower will be tested, however we cannot access MTurk directly as a requester since it doesn't support requesters outside the U.S. by now. Because MTurk is a channel of Crowdflower, Crowdflower can distribute jobs to it, we can access it indirectly through Crowdflower. So Crowdflower is selected as our job publishing platform. When we test MTurk, we will require Crowdflower to distribute our jobs to the MTurk channel only; and when we test Crowdflower, we will conduct Crowdflower to distribute our jobs to all of its channels.

The jobs published on Crowdflower can be divided into two types according to the existence or absence of data-sets to be processed. A job without a data-set to be processed is a survey. The survey type fits our objectives best since we want to collect data from different individuals and each person can only participate in any one of the tests/experiments once. In order to ensure this one-time participation, we use the following constrains: (1) each worker account can only submit one response, and (2) each IP address can only submit one response.

Crowdflower allows us to specify 'included countries' or 'excluded countries' as an access control method. We only use 'included countries' in our tests/experiments, only the countries and regions we

[1]However, this has become history, in December 2013, Crowdflower announced that Amazon Mechanical Turk would no longer be a partner channel, see http://www.crowdflower.com/blog/2014/01/crowdflower-drops-mechanical-turk-to-ensure-the-best-results-for-its-customers (Retrieved May 24, 2014).

selected are allowed to access our jobs. According to the distribution of Chinese people, we select the following countries and regions: Mainland China, Hong Kong, Macau, Taiwan, Singapore, Malaysia, Japan, Korea, Australia, Canada, the United States, the United Kingdom, France, Germany, Russia, and New Zealand.

2.2 Test 1

The objective is to evaluate the feasibility to collect Chinese language data from MTurk. We published a survey on Crowdflower and it was distributed to the MTurk channel only. The questionnaire contains 3 questions: (1) what place of China do you come from, (2) what country or region are you in now, (3) what dialect of Chinese do you speak; all the questions are in Chinese. There is a text-box for each question which allows the participants to input their answers. All the questions must be answered or the data are not allowed to be submitted. It only takes 10 to 15 seconds to fill up the questionnaire. The unit price of this survey is one cent. This test only collected 2 responses in 21 hours. Judged from the speed, we can preliminarily conclude that MTurk is not a feasible platform for Chinese language data collecting tasks. Because of the properties of crowdsourcing environment, this result can be accidental, so we will continue to open the MTurk channel to validate this result.

2.3 Test 2

We published a new test on Crowdflower in order to evaluate the feasibility of collecting Chinese language data from Crowdflower. It is mostly the same as Test 1, the only difference is that all the channels of Crowdflower are enabled so we can access a much larger worker pool. This time, we collected 23 responses in 2 hours. Nine of them (39.1%) are valid, 14 (60.9%) are invalid. The speed is good, but the data quality is not acceptable. In this test, we didn't use powerful quality control method, thus large number of invalid responses were submitted. Invalid responses deteriorated data quality. This demonstrates that quality control is essential to crowdsourcing practices.

2.4 Test 3

It's important to detect and identify invalid responses. "Checkpoint questions" (see section 5) can be used to distinguish valid responses from invalid responses. This test attempts to test the effectiveness of checkpoint questions. We added a Chinese character identification question to the questionnaire of Test 2. The participants are required to identify a Chinese character in a picture and then input this character into a text-box. The frequency of that character is very high, so it is easy for Chinese native speakers to identify. Because this an open-ended question, so it is robust enough. This question satisfies the conditions to be a checkpoint question (see section 5). Then the test was resumed. After 2 hours, 9 new responses were received (23 responses had been received since Test 2). Four of them (44.4%) are valid responses, 5 (55.6 %) are invalid. All of the responses with correct answers to the checkpoint question were checked to be valid responses.

Logically, correctly answering checkpoint questions doesn't definitely mean the other questions are also carefully answered. But human behaviors have certain consistency to some extent. If they carefully answered checkpoint questions, then they are likely to answer the other questions carefully. Although it is not 100% reliable to identify invalid/valid responses by checkpoint questions, it is acceptable if there is no better method.

2.5 Test 4

Checkpoint questions can identify invalid responses but they cannot block them. We can set some conditions for the submission of responses. Only the responses which satisfy these conditions can be submitted. We call these submission conditions "validations" (see section 5). Since checkpoint questions can be used to identify invalid responses, we can set validations on them in order to block the submission of invalid responses. We set a validation on the Chinese character identification checkpoint question: the response can be submitted only when the checkpoint question is correctly answered. Then the test was resumed and 28 new responses were received (a total of 60 responses had been received since Test 2). 26 of them (92.9%) are valid responses, 2 (7.1%) are invalid. Before the adoption of validation, the proportion of

valid response is only 40.6%, after the adoption, it's 92.9%. This basically shows that it can effectively block invalid responses to set validation on checkpoint questions.

