Mechanical Turk-based Experiment vs Laboratory-based Experiment: A Case Study on the Comparison of Semantic Transparency Rating Data

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

In this paper, we conducted semantic transparency rating experiments using both the traditional laboratory-based method and the crowdsourcing-based method. Then we compared the rating data obtained from these two experiments. We observed very strong correlation coefficients for both overall semantic transparency rating data and constituent semantic transparency data ($rho > 0.9$) which means the two experiments may yield comparable data and crowdsourcing-based experiment is a feasible alternative to the laboratory-based experiment in linguistic studies. We also observed a scale shrinkage phenomenon in both experiments: the actual scale of the rating results cannot cover the ideal scale $[0, 1]$, both ends of the actual scale shrink towards the center. However, the scale shrinkage of the crowdsourcing-based experiment is stronger than that of the laboratory-based experiment, this makes the rating results obtained in these two experiments not directly comparable. In order to make the results directly comparable, we explored two data transformation algorithms, z-score transformation and adjusted normalization to unify the scales. We also investigated the uncertainty of semantic transparency judgment among raters, we found that it had a regular relation with semantic transparency magnitude and this may further reveal a general cognitive mechanism of human judgment.

1 Introduction

In experimental linguistic studies, researchers are frequently frustrated by the problem of linguistic data bottleneck which constantly limits the feasibility, efficiency, and reliability of various research projects. It's caused by the practical difficulties of conducting traditional laboratory-based linguistic experiments. Firstly, it's very difficult to obtain large samples using laboratory-based experiments for they are usually very time-consuming and expensive. In order to solve this problem, we need to find a more efficient and economic way to conduct linguistic experiments. Secondly, what's more difficult is to recruit highly diverse subjects due to the difficulties in subject recruitment and the spacial limitations of laboratory-based experiments. As a result, researchers heavily and even blindly rely on relatively small sample size which is 30 or so (Sprouse, 2011) and the undergraduate subject pool. From the point of view of sampling, this is not a good practice, since it raises the concern of external validity, i.e., the extent to which the experimental results can be generalized, because a small and homogeneous sample usually cannot be representative enough. In fact the external validity problem that results from using mainly undergraduate subjects is a typical one and has a dedicated term called the college sophomore problem (Stanovich, 2007; Jackson, 2012). Although there are several responses to this criticism (Stanovich, 2007), the really convincing way to resolve this problem is to use a more diverse subject pool.

Mechanical Turk (MTurk) has emerged in recent years to be a promising solution to the problem of linguistic data bottleneck by providing a new paradigm for linguistic experiments, i.e., the MTurk-based experiment (Mason and Suri, 2012; Horton et al., 2011;

29th Pacific Asia Conference on Language, Information and Computation pages 53 - 62
Shanghai, China, October 30 - November 1, 2015

Paolacci et al., 2010; Schnoebelen and Kuperman, 2010; Buhrmester et al., 2011; Sprouse, 2011; Berinsky et al., 2012), which can hopefully address all the problems mentioned above. MTurk is qualified as a genre of both crowdsourcing which refers to the activities to outsource tasks to undefined and generally large crowds on the web via open call (Howe, 2006; Estellés-Arolas and González-Ladrón-de Guevara, 2012; Wang et al., 2013; Schenk and Guittard, 2011; Howe, 2009), and human computation (Quinn and Bederson, 2009; von Ahn, 2005; Quinn and Bederson, 2011). MTurk needs to be implemented through a website, or more precisely, an MTurk platform. An MTurk platform is an on-line crowdsourcing labor marketplace where requesters post small tasks (conventionally called Human Intelligence Tasks, or HITs) and workers undertake tasks for small pay (Mason and Suri, 2012; Sprouse, 2011). The most famous MTurk platform is Amazon's Mechanical Turk (AMT, www.mturk.com) which was lunched publicly in November 2005; it started early and is so popular in the academic world that it is the de facto standard of MTurk implementation, and the genre name MTurk actually originated from its name and is used by some writers to refer to AMT specially. There are other MTurk implementations, for example another well known MTurk platform is Crowdflower (www.crowdflower.com). Relevant demographics shows that the workers on either AMT (Ross et al., 2010; Pavlick et al., 2014; Ipeirotis, 2010) or Crowdflower[1] are come from all over the world, so both can be treated as international MTurk platforms.