2.6 Summary

We collected 60 responses in Tests 1 to 4; among them, there are only two responses from MTurk. This verified the result of Test 1, i.e., it is not quite feasible to collect Chinese language data from MTurk at least by now. However Crowdflower is a feasible choice since it has a much larger worker pool. Because of the nature of Crowdsourcing, noise is everywhere. It is practically unacceptable to collect data without effective quality control methods; otherwise more invalid responses than valid ones will be received. Checkpoint questions can be used to identify valid and invalid responses. Validations are effective to block the submission of invalid responses. It is a good strategy to set validations on checkpoint questions in order to block invalid responses.

3 Experiment

The experiment was divided into two stages, and there was a time interval of about two months between them. Based on the initial calibration tests, the experiment was conducted to test the feasibility of collecting Chinese language data on international crowdsourcing platforms and to identify and solve some quality control and experimental design issues.

Our original plan was to conduct one experiment to collect a sample of 200 responses. But after we had collected 135 responses, we found a serious spammer problem which must be properly solved otherwise the data quality would be greatly threatened and the feasibility of our task would be questionable. Meanwhile, we found the amounts of responses from the region of mainland China and the channel "bitcoinget" were unexpectedly large, we doubted that it might result from the frequent media reports on bitcoin at that time in mainland China. When the media reports ebbed, would the experiment be replicable? Thus we thought it's necessary to pause the experiment to seek a solution for the spammer problem and to evade the strong external factor of media report. Thus the experiment was divided into two stages. We chose to pause the experiment instead of stopping it so that the participants who had already taken part in the experiment (Stage 1) could not take part again when the experiment was resumed (Stage 2). The experiment was resumed after two months, with a spammer monitor program based on the API of Crowdflower which could detect and combat spammers automatically. Other aspects of the experiment remained unchanged. The Stage 2 experiment could be used to check the experimental repeatability and to solve the spammer problem found in Stage 1.

3.1 Experimental Design

Questionnaire

The experiment we ran was a self-paced online questionnaire, consisting of 46 questions divided into three parts. The first part contained 10 screening questions designed to verify that the participants were (1) human and (2) native Chinese speakers. The second part of the experiment was a task of Chinese word segmentation. The participants were presented with 12 Chinese sentences and their task was to put a "/" sign at the word boundary that they perceived. The third part of the experiment was a semantic transparency data collection task. Semantic similarity rating tasks were used to obtain semantic transparency data. 12 di-morphemic Chinese compounds, (e.g., 帮助 *bangzhu*, help-assist, "help") were shown in 12 carrier sentences (one target compound per carrier sentence). The participant's task was to rate, on a 5-point scale, the degree of semantic similarity between the meaning of each character in the target compound and the meaning when it is used alone. In view of the different character systems used in different Chinese-speaking regions, we implemented two versions of the questionnaire: a simplified Chinese character version for participants from Mainland China and a traditional Chinese character version for participants from Hong Kong.

Experiment Control

Experiment control measures are used to ensure the validity of participants and their participations. Because we cannot access the real identities of the participants, we can only use some indirect methods

which are not completely reliable but can satisfy our demands at large. Firstly, all the participants must be native Chinese speakers. The questionnaire was displayed in Chinese characters which can be a natural barrier to non-native Chinese speakers. Ten screening questions are designed in the questionnaire to test the language backgrounds of the participants. By the above measures, we can effectively discriminate native Chinese speakers from non-native ones. Chinese learners are not a major threat due to their small amount and low overall Chinese fluency. We invited two Chinese learners to test our questionnaire; neither of them could finish it. Secondly, one participant can only submit one response. We used the methods which are already explained in 2.1: one account can only submit one response; one IP address can only submit one response.

Quality Control

In addition to experiment control measures, quality control measures are used to further prevent invalid responses. We used checkpoint questions and other measures for data validation. Only those responses that fulfill the following conditions were considered as valid responses: (1) the screening questions in Part 1 were correctly answered, (2) the answers in Part 2 followed the correct format, and (3) the completion time was equal or greater than 5 minutes. Those that failed one or more conditions were considered as invalid. The effectiveness of the validation measures is discussed in 5. After the Stage 1 experiment, we found a serious spammer problem. After adopting the above quality control measures, spammers became the biggest threat to data quality. It can be exhausted to combat spammers manually due to their high speeds and randomness. Thus, based on the API of Crowdflower we wrote a spammer monitor program to detect and combat spammers automatically.

Parameters

The experiment uses the parameters described in 2.1. Besides that, the unit price of our task is set to US$0.25. Pricing strategy should be carefully chosen in crowdsourcing practices. High prices tend to attract cheating, but low prices may fail to attract enough participations, see (Mason and Watts, 2010).

4 Results and Evaluation

Stage 1 of the experiment lasted for about two days, with multiple manual pauses in between to resist spamming attempts. A total of 135 responses were received, out of which 88 (65.19%) were valid and 47 (34.81%) were invalid according to the criteria stated above. Among the valid responses, 81 (92.05%) were contributed by participants who claimed to be from Mainland China and only 7 (7.95%) by participants from Hong Kong. 38 out of the 47 invalid responses (80.85%) were probably produced by spammers because their completion times were very short and/or the validation measures were bypassed. The 3 largest source channels of valid responses were *bitcoinget* (n=52, 59.09%), *prodege* (n=11, 12.50%) and *getpaid* (n=7, 7.95%), while the 3 largest source regions (based on the IP addresses) were Mainland China (n=54, 61.36%), USA (n=14, 15.91%) and Canada (n=6, 6.82%).

Stage 2 of the experiment lasted for about 4 days also with several breaks. 65 responses were received in Stage 2, among which 54 (83.08%) were valid and 11 (11.92%) were invalid. 46 (85.19%) of the valid responses were contributed by participants from Mainland China and 8 (14.81%) by participants from Hong Kong. 6 (54.55%) of the invalid responses were probably produced by spammers. The main contributing source channels and regions of valid data in Stage 2 were slightly different from Stage 1. Top 3 source channels were *prodege* (n=25, 46.30%), *bitcoinget* (n=7, 12.96%) and *instagc* (n=5, 9.26 %); top 3 source regions were Canada (n=22, 40.74%), USA (n=15, 27.78%) and Mainland China (n=11, 20.37%). Despite the different distributions of source channels and regions, the data obtained from Stage 1 and Stage 2 were highly similar, suggesting that the experiment was highly replicable.

In total, we obtained 200 responses in this experiment, among which 142 (71%) were valid. The valid responses showed high consistency in their answers to the language tasks in Part 2 and Part 3. For example, among the 127 valid responses from Mainland China, the answers to the word segmentation questions in Part 2 had an average consistency[2] of 74.30% (SD=12.94%), while the semantic similarity ratings in

[2]Consistency here means the percentages of the majority-voted answers; if we consider the second most frequent answers,

Part 3 had an average consistency of 58.46% (SD=21.97%). Majority-voted answers and ratings were verified by a team of trained linguists as the most likely segmentations/ratings of the given linguistic materials, while the less popular answers were also verified as possible or reasonable alternatives. These results suggest that the language behavioral data acquired in this experiment, when pruned of invalid responses, were largely consistent with expectations for native language users' judgment.

4.1 Chinese Word Segmentation Data Example

In the experiment, the participants were required to segment 12 short Chinese sentences; because of space limitation, we will only present the results of one representative sentence here. The theoretical segmentation result of the target Chinese sentence "只有依靠群众才能做好工作" (*lit., character by character:* only-have-rely on-depend on-crowd-mass-only-can-do-well-job-work, "The job can only be done well by relying on the messes") is "只有/依靠/群众/才/能/做/好/工作" (*lit. word by word:* only/rely on/the messes/only/can/do/well/job) in which the symbol "/" indicates word boundaries. The segmentation results of this sentence obtained in the experiment are listed in Table 1. We can see that the consistency is high, however the majority-voted result "只有/依靠/群众/才能/做好/工作" is different from the theoretical segmentation result. Most participants treat the slice "才能" as one word instead of two words and the same thing happened to the slice "做好". Speakers' intuition can be different from theoretical analysis: this is an important clue to investigate the representation of Chinese words in the mental lexicon of Chinese speakers.