In the early stage of the development of MTurk, it's potential to be an efficient and economic tool for linguistic data collection (e.g., annotation, transcription, translation, etc.) and behavioral research (e.g., survey and experimentation) for social sciences has already been recognized and attempted (Snow et al., 2008; Kittur et al., 2008; Chen et al., 2009). Especially since around 2010, there have been more and more reports on conducting experimental research using MTurk (Mason and Suri, 2012; Rand, 2012; Buhrmester et al., 2011; Horton et al., 2011; Paolacci et al., 2010; Schmidt, 2010; Munro et al., 2010; Schnoebelen and Kuperman, 2010; Sprouse, 2011; Enochson and Culbertson, 2015; Kuperman et al., 2012) and several of them focus on linguistic experiments (Munro et al., 2010; Schnoebelen and Kuperman, 2010; Sprouse, 2011; Enochson and Culbertson, 2015; Kuperman et al., 2012). Experiments conducted on MTurk platforms are usually survey-based and use web questionnaires composed using the GUI toolkits provided by the platforms, and advanced users can make use of HTML, CSS, JavaScript, Adobe Flash (Simcox and Fiez, 2014; Enochson and Culbertson, 2015), etc., to realize additional elements, customized appearance, special control, and apparatus they need. Compared to laboratory-based experiment, the MTurk-based experiment has many attractive merits: 1) the recruitment and compensation of subjects is automatic, painless, on demand, and 24x7 based; 2) MTurk workers are willing to take part in experiments with much less pay than subjects of laboratory-based experiments; 3) it is a lot easier to obtain very large samples; 4) MTurk worker pool is far more diverse than typical undergraduate subject pool widely used in laboratory-based experiment; 5) the anonymous nature of MTurk-based experiment can largely help to avoid experimenter effect, subject crosstalk (Paolacci et al., 2010) and the problem of socially desirable responses.

Data quality is the key concern in conducting research using MTurk-based experiments because the MTurk setting is not so controllable as the laboratory setting, a host of studies have been carried out to address this concern. The comparison between the data obtained from MTurk-based experiments and laboratory-based experiments suggests that MTurk-based experiments can provide comparable or even better data (Munro et al., 2010; Schnoebelen and Kuperman, 2010; Sprouse, 2011; Horton et al., 2011). And a large set of classic effects discovered previously in laboratory-based experiments have been successfully replicated using MTurk-based experiments even in the case of the experiments which require millisecond accuracy timing (Enochson and Culbertson, 2015; Simcox and Fiez, 2014; Crump et al., 2013; Horton et al., 2011). These positive results repeatedly confirm that MTurk is a reliable tool to conduct experimental research which not only yields

[1] For the demographics of Crowdflower's worker pool, see https://success.crowdflower.com/hc/en-us/articles/202703345-Contributors-Crowd-Demographics, retrieved on Apr. 22, 2015.

valid data but also minimizes the cost in time, effort, and expense. Conducting research using MTurk-based experiments lets researchers concentrate on data analysis, creative thinking, and writing instead of being disturbed by various administrative tasks of laboratory-based experiments from time to time, therefore increases their academic productivity. Although, this methodology has not been completely established, its future seems to be guaranteed (Horton et al., 2011).

In order to evaluate a new method, it is a common strategy to compare the results yield by the new method with the results yield by the established method to see their agreement. Although neither method is perfect or completely reliable, since the established method is well acceptable, if the new method agrees well enough with it, then the new method is also acceptable to be an alternative. We conducted two similar semantic transparency rating experiments using the Mechanical Turk-based method and the traditional laboratory-based method. We will compare the results from these two experiments to see their agreement hence we can further evaluate the Mechanical Turk-based experimentation.

2 Method

2.1 MTurk-based Semantic Transparency Rating Experiment[2]

2.1.1 Materials

We selected a total of $1,176$ disyllabic Chinese nominal compounds which have mid-range word frequencies and appear in both Sinica Corpus 4.0 (Chen et al., 1996) and the "Lexicon of Common Words in Contemporary Chinese 现代汉语常用词表", see Wang et al. (2014) for details.

2.1.2 Experimental Design

Normally, a crowdsourcing experiment should be reasonably small in size. We randomly divide these 1,176 words into 21 groups, G_i $(i = 1, 2, 3, ..., 21)$; each group has 56 words.

Questionnaires We collect overall semantic transparency (OST) and constituent semantic transparency (CST) data of these words. In order to avoid interaction, we designed two kinds of questionnaires to collect OST data and CST data respectively. So G_i $(i = 1, 2, 3, ..., 21)$ has two questionnaires, one OST questionnaire for OST data collection and one CST questionnaire for CST data collection. Besides titles and instructions, each questionnaire has 3 sections. Section 1 is used to collect identity information includes gender, age, education and location. Section 2 contains four very simple questions about the Chinese language; the first two questions are open-ended Chinese character identification questions, the third question is a close-ended homophonic character identification question, and the fourth one is a close-ended antonymous character identification question; different questionnaires use different questions. Section 3 contains the questions for semantic transparency data collection. Suppose AB is a disyllabic nominal compound, we use the following question to collect its OST rating scores: "How is the sum of the meanings of A and B similar to the meaning of AB?" And use the following two questions to collect its CST rating scores of its two constituents: "How is the meaning of A when it is used alone similar to its meaning in AB?" and "How is the meaning of B when it is used alone similar to its meaning in AB?". 7-point scales are used in section 3; 1 means "not similar at all" and 7 means "almost the same".

In order to evaluate the data received in the experiments, we embedded some evaluation devices in the questionnaires. We mainly evaluated intra-group and inter-group consistency; and if the data have good intra-group and inter-group consistency, we can believe that the data quality is good. In each group we choose two words and make them appear twice, we call them intra-group repeated words and we can use them to evaluate the intra-group consistency. We insert into each group two same extra words, w_1"地步", w_2"高山", to evaluate the inter-group consistency.

Quality Control Measures On a crowdsourcing platform like Crowdflower, the participants are anonymous, they may try to cheat and submit invalid data, and they may come from different countries and speak different languages rather than the required one. There may be spammers who continuously submit invalid data at very high speed and

[2]We have reported this experiment in Wang et al. (2014).

they may even bypass the quality control measures to cheat for money. In order to ensure that the participants are native Chinese speakers and to improve data quality, we use the following measures, (1) a participant must correctly answer the first two Chinese character identification questions in the section 2s of the questionnaires, and he/she must correctly answer at least one of the last two questions in these section 2s; (2) If a participant do not satisfy the above conditions, he/she will not see Section 3s; (3) each word stimulus in section 3s has an option which allows the participants to skip it in case he/she does not recognize that word; (4) all the questions in the questionnaires must be answered except the ones which allow to be skipped and are explicitly claimed to be skipped; (5) we wrote a monitor program to detect and resist spammers automatically; (6) after the experiment is finished, we will analyze the data and filter out invalid data, and we will discuss this in detail in section 2.1.3.

	OST		CST	
G_i	n	%	n	%
G_1	62	68.89	70	77.78
G_2	60	66.67	64	71.11
G_3	61	67.78	58	64.44
G_4	57	63.33	58	64.44
G_5	51	56.67	59	65.56
G_6	55	61.11	54	60
G_7	54	60	55	61.11
G_8	60	66.67	48	53.33
G_9	52	57.78	55	61.11
G_{10}	58	64.44	59	65.56
G_{11}	52	57.78	56	62.22
G_{12}	55	61.11	63	70
G_{13}	52	57.78	57	63.33
G_{14}	56	62.22	54	60
G_{15}	54	60	53	58.89
G_{16}	58	64.44	56	62.22
G_{17}	52	57.78	50	55.56
G_{18}	53	58.89	51	56.67
G_{19}	53	58.89	50	55.56
G_{20}	53	58.89	51	56.67
G_{21}	52	57.78	51	56.67
Min	51	56.67	48	53.33
Max	62	68.89	70	77.78
Median	54	60	55	61.11
Mean	55.24	61.38	55.81	62.01
SD	3.4	3.78	5.32	5.91

Table 1: Amount of valid response in the OST and CST datasets of each group.

Experimental Platform and Procedure We choose Crowdflower as our experimental platform, because according to our previous experiments, it is a feasible crowdsourcing platform to collect Chinese language data. We create one task for each questionnaire on the platform; there are 21 groups of word and each group has one OST questionnaire and one CST questionnaire, so there are a total of 42 tasks T_i^{ost}, T_i^{cst} $(i = 1, 2, 3, ..., 21)$. We publish these 42 tasks successively, and for each task we create a monitor program to detect and resist spammers. All of these tasks use the following parameters: (1) each task will collect 90 responses; (2) we pay 0.15USD for each response of OST questionnaire and pay 0.25USD for each response of CST questionnaire; (3) each worker account of Crowdflower can only submit one response for each questionnaire and each IP address can only submit one response for each questionnaire; (4) we only allow the workers from the following regions (according to IP addresses) to submit data: Mainland China, Hong Kong, Macau, Taiwan, Singapore, Malaysia, USA, UK, Canada, Australia, Germany, France, Italy, New Zealand, and Indonesia; and we can dynamically disable or enable certain regions on demand in order to ensure both data quality and quantity.

2.1.3 Data Cleansing and Result Calculation

The OST dataset produced by the OST task T_i^{ost} $(i = 1, 2, 3, ..., 21)$ is D_i^{ost}. The CST dataset produced by the CST task T_i^{cst} is D_i^{cst}. Each dataset contains 90 responses. Because of the nature of crowdsourcing environment, there are many invalid responses in each dataset; so firstly we need to filter them out in order to refine the data. A response is invalid if (1) its completion time is less than 135 seconds (for OST responses); its completion time is less than 250 seconds (for CST responses); or (2) it failed to correctly answer the first two questions of section 2s of the questionnaires; or (3) it wrongly answered the last two questions of section 2s of the questionnaires; or (4) it skipped more than six words in section 3s of the questionnaires; or (5) it used less than three numbers on the 7-point scales in section 3s of the questionnaires. We also filtered out the responses from the workers who appeared in more than one countries/regions according to their IP addresses. The statistics of valid response are shown

in Table 1.

The OST dataset D_i^{ost} $(i = 1, 2, 3, ..., 21)$ contains n_i valid responses; it means word w in the OST dataset of the ith group has n_i OST rating scores; the arithmetic mean of these n_i OST rating scores is the OST result of word w. The CST results of the two constituents of word w are calculated using the same algorithm.

2.2 Laboratory-based Semantic Transparency Rating Experiment

2.2.1 Material

The Mechanical Turk-based semantic transparency rating experiment is a large-scale experiment, it collected the overall and constituent semantic transparency rating data for $1,176$ compounds. This scale is beyond the capacity of common laboratory-based experiment given the time and resource limitations. So it is impossible for us the conduct a completely parallel semantic transparency rating experiment in the laboratory setting. As a practically and statistically feasible alternative, we extracted a representative sample of reasonable size for laboratory-based experiment from the $1,176$ compound stimuli of the Mechanical Turk-based experiment. Then the method comparison will be conducted on the basis of the sample.

The compound stimuli of the Mechanical Turk-based semantic transparency rating experiment belong to three structural categories, i.e., NN, AN, VN, the sample should cover all these category types. According the overall semantic transparency value and constituent semantic transparency value of compound, compounds are usually divided into four categories: 1) TT, the compounds with the largest overall semantic transparency values and the most balanced constituent semantic transparency values, 2) TO, the compounds with the mid-range overall semantic transparency values and the most unbalanced constituent semantic transparency values and the CST of the first morpheme is larger than that of the second, 3) OT, the compounds with the mid-range overall semantic transparency values and the most unbalanced constituent semantic transparency values and the CST of the second morpheme is larger than that of the first, and 4) OO, the compounds with the lowest overall semantic transparency values and the most balanced constituent semantic transparency values. The sample should also cover all these four semantic transparency types. A total of 152 compounds were selected; all of the compounds have the modifier-head structure.

2.2.2 Questionnaire

The questionnaire is divided into three parts. Part I is the demographic questions, we ask the subjects to provide their demographic information on 1) gender, 2) age, 3) language background, 4) native place, and 5) email address (optional). Part II is the overall semantic transparency rating task, the subjects are asked to rate the overall semantic transparency of the compound stimuli one by one, and we use the same questions and rating scales as the Mechanical Turk-based experiment. Part III is the constituent semantic transparency rating task, the subjects are asked to rate the constituent semantic transparency of the compound stimuli one by one, and we also use the same questions and rating scales as the Mechanical Turk-based experiment. We make "笑脸", "蓝本", "火灾", "脾气" appear twice in the questionnaire, so we can use them to check the consistency of ratings. The questionnaire has a simplified Chinese character version and a traditional Chinese character version. And the questionnaires are implemented using Google Form, the whole questionnaire is divided into pages, each page contains six stimuli. At the end of each quarter of the questionnaire, we show the subjects a notice to tell them that they can take a short break (three to five minutes) if they feel tired. It takes about 45 minutes to fill out the questionnaire.

2.2.3 Subjects

We recruited a total of 78 students at the Hong Kong Polytechnic University. Seventy-four of them are undergraduates, and four of them are postgraduates. Thirty-nine of them are from mainland China and the other 39 are Hong Kong local. The subjects from mainland China came from 19 different provinces: Anhui 安徽, 3; Chongqing 重庆, 3; Fujian 福建, 2; Gansu 甘肃, 1; Guangdong 广东, 3; Guizhou 贵州, 2; Hebei 河北, 1; Heilongjiang 黑龙江, 2; Henan 河南, 1; Hubei 湖北, 1; Jiangsu 江苏, 2; Jilin 吉林, 1; Liaoning 辽宁, 1; Neimenggu 内蒙古, 3; Shandong 山东, 5; Shanghai 上海, 1; Shanxi 陕西, 5; Tianjin 天津, 1; Zhejiang 浙江, 1. Fourty-

one subjects are 16 to 20 years old; 33 are 21 to 25; 4 are 26 to 30. Twenty-two of them are male, the other 56 are female. Their mother tongue is Chinese and all of them can speak Putonghua.

2.2.4 Procedure

The subjects were invited into the laboratories. Because the subjects from mainland China and Hong Kong would use different versions of questionnaire, two laboratories were prepared for the experiment, one was for the subjects from mainland China and the simplified character version questionnaire would be used, and the other laboratory was for the subjects from Hong Kong and the traditional character version questionnaire would be used. Each subject was assigned a unique code (or seat number). When the subjects arrived, they were guided to their desks according to their codes. On each desk there was a computer which was displaying a brief introduction to the experiment and at the bottom of the introduction, there were two buttons: "I Agree" and "I Disagree" respectively. We briefly explained the experiment to the subjects orally, and then asked them to sign the consent forms on their desks first if they agreed to participate in our experiment. After they signed the consent forms, they could then read the introduction on the screen and press "I Agree" to start to fill in the questionnaire. Once a subject finished the experiment, he/she would get an allowance of 100 Hong Kong dollars. All the 78 subjects finished the experiment, so we collected 78 responses.

2.2.5 Data Cleansing and Result Calculation

We firstly checked the responses one by one and filtered out invalid ones. A response is considered invalid if 1) more than 15 words were skipped (i.e., the subject claimed that he/she didn't know these words), or 2) less than three numbers of the 7-point rating scale were used. Only two invalid responses were identified, one was from a mainland subject, the other was from a Hong Kong subject. So there are a total of 76 valid responses, this means each word was rated by 76 subjects. The OST and CST results of these words were calculated based on these 76 responses, the calculation method was the same as the Mechanical Turk-based experiment.

3 Results and Discussion

3.1 Correlation

We can evaluate the semantic transparency rating results from the Mechanical Turk-based experiment by examining to what extent they correlates with the results from the laboratory-based experiment. This is a commonly used practice in psycholinguistics.

Strictly speaking, the distributions of the overall and constituent semantic transparency of compound are not normal and do not satisfy the requirement of Pearson's product-moment correlation coefficient, so the Spearman's rank-order correlation coefficient is used. We calculated three correlation coefficients, 1) the correlation coefficient between the normalized OST results from the two experiments: 0.94, 2) the correlation coefficient between the normalized CST results of the first morphemes of the compounds from the two experiments: 0.93, and 3) the correlation coefficient between the normalized CST results of the second morphemes of the compounds from the two experiments: 0.92. All of the correlation coefficients are larger than 0.9 which indicates that the results from the Mechanical Turk-based experiment correlate strongly with the results from the laboratory-based experiment. From the scatter plots (see Figure 1), we can see that although these two kinds of results correlates strongly with each other, we cannot say that they agree with each other very well, because the dots do not distribute around the line of equality (the dashed line).

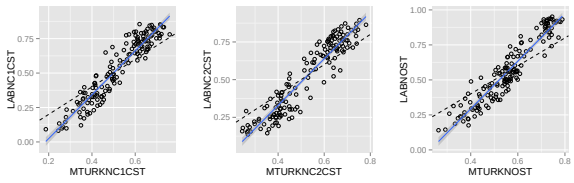

Figure 1: Correlations between normalized OST and CST results from the MTurk-based experiment and the lab-based experiment.

3.2 Scale Shrinkage Issue

We also checked and compared the distributions of the semantic transparency rating results from the Mechanical Turk- and laboratory- based experiments, see Figure 2. We can see that the distributions of the results from the two experiments have

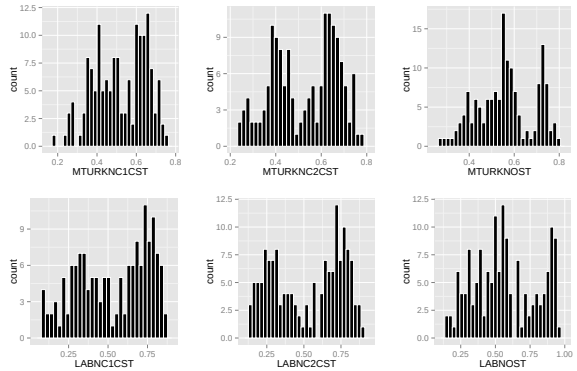

Figure 2: Distributions of semantic transparency rating results from the MTurk-based experiment and the lab-based experiment.

the similar overall forms, but the two kinds of results distribute on different scales. The scale of the Mechanical Turk OST results is from 0.26 to 0.79, however the scale of the laboratory OST results is from 0.14 to 0.95; the scale of the Mechanical Turk C1CST results is from 0.19 to 0.76, while the scale of the laboratory C1CST results is from 0.08 to 0.86; the scale of the Mechanical Turk C2CST results is from 0.24 to 0.78, but the scale of the laboratory C2CST results is from 0.14 to 0.89. Since in our compound stimuli, there are completely transparent compounds and completely opaque compounds, so ideally, two kinds of results should share and cover the same scale from 0 to 1. But virtually, for this kind of subjective rating tasks, subjects rarely totally agree with each other and there is always some noise or errors of varied degrees. Consequently, the distributions of the results of subjective rating tasks rarely cover the whole scale. The actual scales usually shrink towards the center. The scale shrinkage of the results from the Mechanical Turk-based experiment is larger than that of the results from the laboratory-based experiment; this is perhaps because that the Mechanical Turk-based experiment has higher noise level than the laboratory-based experiment.

3.3 Data Transformation

Because the semantic transparency results from the Mechanical Turk- and laboratory-based experiments use different scales and have different units, they are not directly comparable. In order to make the kinds of results comparable, we need to transform the results so that they will use the same scale. We are going to examine two kinds of data transformation methods: 1) Z-score transformation (standardization), 2) adjusted normalization; next we are going to discuss them one by one.

Z-score Transformation

The z score is calculated by the following formula:

$$z\ score = \frac{raw\ score - mean}{standard\ deviation}$$

The raw scores (normalized OST and CST results) from Mechancial Turk- and laboratory-based experiments are transformed into z scores according to the above formula; we call the z-score transformed normalized OST and CST results the standardized OST and CST results. After z-score transformation, the standardized semantic transparency results from the two experiments will share the same scale.

Then we can further examine the agreement of the standardized semantic transparency results from the two experiments, see Figure 3. On these scatter plots, we can see that now all the dots distribute around the line of equality (the dashed line) and the regression line basically coincides with the line of equality; compared with the scatter plots based on the raw scores (see Figure 1), the standardized results agree with each other better which makes the results from the two experiments comparable.

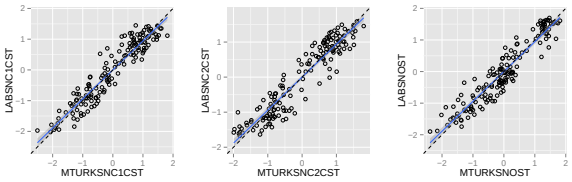

Figure 3: Correlations between standardized OST and CST results from the MTurk-based experiment and the lab-based experiment.

Adjusted Normalization

The adjusted normalized score is calculated according to the following formula:

$$ANscore = \frac{raw\ score - min\ raw\ score}{max\ raw\ score - min\ raw\ score}$$

Since the actual scales of the raw scores shrink towards the center, we can use the above formula to

stretch the scales to $[0, 1]$ again. When using this formula, we need to make sure that the maximum and minimum raw scores are not outliers, otherwise this transformation will fail. The results from the two experiments are both transformed using this formula, after this, they will again share the same scale. See Figure 4 for the relations between the adjusted normalized semantic transparency results from both experiments.

Compared with the raw scores (see Figure 1), the adjusted normalized results from the Mechanical Turk- and laboratory-based experiments agree with each other better, but the agreement is not as good as the standardized results (see Figure 3). However the adjusted normalization method has an advantage over the standardization method, that is the adjusted normalization will yield results from 0 to 1 and this scale is accord with the definition of semantic transparency value.

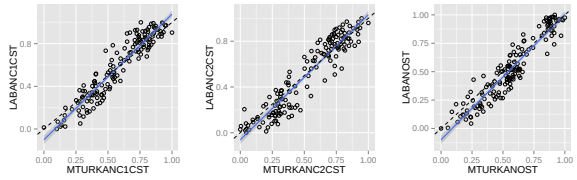

Figure 4: Correlations between adjusted normalized OST and CST results from the MTurk-based experiment and the lab-based experiment.

3.4 Uncertainty of Semantic Transparency Judgment among Raters

Semantic transparency rating task is a subjective rating task. In such a task, the subjects rarely totally agree with each other and there are usually errors of varied degrees. So we can say that there is usually some uncertainty or inconsistency of judgment among raters. Next we are going to measure the uncertainty of judgment among raters and to examine its relationship with the semantic transparency value.

In our semantic transparency rating tasks, 7-point scales are used as the measurement instrument. For a di-morphemic word ab, suppose that m raters rated its overall semantic transparency (OST) and constituent semantic transparency (CST), so ab has m OST ratings scores and also has m C1CST rating scores and m C2CST rating scores; each rating score can only be one of $\{1, 2, 3, 4, 5, 6, 7\}$. For the m OST rating scores, suppose the possibilities of the numbers on the 7-point scale to be chosen are $p_1, p_2, ..., p_7$ respectively, the resultant OST value is the mean of these m rating scores and the uncertainty of judgment among raters can be calculated using the formula of information entropy:

$$OSTRIE = -\sum_{i=1}^{7} p_i \log_2 p_i$$

using the same formula, C1CSTRIE and C2CSTRIE can also be calculated. See Figure 5 for the relationship between semantic transparency value and uncertainty of judgment among raters; both Mechanical Turk data and laboratory data are used to draw the figures. We can observe a very strong and regular relation between them. In terms of this relationship, laboratory data show stronger and more regular relationship than Mechanical Turk data. This kind of curve may reveal some kind of general cognitive mechanism of human subjective judgment.

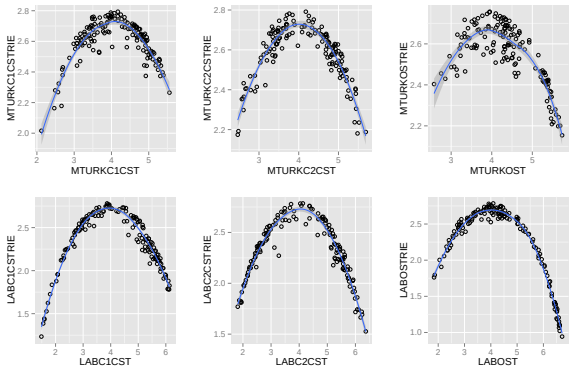

Figure 5: Uncertainty of semantic transparency judgments among the raters of the MTurk-based experiment and the lab-based experiment.

Acknowledgments

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. 2012. Evaluating online labor markets for experimen-

tal research: Amazon. com's mechanical turk. *Political Analysis*, 20(3):351–368.

Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5.

Keh-Jiann Chen, Chu-Ren Huang, Li-Ping Chang, and Hui-Li Hsu. 1996. Sinica corpus: Design methodology for balanced corpora. In B.-S. Park and J.B. Kim, editors, *Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation*, pages 167–176. Seoul:Kyung Hee University.

Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei. 2009. A crowdsourceable qoe evaluation framework for multimedia content. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 491–500. ACM.

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.

Kelly Enochson and Jennifer Culbertson. 2015. Collecting psycholinguistic response time data using amazon mechanical turk. *PLoS ONE*, 10(3):e0116946, 03.

Enrique Estellés-Arolas and Fernando González-Ladrón-de Guevara. 2012. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200.

John J Horton, David G Rand, and Richard J Zeckhauser. 2011. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14(3):399–425.

Jeff Howe. 2006. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

Jeff Howe. 2009. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Three Rivers Press.

Panagiotis G Ipeirotis. 2010. Demographics of mechanical turk.

Sherri Jackson. 2012. *Research methods and statistics: A critical thinking approach*. Cengage Learning.

Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44(4):978–990.

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

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.

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

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of amazon mechanical turk. *Transactions of the Association for Computational Linguistics*, 2:79–92.

Alexander J Quinn and Benjamin B Bederson. 2009. A taxonomy of distributed human computation. *Human-Computer Interaction Lab Tech Report, University of Maryland*.

Alexander J Quinn and Benjamin B Bederson. 2011. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1403–1412. ACM.

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.

Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 2863–2872. ACM.

Eric Schenk and Claude Guittard. 2011. Towards a characterization of crowdsourcing practices. *Journal of Innovation Economics & Management*, 7(1):93–107.

L Schmidt. 2010. Crowdsourcing for human subjects research. *Proceedings of CrowdConf*.

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

Travis Simcox and Julie A Fiez. 2014. Collecting response times using amazon mechanical turk and adobe flash. *Behavior research methods*, 46(1):95–111.

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.

Jon Sprouse. 2011. A validation of amazon mechanical turk for the collection of acceptability judgments in linguistic theory. *Behavior research methods*, 43(1):155–167.

Keith E Stanovich. 2007. *How to think straight about psychology*. HarperCollins Publishers.

Luis von Ahn. 2005. *Human Computation*. Ph.D. thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA, 12.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2013. Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, 47:9–31.

Shichang Wang, Chu-Ren Huang, Yao Yao, and Angel Chan. 2014. Building a semantic transparency dataset of chinese nominal compounds: A practice of crowdsourcing methodology. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 147–156, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.