Segmentation Result	n	%
只有/依靠/群众/才能/做好/工作	100	78.74
只有/依靠/群众/才能/做/好/工作	11	8.66
只有/依靠/群众/才/能/做好/工作	5	3.94
只有/依靠/群众/才/能/做/好/工作	4	3.15
只有/依靠/群众/才/能做好/工作	2	1.57
只有/依靠/群众/才能做好工作	1	0.79
只有/依靠/群众/才能/做好工作	1	0.79
只有/依靠群众/才能/做好工作	1	0.79
只有/依靠群众/才能/做/好/工作	1	0.79
只/有/依靠/群众/才/能/做/好/工作	1	0.79
Total	127	100

Table 1: Chinese Word Segmentation Data Example

4.2 Semantic Similarity Rating Data Example

Semantic transparency affects the representation and processing of compounds (Libben, 1998; Han et al., 2014). In the experiment, we use semantic similarity rating tasks to collect semantic transparency data of 12 compounds which can be used in the studies of mental lexicon. Here we will only discuss two of them in detail. In Chinese, "东西" (*dongxi*, east-west, "thing") is a typical semantically opaque word, because its literal meaning is "east and west" but its actual meaning is "thing": we can hardly find any link between the two. In contrast, "帮助" (*bangzhu*, help-assist, "help") is a typical semantically transparent word, for its literal meaning equals its actual meaning. In our experiment, for each target word, we ask the participants to rate to what extent the meaning of each character when it is used alone is similar to its meaning in the target word. This kind of semantic similarity rating task enables us to estimate the semantic transparency of the target words. The semantic similarity rating data of the above two words are shown in Table 2, and for the results of all the words, see Table 3.

the consistency numbers can be much larger than the reported ones, especially the ones of semantic similarity rating results (see Table 3).

	东西 *dongxi*, east-west, “thing”		帮助 *bangzhu*, help-assist, “help”	
Rating Score	东 *dong*, “east”	西 *xi*, “west”	帮 *bang*, “help”	助 *zhu*, “assist”
1	115	121	6	4
2	2	2	2	13
3	1	1	8	7
4	0	1	23	38
5	8	1	88	63
?	1	1	0	2
Total	127	127	127	127

Table 2: Semantic Similarity Rating Data Example

In the tables, the rating scores 1 to 5 and “?” mean “not similar at all”, “slightly similar”, “moderately similar”, “very similar”, “identical”, and “unable to rate” respectively. The consistency of the semantic similarity rating data is also very high. For example, most participants (115 out of 127) think the meaning of “东” (*dong*, “east”) when it is used alone is not similar at all to its meaning in the word “东西” (*dongxi*, east-west, “thing”), and most participants (121 out of 127) think the meaning of “西” (*xi*, “west”)when it is used alone is not similar at all to its meaning in the word “东西” (*dongxi*, east-west, “thing”). The consistency of the rating data of “帮助” (*bangzhu*, help-assist, “help”) is not as high as “东西” (*dongxi*, east-west, “thing”), but most participants choose 5 which is our expectation and it is also normal that many participants choose 4, since it is next to 5. The semantic transparency estimation of the two words based on these data is quite consistent with our expectation.

5 The Quality Control Issues

In order to obtain high quality data in crowdsourcing environments, it is fundamental to identify invalid responses. Checkpoint questions can be used to identify them. Checkpoint questions should satisfy two conditions. Firstly, a checkpoint question should be super easy, since making wrong judgments to super easy questions is a clear signal of carelessness. Secondly, a checkpoint question should have a publicly recognized correct answer or it cannot act as a standard. Checkpoint questions can be open-ended or close-ended. Open-ended questions are usually more robust than close-ended ones, since their answers are difficult to guess.

There are at least 3 basic measures to deal with invalid responses: (1) blocking the submission of invalid responses; (2) rejecting the invalid responses that have been submitted; (3) refining the data-set received and filter out invalid responses before analysis. Adopting validations on checkpoint questions is a good strategy. A validation is a submission condition and the submission of responses will be blocked if the validations of them are failed. Since checkpoint questions can identify invalid responses, using validations on checkpoint questions can block the submissions of invalid responses. Crowdflower supports validation but it is implemented on the client end, so can be bypassed; but average participants usually don’t have the required expertise to do that, so it is largely reliable.

After the adoption of the above quality control measures, spammers are the major threats to data quality. It can be exhausted to combat spammers manually, because of their high speed and randomness, so automatic monitor programs should be used to combat them. Monitor programs use patterns to detect spammers. Patterns may depend on the specifics of different crowdsourcing practices, but there are some general patterns which are based on the typical behaviors of spammers and can be applied to almost all crowdsourcing practices. One pattern is the “temporal pattern”, abnormal high speed is an obvious feature of spammers and can be used as a general pattern. There are two cases. One case is that the completion time of a response is abnormally short. For instance, the normal completion time of a response is around 9 minutes, but the human spammers only needed an average of 138 seconds and the robot spammers only needed an average of 20 seconds. The other case is that the time interval between 2 responses is

Word	Character	Rating Score						Total
		1	2	3	4	5	?	
东西	东	115	2	1	0	8	1	127
	西	121	2	1	1	1	1	127
地步	地	94	12	8	3	9	1	127
	步	100	11	8	2	4	2	127
漂亮	漂	79	15	10	9	11	3	127
	亮	63	32	15	7	5	5	127
风度	风	109	8	3	0	7	0	127
	度	84	29	7	2	3	2	127
出息	出	97	13	4	3	8	2	127
	息	110	7	3	0	3	4	127
利索	利	80	15	15	3	9	5	127
	索	98	12	6	0	3	8	127
帮助	帮	6	2	8	23	88	0	127
	助	4	13	7	38	63	2	127
衣服	衣	2	8	12	27	78	0	127
	服	32	29	19	24	20	3	127
告诉	告	20	23	24	26	32	2	127
	诉	19	41	30	21	13	3	127
制作	制	4	22	20	43	36	2	127
	作	12	25	31	33	24	2	127
兑换	兑	3	13	13	44	54	0	127
	换	3	8	16	42	56	2	127
灾祸	灾	3	5	16	41	62	0	127
	祸	2	11	21	43	50	0	127

Table 3: The Complete List of Semantic Similarity Rating Data

abnormally short and several such events take place one after another. This temporal pattern can be used to detect concurrent attacks. The other pattern is the "violation of validations". If the validations of a response failed but it was still submitted, then the validations were bypassed and this is a typical behavior of spammers. Once a spammer is detected, we can block it and reject all the responses it submitted if the crowdsourcing platform supports these methods, otherwise we can just pause the task for a while in order to avoid or reduce its attack.

The effect of any single quality control measures is limited; multiple measures should be used at the same time to form a quality control system with much more control power. A reasonable quality control system should notice two key points: (1) maximally block the submission of invalid responses, and (2) maximally filter invalid responses out.

6 Conclusion

Our study showed that crowdsourcing is a very powerful experimental design for exploration cognitive access to the shared Mental Lexicon of the speakers of the same language. We showed that Mandarin speakers shared the same strategy in determination of lexical units. The strategy seems to be match more closely with distributional information. This suggests an empirical approach to lexical unit determination which is then subject to the influence of language use and can lead to changes in the mental lexicon. Although our study is far from conclusive as a proof for the shared lexical access strategy, it does point out to the great potential of pursuing this issue using crowdsourcing experiments.

Acknowledgements

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 544011).

References

Adam J Berinsky, Gregory A Huber, and Gabriel S Lenz. 2011. Using mechanical turk as a subject recruitment tool for experimental research. *Submitted for review*.

Chris Biemann. 2013. Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12. Association for Computational Linguistics.

Matthew JC Crump, John V McDonnell, and Todd M Gureckis. 2013. Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, 8(3):e57410.

Iryna Gurevych and Torsten Zesch. 2013. Collective intelligence and language resources: introduction to the special issue on collaboratively constructed language resources. *Language Resources and Evaluation*, 47(1):1–7.

Yi-Jhong Han, Shuo-chieh Huang, Chia-Ying Lee, Wen-Jui Kuo, and Shih-kuen Cheng. 2014. The modulation of semantic transparency on the recognition memory for two-character chinese words. *Memory & Cognition*, pages 1–10.

Chu-Ren Huang, Elanna I. J. Tseng, Dylan B. S. Tsai, and Brian Murphy. 2003. Cross-lingual Portability of Semantic relations: Bootstrapping Chinese WordNet with English WordNet Relations. *Language and Linguistics*, 4.3:509–532.

Chu-Ren Huang, Ru-Yng Chang, and Shiang bin Li, 2010. *Ontology and the Lexicon*, chapter Sinica BOW: Integration of Bilingual WordNet and SUMO, pages 201–211. Cambridge University Press, Cambridge.

Gary Libben. 1998. Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. *Brain and Language*, 61(1):30 – 44.

Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on amazon's mechanical turk. *Behavior research methods*, 44(1):1–23.

Winter Mason and Duncan J Watts. 2010. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, 11(2):100–108.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130. Association for Computational Linguistics.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, pages 2–9. ACM.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419.

David G Rand. 2012. The promise of mechanical turk: How online labor markets can help theorists run behavioral experiments. *Journal of theoretical biology*, 299:172–179.

Tyler Schnoebelen and Victor Kuperman. 2010. Using amazon mechanical turk for linguistic research. *Psihologija*, 43(4):441–464.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics.