Lingua Sinica

ORIGINAL ARTICLE Open Access

Word intuition agreement among Chinese speakers: a Mechanical Turk-based study

Shichang Wang[1*], Chu-Ren Huang[2], Yao Yao[2] and Angel Chan[2]

* Correspondence:
wangshichang@sdu.edu.cn
[1]School of Literature, Shandong University, Jinan, China
Full list of author information is available at the end of the article

Abstract

Word intuition is speakers' intuitive knowledge on wordhood. Collective word intuition is the word intuition of the whole language community. Given this definition, the optimal word segmentation result in Chinese NLP should reflect collective word intuition. It is also believed that an ideal definition of Chinese word should accord with the collective word intuition of Chinese speakers. To test the validity and feasibility of modeling collective word intuition, it is important to know to what extent Chinese speakers agree with each other on what is a word. In this study, we measured word intuition agreement using Mechanical Turk-based Chinese word segmentation experiment. Three metrics were used: proportionate agreement, Cohen's kappa, and Fleiss' kappa. The results show that Chinese speakers agree with each other almost perfectly on what is a word. And we found no evidence to support an effect of semantic transparency on word intuition agreement. Such high word intuition agreement among Chinese speakers supports the psychological reality of Chinese word and also suggests that that it is quite feasible to formulate a definition of Chinese word by modeling the collective word intuition of Chinese speakers.

Keywords: Word intuition, Word segmentation, Word intuition agreement, Semantic transparency, Mechanical Turk

1 Introduction

Word intuition is speakers' intuitive knowledge of what a word is and can be defined individually or collectively. The individual word intuition is the word intuition of individual speakers. Whereas, the collective word intuition is the word intuition of the whole language community. In English and other languages with conventionalized word boundaries in orthography, collective word intuition can be roughly modeled by the collective behavior in marking orthographic wordbreaks. However, in languages without conventionalized word boundaries in orthography, explicit empirical evidence for both individual and collective word intuition is very difficult to obtain.

A priori, a general method to the description of the collective word intuition of Chinese speakers is to summarize the individual word intuition of all the Chinese speakers. In principle, collective word intuition of Chinese speakers can be measured using word segmentation tasks (Hoosain 1992; Wang 王立 2003), assuming that speakers' word segmentation behaviors reflect their word intuition. This quantifiable result of segmentation consistency can be a convenient measurement of Chinese speakers' word intuition. This leads to a potentially very attractive feature of collective word intuition: that it can be quantified with probabilistic values.

In computational and Chinese language processing, the difficulty in capturing collective word intuition leads to the difficulty in consensus in articulating consensus word segmentation standard (Huang et al. 1996; Liu et al. 刘源等 1994) and in achieving optimal results in word segmentation (Huang and Xue 2012; Huang and Xue 2015; Huang and Zhao 黄昌宁, 赵海 2007). Presumably, a clearly articulated collective word intuition for Chinese speakers would be the ideal word segmentation standard for Chinese language processing and would also make modeling and evaluation of Chinese word segmentation explicit and straightforward. And the probabilistic representation of word intuition in fact is even better suited as word segmentation standards given the predominance of stochastic models in computational word segmentation.

Collecting word intuition experimental data, however, is not an easy task. It is a very time- and resource-intensive process in traditional laboratory setting. In addition, given the possible variations in such experimental results, one may even question the psychological reality of word as a natural linguistic unit in Chinese (Hoosain 1992; Huang and Xue 2012). Only when the word intuition agreement among Chinese speakers is reasonably high, can the psychological reality of word in Chinese be supported and meanwhile can word intuition be the solid foundation of the definition of Chinese word. It is interesting to note that in fact, the first studies tackling agreements in human word segmentation results were done by computational linguists rather than psycholinguists, such as Sproat et al. (1996) and Liu and Liang 刘源, 梁南元 (1986). Liu and Liang 刘源, 梁南元 (1986) reported agreement ranging from 60% (before training) to 80% (after training), without describing how agreement is measured and cannot be correctly interpreted. Another weakness of their study is that the text stimuli are short and out of context, and hence do not model realistic context of word segmentation. Sproat et al. (1996) conducted a similar experiment. They extracted 100 sentences (4372 Chinese characters in total) randomly from a corpus and then asked six Chinese native speakers (three from Mainland and three from Taiwan) to segment these sentence stimuli. They used the arithmetic mean of precision and recall to measure interjudge similarity between each unique subject pair. They reported minimum interjudge similarity of 0.69, maximum similarity 0.89, and the mean of 0.76. The segmentation agreement in terms of the arithmetic mean of precision and recall, however, is rarely (if ever) used in psychological or linguistic studies. Instead, various kappa statistics are more "standard" for current studies in cognitive and social sciences. Their number of subjects (6) is also considered to be too small for this study (typically 20 or more will be required.)

The possible link between collective word intuition and word segmentation standard, as well as the fact that previous studies failed to recruit enough subjects for valid results lead to our proposal to use the NLP technique of crowdsourcing to tackle this issue. 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 (Berinsky et al. 2012; Buhrmester et al. 2011; Horton et al. 2011; Mason and Suri 2012; Paolacci et al. 2010; Schnoebelen and Kuperman 2010; Sprouse 2011). 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 (Horton et al. 2011; Munro et al. 2010; Schnoebelen and Kuperman 2010; Sprouse 2011). A large set of classic effects discovered previously in laboratory-based experiments have been successfully replicated using MTurk-based experiments (Crump et al. 2013; Enochson and Culbertson 2015; Horton et al. 2011; Simcox and Fiez 2014). Last but not the least, MTurk has been successfully implemented for research on Chinese language resources (Wang et al. 2014b). Our current study hence designs an MTurk-based experiment on word intuition, with the hope of constructing word segmentation resources to inform computational word segmentation in the future.

"A word is a minimum free form". This is perhaps the most classic definition of word which is suggested by Bloomfield (1933: 178). However, according to this definition, the Chinese forms 江水 *jiāngshuǐ* 'river water' and 龙眼 *lóngyǎn* 'longan' are not words because 江 *jiāng* 'river', 水 *shuǐ* 'water', 龙 *lóng* 'dragon', and 眼 *yǎn* 'eye' are all free forms (or free morphemes). This is quite counter-intuitive. If we inspect carefully the two forms, we can find that in fact, they are very different. The form *jiāngshuǐ* is semantically transparent, but *lóngyǎn* is semantically opaque. Because of this, among Chinese linguists, there is rarely debate on the wordhood of *lóngyǎn*, but there are still some debates on the wordhood of *jiāngshuǐ*. This leads to the hypothesis that semantic transparency may affect Chinese speakers' word intuition. Is the word intuition agreement on semantically transparent forms significantly lower than semantically opaque forms? This study will also probe into this research question.

2 Method

2.1 Materials

The materials of word segmentation tasks are at least phrases, but we prefer naturally occurred sentences. In order to cover more linguistic phenomena to better support the studies of word intuition, we decided to use more than 150 long sentences (the crowdsourcing method makes this possible). Meanwhile, the resultant dataset will also be used to examine the effect of semantic transparency on word intuition, so these sentences should contain the words to be used in the examination of semantic transparency effect. Hence, the material selection procedure consists of two steps: (1) word selection, i.e., to select an initial set of words which would be used in the examination of semantic transparency effect, and (2) sentence selection, i.e., to select a set of sentences which contains the words selected in step 1 (each sentence carries one word) and at the same time satisfy other requirements.

2.1.1 Word selection

We have already created a semantic transparency dataset SimTransCNC 1.0 which contains the overall and constituent semantic transparency rating data of about 1200 Chinese bimorphemic nominal compounds which have mid-range word frequencies and consist of free morphemes (Wang et al. 2014a). Based on this dataset, 152 words are selected. These words have two functions in this study: (1) they are used as indexes to extract sentences from corpus; (2) they are used as the word stimuli in the study of the effect of semantic transparency on word intuition agreement. However, for function 2, we will not use all the 152 words; instead, these

152 words will firstly undergo a laboratory-based semantic transparency rating experiment (Wang et al. 2015) to further ensure the accuracy of their semantic transparency scores and then a subset of words will be selected according to the laboratory experiment results to examine the semantic transparency effect on word intuition agreement (see Section 3.5).

In the study of the semantic transparency effect on word intuition agreement, the independent variable is the semantic transparency of words and the dependent variable is the word intuition agreement of these words. We control the length, part-of-speech, frequency, morphological structure, and the nature of constituent of these words. All the words are bimorphemic nominal compounds which consist of free morphemes and have the structure of modifier-head and mid-range frequencies. But the modifier-head structure can be further mainly divided into three substructures: NN, AN, and VN. These words cover all the three substructures to enable us to see if these substructures make differences. Following Libben et al. (2003), we differentiate four transparency types: TT, TO, OT, and OO; "T" means "transparent" and "O" means "opaque". TT words show the highest OST (overall semantic transparency: the semantic transparency of a whole compound) scores and the most balanced CST (constituent semantic transparency: the semantic transparency of a constituent of a compound) scores, e.g., 江水 *jiāngshuǐ* 'river water'; OO words have the lowest OST scores and the most balanced CST scores, e.g., 脾气 *píqi* 'temperament'; TO and OT words bear mid-range OST scores and the most imbalanced CST scores, e.g., 音色 *yīnsè* 'timbre' (TO) and 贵人 *guìrén* 'magnate' (OT). See Table 1 for the distribution of the selected words.

2.1.2 Sentence selection

The words selected in step 1 were used as indexes, and all the sentences carrying them in Sinica corpus 4.0 were extracted. One sentence was selected for each word roughly according to the following criteria: (1) the length of sentence should be between 20 and 50 characters (punctuations excluded); (2) the sentence should not contain too many punctuations; (3) preferred concrete and narrative sentences to abstract ones which are difficult to understand; and (4) if we could not find proper sentences from Sinica corpus for some words, we used other corpora (only 5 sentences). In this way, a total of 152 sentences are selected, for the length (in character) distribution, see Table 2.

2.2 Crowdsourcing task design

Since a crowdsourcing task should be short, these 152 sentences are evenly and randomly divided into eight sentence groups; each sentence group has 19 sentences. We created one crowdsourcing task for each sentence group on Crowdflower; according to our previous studies, compared to Amazon Mechanical Turk (MTurk), Crowdflower is a more

Table 1 Distribution of types of selected words

Transparency type	Word structure		
	NN	AN	VN
TT	20	10	10
TO	20	6	10
OT	20	10	10
OO	20	10	6

Table 2 Length distribution of selected sentences

	Length of sentence
Min	20
Max	46
Sum	4946
Mean	32.54
SD	5.46

feasible platform for Chinese linguistic data collection since it is more accessible and can reach more Chinese speakers than Amazon Mechanical Turk (Wang et al. 2014b).

2.2.1 Questionnaires

The core of each crowdsourcing task is a questionnaire. Each questionnaire consists of five sections: (1) title, (2) instructions, (3) demographic questions, (4) screening questions, and (5) segmentation task; both simplified and traditional Chinese character versions are provided. Section 3, demographic questions, asks the online subjects (all of the subjects are volunteers) to provide their identity information on gender, age, level of education, and email address (optional). Section 4, screening questions, consists of four simple questions on the Chinese language which can be used to test if a subject is a Chinese speaker or not; the first two questions are open-ended Chinese character identification questions. Each question shows a picture containing a simple Chinese character and asks the subject to identify and type it in a text box below it. The third question is a close-ended homophonic character identification question. It shows the subject a character and asks him/her to identify its homophonic character in 10 different characters. The fourth one is a close-ended antonymous character identification question asking the subject to identify the antonymous character of the given one from 10 different characters. The sections 4s of the eight crowdsourcing tasks share the same question types but have different question instances. Section 5, the segmentation task, shows the subjects 19 sentences and asks them to insert a word boundary symbol ("/") at each word boundary they perceive. The subjects are required to insert a "/" behind each punctuation and the last character of a sentence. The subjects are also informed that they need not to care about if their judgments are right or wrong, but just follow their intuitions.

2.2.2 Parameters of tasks

These eight crowdsourcing tasks are created with the following parameters: (1) each subject account can only submit one response to one task; (2) each IP address can only submit one response to one task; (3) we only accept the responses from mainland China, Hong Kong, Macao, Taiwan, Singapore, Indonesia, Malaysia, Thailand, Australia, Canada, Germany, USA, and New Zealand since there areas are main Chinese speaking areas; and (4) we pay 0.25 USD for one response.

2.2.3 Quality control measures

The following quality control measures are used: (1) Section 4, screening questions, was used to discriminate Chinese speakers from non-Chinese speakers and to block bots; (2) Section 5, the segmentation task, was kept invisible unless the first two screening questions were correctly answered; (3) the answers to the segmentation questions in section 5 must comply with the prescribed format to prevent random string: (a) the segmentation answer to each sentence must be only composed by the original sentence with one or zero "/" behind each Chinese character and each punctuation; (b) in the answers behind each punctuation, there must be a “/”; (c) the end of an answer must be a “/”; (4) the submission attempts were blocked unless all the required questions are answered and the answers satisfy the above conditions; and (5) data cleansing was conducted after data collection to rule out invalid responses.

2.3 Procedure

We firstly ran a small pretest task to test if the tasks were correctly designed, and it turned out that the pretest task could run smoothly. Then, we launched the first task and let it run alone for about 2 days to further test the task design. After we finally confirmed that the tasks could really run smoothly, we launched the other seven tasks and let them run concurrently. Our aim was to collect 200 responses for each task. The speed was amazingly fast in the beginning, and all eight tasks received their first 100 responses in the first 3 to 6 days; then the speed became slower and slower, it eventually took us about 1.3 months to reach our target number. After all, Crowdflower is not a Chinese native crowdsourcing platform; this kind of speed is understandable.

3 Results

3.1 Data cleansing

All tasks successfully obtained 200 responses. However, not all responses are valid. Compared to the laboratory setting, the crowdsourcing environment is quite noisy by nature, so before the newly collected data can be used in any serious analysis to draw reliable conclusions, data cleansing must be conducted. The raw responses underwent rule-based data cleansing. A response is considered invalid if it has at least one of the following five features: (1) at least one of the four screening questions are incorrectly answered; (2) the lengths of the resultant segments of at least one of its 19 sentences are all one character; (3) at least one segment longer than seven characters is observed in the resultant segments of its 19 sentences; (4) the completion time of the response is shorter than 5 min; and (5) the completion time of the response is longer than 1 h. Invalid responses were ruled out. The numbers of valid response of the eight tasks are listed in Table 3. The resultant dataset contains the manual Chinese word segmentation

Table 3 Numbers of valid response of the tasks

Task	Valid response	Percent
1	142	71
2	143	71.5
3	138	69
4	135	67.5
5	133	66.5
6	127	63.5
7	123	61.5
8	127	63.5
Min	123	61.5
Max	143	71.5
Mean	133.5	66.75
SD	7.37	3.68

data of 152 sentences, length of which ranges from 20 to 46 characters ($M = 32.54$, $SD = 5.46$), and each sentence is segmented by at least 123 and at most 143 subjects ($M = 133.5$, $SD = 7.37$).

3.2 Evaluation of experimental data

Although Fleiss' kappa can be used to measure the agreement between raters, high agreement does not necessarily mean high data quality especially in the situation of intuition measurement where variations among subjects are expected. And it cannot show directly how many errors the resultant dataset actually contains either. Knowing how many errors the dataset contains is very important to assess the reliability of the conclusions drawn from the dataset. We firstly define two kinds of manual segmentation errors, and based on that, an evaluation method called manual segmentation error rate (*MSER*) was proposed to evaluate the resultant dataset.

3.2.1 Types of manual segmentation errors

In Chinese phrases/sentences, there are three types of non-monosyllabic segments from the point of view of manual word segmentation: ridiculous segments, indivisible segments, and modest segments. A ridiculous segment usually cannot be treated as one valid unit/word because it makes no sense in the context of the phrase/sentence; for example, in the phrase 这是好东西 *zhè shì hǎo dōngxī* 'this is a good thing', the segment 好东 *hǎo dōng* 'good-east' (NONSENSE) cannot be treated as one unit/word because it is incomprehensible. An indivisible segment usually cannot be divided because it is a fixed unit and its lexical meaning cannot be derived easily from the lexical meanings of its constituents (in other word, semantically opaque); it will become incomprehensible if it is divided; for example, in the phrase example, the segment 东西 *dōngxī* 'thing' is of this type. A modest segment can be either treated as one unit/word or divided into two or more units/words because it is equally comprehensible no matter divided or not; the segment 这是 *zhè shì* 'this is' in the phrase example is of this type.

Two circumstances can be treated as errors of manual word segmentation; firstly, if a ridiculous segment appears in segmentation results, it can be treated as an error (type I error); and secondly, if an indivisible segment is divided in segmentation results, it can also be treated as an error (type II error). These two circumstances are not compatible with our general word intuition even to the least extent because they are simply incomprehensible, and they cannot be explained by variations of word intuition among speakers; normally, when the subjects do word segmentation tasks carefully according to their word intuition, these would not occur and thus we can treat them as errors. Human word segmentation errors will occur when the subjects try to cheat by segmenting randomly or make accidental mistakes.

3.2.2 Manual segmentation error rate

A subject divides the phrase/sentence S into n ($n \in N^+$) segments by n segmentation operations (not $n - 1$; the subject left the remaining segment at the tail as one word; it means the subject had "confirmed" that this is a segmentation operation too). A segmentation operation can only yield one of the following four possible results: one type I error, one type II error, one type I error plus one type II error (two errors; e.g., 好东/西 *hǎo dōng/ xī* 'good-east/west' (NONSENSE'), or no error. Suppose $e'(e' \in N)$ is the number of times the type I error occurred during the segmentation process, and

$e^{''}(e^{''} \in N)$, the number of times the type II error occurred, then we can define manual segmentation error rate (*MSER*):

$$MSER = (e' + e'')/n$$

In extreme cases, *MSER* could be greater than one, for example, in the segmentation result 去哈/尔滨/ *qù Hā/ ěrbīn/* 'go to Ha/er-bin/' (NONSENSE), $e^{'} = 2$, $e^{''} = 1$, $n = 2$, so $MSER = 3/2$. If this happens, we just assume that $MSER = 1$. *MSER* can be used to evaluate manual word segmentation results. Lower *MSER* means better data quality. Let us consider its collective form. If *S* is segmented by m $(m \in N^{+})$ subjects, and the ith $(1 \le i \le m)$ subject's type I error count, type II error count, and segmentation operation count are e_i', e_i'', n_i respectively, then the collective form of *MSER* is:

$$MSER = \frac{\sum_{i=1}^{m} (e_i' + e_i'')}{\sum_{i=1}^{m} n_i}$$

As a convenient way, we can find type I errors and their counts in the unigram frequency list of the segmentation results, and find type II errors and their counts in the bigram frequency list of the segmentation results.

Among the 19 sentences of each task, three sentences were sampled for evaluation: the first sentence, the middle (10th) sentence, and the last (19th) sentence. We calculated the *MSER* for each of them, see Table 4 for details. The *MSER*s of the segmentation results of these sentences are all very low (< .05), and their mean is only .013 (*SD* = .004). This means the resultant dataset only contains few errors and indicates that the data quality is good.

3.3 Representation of word segmentation results

Characters are written symbols which are used to record linguistic units (e.g., morphemes, words, phrases, and sentences). The characters which are used to record Chinese sentences include Chinese characters, punctuations, numbers, and Latin letters. All the characters constituent a character set. A sentence (more precisely, a written sentence) can be treated as a string of characters which follow proper grammatical rules. A grammar is a set of rules which combine characters into sentences. A language is a set of all possible sentences given a character set and a grammar.

In a written sentence, after each character, there is an interval and we call it a character interval. A character interval can be a word-boundary which indicates the end of a word and at the same time the start of the next word if there exists one. And it can also be a non-word boundary which locates inside a word. Each character interval can be treated as a binary variable. When a character interval is a word boundary, we say that it has the value of *one*; when it is a non-word-boundary, it has the value of *zero*. See Huang et al. (2007) and Li and Huang (2009) for the source of this abstraction.

The sentence *S* which consists of n $(n > 0)$ characters $C_1, C_2, C_3, \ldots, C_n$ can be represented as follows:

Table 4 Manual segmentation error rates (MSER) of the segmentation results of the eight tasks

Task	Sentence	$\sum n$	$\sum e^{'}$	$\sum e^{''}$	*MSER*
1	S_1	2864	13	20	.012
	S_{10}	3904	18	16	.009
	S_{19}	4046	12	7	.005
2	S_1	2993	29	19	.016
	S_{10}	2000	9	6	.008
	S_{19}	2529	19	26	.018
3	S_1	6634	32	27	.009
	S_{10}	2834	21	14	.012
	S_{19}	2894	43	22	.022
4	S_1	2612	24	22	.018
	S_{10}	1836	14	8	.012
	S_{19}	2640	26	20	.017
5	S_1	2361	15	14	.012
	S_{10}	2829	14	7	.007
	S_{19}	2489	14	15	.012
6	S_1	2906	35	22	.020
	S_{10}	2758	21	8	.011
	S_{19}	1711	20	13	.019
7	S_1	1857	19	11	.016
	S_{10}	3125	35	14	.016
	S_{19}	2808	28	10	.014
8	S_1	2465	23	14	.015
	S_{10}	3238	23	11	.011
	S_{19}	2042	15	7	.011
	Min	1711	9	6	.005
	Max	6634	43	27	.022
	Sum	68,375	522	353	
	Mean	2848.96	21.75	14.71	.013
	SD	989.76	8.51	6.3	.004

$$S = C_1I_1C_2I_2C_3I_3\ldots C_nI_n$$

$I_1, I_{2,}I_3, \ldots, I_n$ represent character intervals. Suppose that $\{S_1, S_2, S_3, \ldots, S_n\}$ is a set of sentences of the language L. And $\{H_{1,}H_{2,}H_3, \ldots H_m\}$ is a set of human subjects who speak the language L. An intuitive word segmentation task virtually requires each human subject to assign *one* to the character intervals of each sentence which are word boundaries and to assign *zero* to character intervals of each sentence which are non-word boundaries according to intuition.

The results of an intuitive word segmentation task can be summarized into a table in the format illustrated in Table 5. Each row in the table stores the segmentation results of one human subject, and each column stores the segmentation results of one sentence. The x^{th} human subject segments the y^{th} sentence and the segmentation result is $R_{x,y}$. Since the sentence to be segmented is known, it is sufficient to represent the

Table 5 Results of an intuitive word segmentation task

	S_1	S_2	S_3	...	S_n
H_1	$R_{1,1}$	$R_{1,2}$	$R_{1,3}$	...	$R_{1,n}$
H_2	$R_{2,1}$	$R_{2,2}$	$R_{2,3}$	...	$R_{2,n}$
H_3	$R_{3,1}$	$R_{3,2}$	$R_{3,3}$	...	$R_{3,n}$
...	...	...	...	...	...
H_m	$R_{m,1}$	$R_{m,2}$	$R_{m,3}$	...	$R_{m,n}$

segmentation result by just listing the values of the character intervals in the sentence. So $R_{x,y}$ can be treated as a vector $(i_1, i_2, i_3, \ldots)$ in which $i_1, i_2, i_3, \ldots$ represent the values of the first, second, third,... character interval. The number of components of the vector equals the length of the sentence to be segmented *len*(*S*) (in character). And we call this kind of vector the segmentation result vector (*SRV*). Normally, this table should be analyzed column by column (hence sentence by sentence). The $i^{th}(i = 1, 2, 3, \ldots)$ column stores the segmentation results of the i^{th} sentence and it can be treated as a $m \times len(S_i)$ matrix and we call it the segmentation result matrix (*SRM*).

3.4 Calculation of word intuition between/among Chinese speakers

Word segmentation agreement between human subjects reflects the agreement of word intuition between Chinese speakers, i.e., to what extent the Chinese speakers agree with each other on what is a word intuitively. Since there is no single best way to measure agreement, we used several metrics to provide more information: (1) proportionate agreement, (2) Cohen's kappa, and (3) Fleiss' kappa. Suppose that a and b are two human subjects and s is the sentence to be segmented and $len(s) = n$, and the SRVs generated by a and b on sentence s are $R_{a,s} = (a_1, a_2, a_3, \ldots, a_n)$ and $R_{b,s} = (b_1, b_2, b_3, \ldots, b_n)$. We can use proportionate agreement and Cohen's kappa to measure the segmentation agreement between a and b based on their SRVs. When there are more than two human subjects, we can measure the segmentation agreements of all the unique subject pairs one by one and then see the distribution, or alternatively, we can use Fleiss' kappa to measure the overall agreement as a summary.

3.4.1 Proportionate agreement

The proportionate agreement between these two SRVs is defined as the number of same judgments between these two vectors normalized by the maximum possible number of different judgments between two SRVs whose length equals n:

$$PA(R_{a,s}, R_{b,s}) = 1 - \frac{\sum_{i=1}^{n} (a_i - b_i)^2}{n}$$

The range of $PA(R_{a,s}, R_{b,s})$ is [0,1], where 0 means complete disagreement (0% agreement) while 1 means complete agreement (100% agreement).

The crowdsourcing Chinese word segmentation experiment consists of eight sessions (i.e., eight crowdsourcing word segmentation tasks); each session have the same human subject set (see Table 3 for subject numbers of the eight sessions) and the same sentence set (19 sentences per session). Different sessions have different human subject group (there are perhaps partial overlaps) and different sentence set (without overlap).

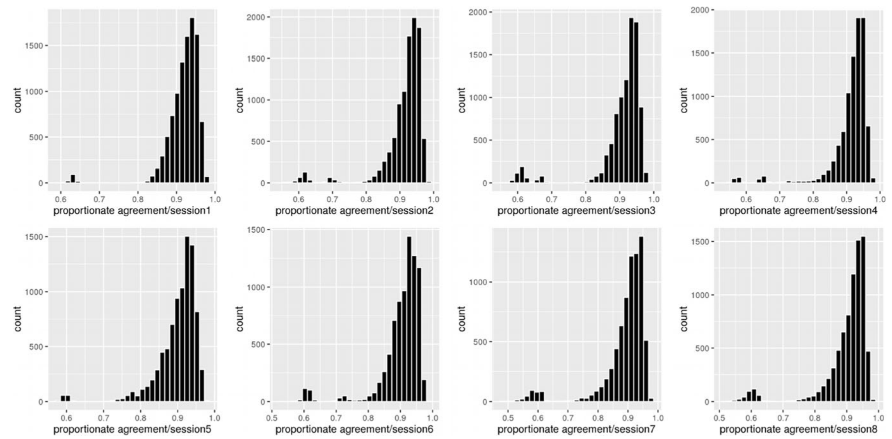

Fig. 1 Word segmentation agreement between human subjects: proportionate agreement

For each session, we calculate the proportionate agreement for each unique human subject pair. A session has 19 sentences; hence, a human subject generates 19 SRVs. These SRVs are concatenated into one general SRV to represent the segmentation behavior of the human subject; the calculation is based on the general SRVs of the human subjects. See Fig. 1 for the distributions of proportionate agreement statistics of the eight sessions, and see Table 6 for the proportionate agreement statistics of the eight sessions. Because of the existence of outliers (see Fig. 1), the medians summarize the statistics better than the means. As a summary, the word intuition agreement measured by proportionate agreement ranges from 0.91 to 0.93 ($M = 0.92$, $SD = 0.001$).

3.4.2 Cohen's kappa

Cohen's kappa (Cohen 1960) is another metric which measures the rating agreement between two raters who classify the same set of objects into several categories. It is believed to be more reasonable than proportionate agreement since it takes the agreement by chance into account. For this reason, Cohen's kappa is more conservative than proportionate agreement. It is calculated using the following formula:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

in which " p_o is the observed proportion of agreement, and p_e is the proportion of

Table 6 Summaries of proportionate agreement statistics of the eight sessions

Session	Min	Max	Mean	Median	SD
1	0.59	0.98	0.92	0.93	0.046
2	0.54	0.99	0.91	0.93	0.066
3	0.55	0.99	0.91	0.93	0.076
4	0.54	0.99	0.91	0.93	0.065
5	0.58	0.98	0.9	0.91	0.057
6	0.52	0.99	0.9	0.92	0.068
7	0.48	0.98	0.89	0.91	0.081
8	0.52	0.98	0.9	0.92	0.078

Table 7 Interpretation of kappa statistics

Kappa statistic	Strength of agreement
< 0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

agreement expected by chance" (Cohen 1968). See Cohen (1960) for the details of the definition and calculation of Cohen's kappa. To interpret it, two subjects are in complete agreement when $\kappa = 1$, and in complete disagreement when $\kappa \leq 0$. Landis and Koch (1977) provide a scheme for the interpretation of κ statistic (see Table 7). But it is worth noting that the authors also pointed that "these divisions are clearly arbitrary", but "they do provide useful 'benchmarks'" and this scheme can help to "maintain consistent nomenclature when describing the relative strength of agreement associated with kappa statistics" (Landis and Koch 1977).

Following the same procedure of the calculation of proportionate agreement, we calculated Cohen's kappa for the eight sessions. See Fig. 2 for the distributions of the Cohen's kappa statistics of the eight sessions, and see Table 8 for the statistics. Because of the existence of outliers (see Fig. 2), the medians summarize the statistics better than the means. As a summary, the word intuition agreement measured by Cohen's kappa ranges from 0.82 to 0.86 ($M = 0.84$, $SD = 0.02$). According to Table 7, this means almost perfect agreement.

3.4.3 Fleiss' kappa

The agreement metrics we discussed above, proportionate agreement and Cohen's kappa, are only used to measure the agreement between two subjects. When measuring the agreement among three or more subjects, Fleiss' kappa (Fleiss 1971) should be used instead. It is calculated using the following formula:

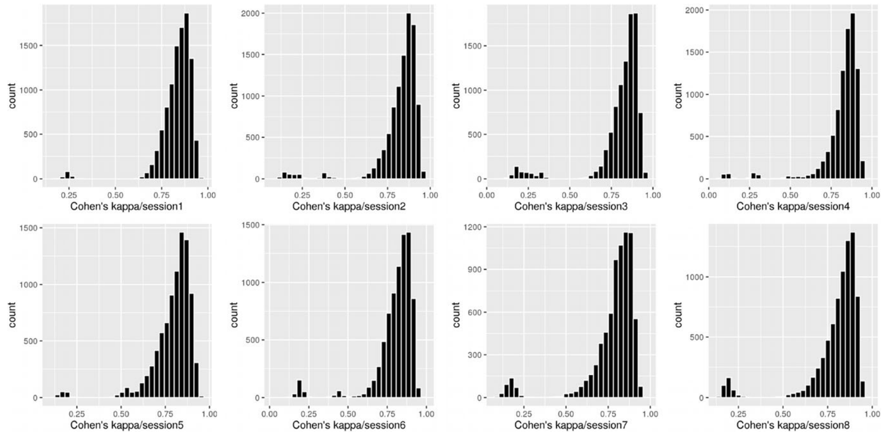

Fig. 2 Word segmentation agreement between human subjects: Cohen's kappa

Table 8 Summaries of Cohen's kappa statistics of the eight sessions

Session	Min	Max	Mean	Median	SD
1	0.15	0.96	0.83	0.85	0.092
2	0.097	0.97	0.82	0.86	0.14
3	0.075	0.97	0.8	0.85	0.15
4	0.058	0.97	0.82	0.85	0.13
5	0.11	0.96	0.79	0.82	0.11
6	0.019	0.97	0.8	0.83	0.14
7	0.074	0.96	0.78	0.82	0.16
8	0.13	0.97	0.8	0.84	0.15

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

"$1 - \bar{P}$ measures the degree of agreement attainable over and above what would be predicted by chance" (Fleiss 1971), $\bar{P} - \bar{P}_e$ measures "the degree of agreement actually attained in excess of chance" (Fleiss 1971), and the whole equation is a "normalized measure of overall agreement, corrected for the amount expected by chance" (Fleiss 1971). The interpretation of Fleiss' kappa statistics accords with that of Cohen's kappa, and Table 7 is applicable too. See Table 9 for the Fleiss' kappa statistics of the eight sessions. The mean of the Fleiss' kappa statistics is 0.78 ($SD = 0.02$). According to Table 7, this means substantial agreement but it is very close to the threshold value of almost perfect agreement (0.81). Because there are outliers in the word segmentation results (see Figs. 1 and 2) which would reduce the Fleiss' kappa statistics, we could interpret the Fleiss' kappa value 0.78 as almost perfect agreement.

3.5 Role of semantic transparency in word intuition agreement

In order to examine the role of semantic transparency in the word intuition of Chinese speakers, we firstly extracted the most typical compound stimuli of all the semantic transparency types from the compound stimuli of the Chinese word segmentation

Table 9 Fleiss' kappa statistics of the eight sessions

Session	Fleiss' kappa
1	0.81
2	0.8
3	0.78
4	0.79
5	0.77
6	0.77
7	0.75
8	0.77
Min	0.75
Max	0.81
Mean	0.78
Median	0.78
SD	0.02

Table 10 Distribution of types of most typical compound

Transparency type	Word structure		
	NN	AN	VN
TT	6	3	3
TO	6	3	3
OT	6	3	3
OO	6	3	3

experiment (see Table 1) according the semantic transparency rating data of a laboratory-based semantic transparency rating experiment (Wang et al. 2015). See Table 10 for the distribution of the extracted word stimuli among semantic transparency types and structural types.

Then, we checked the segmentation results of these compounds to see if the transparent compounds have higher probability to be explicitly segmented into two units than the opaque compounds. Tables 11, 12, 13, and 14 show the segmentation results of four types of compound stimuli: transparent compounds (TT), two types of partially transparent compounds (TO and OT), and opaque compounds (OO). In each table, the first column ("Type") indicates the semantic transparency type and structural types of the compound stimuli. The second column ("Compound") lists the compound stimuli. The third column ("#Resp") means the number of the valid responses the sentences contain the compound stimuli received. Column four ("#AB") shows how many times the compound stimuli were explicitly segmented as one unit (or one word). Column five ("#A/B") indicates how many times the compound stimuli were explicitly segmented as two units (or two words). And there could also be other segmentation results, column six ("#Other") shows the counts of other segmentation results (see Wang (2016: 145) for the details of other segmentation results).

The possibilities of the four types of compound stimuli to be segmented into two units are all very close to zero; we found no evidence to support the hypothesis that transparent compounds have more chances to be segmented into two units than opaque ones. For example the word intuition agreement of the transparent compound *jiāngshuǐ* is 0.99 (141/142) and

Table 11 Segmentation results of typical TT compounds

Type	Compound	#Resp	#AB	#A/B	#Other
TT-AN	好事 *hǎoshì* 'good deed'	142	138	2	2
TT-AN	亮光 *liàngguāng* 'light'	133	131	0	2
TT-AN	冷水 *lěngshuǐ* 'cold water'	127	103	2	22
TT-NN	草地 *cǎodì* 'grassland'	142	78	1	63
TT-NN	江水 *jiāngshuǐ* 'river water'	142	141	0	1
TT-NN	琴声 *qínshēng* 'tweedle'	143	140	2	1
TT-NN	米饭 *mǐfàn* 'rice'	133	131	1	1
TT-NN	火灾 *huǒzāi* 'fire disaster'	133	132	0	1
TT-NN	煤矿*méikuàng* 'coal mine'	123	115	0	8
TT-VN	裂缝 *lièfèng* 'fissure'	138	134	0	4
TT-VN	笑声 *xiàoshēng* 'laughter'	123	120	1	2
TT-VN	借款 *jièkuǎn* 'borrowed money'	127	115	0	12

Table 12 Segmentation results of typical TO compounds

Type	Compound	#Resp	#AB	#A/B	#Other
TO-AN	甜点 *tiándiǎn* 'dessert'	142	137	0	5
TO-AN	新星 *xīnxīng* 'nova'	133	129	1	3
TO-AN	歪风 *wāifēng* 'unhealthy tendency'	127	123	0	4
TO-NN	灯泡 *dēngpào* 'lamp bulb'	143	140	1	2
TO-NN	人海 *rénhǎi* 'sea of faces'	138	126	4	8
TO-NN	音色 *yīnsè* 'timbre'	138	128	0	10
TO-NN	福气 *fúqi* 'good fortune'	135	132	1	2
TO-NN	河床 *héchuáng* 'riverbed'	133	111	0	22
TO-NN	梦乡 *mèngxiāng* 'dreamland'	127	125	0	2
TO-VN	救星 *jiùxīng* 'savior'	142	140	0	2
TO-VN	助手 *zhùshǒu* 'assistant'	133	131	0	2
TO-VN	销路 *xiāolù* 'sale'	127	122	0	5

that of the opaque compound *lóngyǎn* is 0.97 (129/133); there is no significant difference. Based on our data, we cannot say that semantic transparency plays no role in word intuition of Chinese speakers, but even if semantic transparency affects the word intuition of Chinese speakers, its role is rather restricted. We also found some special forms, for example monosyllabic verb + disyllabic noun (没冷水 *méi lěngshuǐ* 'no cold water', 拿借款 *ná jièkuǎn* 'fetch borrowed money', 喝开水 *hē kāishuǐ* 'drink boiled water', 抽大麻 *chōu dàmá* 'smoke marijuana'), disyllabic noun + monosyllabic localizer (草地上 *cǎodì shàng* 'on the grassland', 天桥上 *tiānqiáo shàng* 'on the overpass'), and some other forms (脾气大 *píqi dà* 'violent-tempered', 天王地位 *tiānwáng dìwèi* 'super star status', and 幕后黑手 *mùhòu hēishǒu* 'black hand behind the scenes'). These forms are usually treated as phrases theoretically; however, they all show significant chances to be treated as one word intuitively. These forms await further studies.

4 Discussion

We measured word intuition agreement among Chinese speakers based on the measurement of word segmentation agreement. Various metrics show that Chinese speakers agree

Table 13 Segmentation results of typical OT compounds

Type	Compound	#Resp	#AB	#A/B	#Other
OT-AN	白菜 *báicài* 'Chinese cabbage'	135	134	0	1
OT-AN	金鱼 *jīnyú* 'goldfish'	133	132	0	1
OT-AN	贵人 *guìrén* 'magnate'	127	123	2	2
OT-NN	天才 *tiāncái* 'genius'	142	142	0	0
OT-NN	天王 *tiānwáng* 'heavenly king'	143	128	0	15
OT-NN	法宝 *fǎbǎo* 'magic weapon'	143	140	0	3
OT-NN	天桥 *tiānqiáo* 'overpass'	123	74	0	49
OT-NN	轮船 *lúnchuán* 'steamer'	123	110	0	13
OT-NN	花灯 *huādēng* 'festival lantern'	123	116	0	7
OT-VN	发票 *fāpiào* 'invoice'	143	136	0	7
OT-VN	拖鞋 *tuōxié* 'slippers'	133	129	0	4
OT-VN	开水 *kāishuǐ* 'boiled water'	127	105	0	22

Table 14 Segmentation results of typical OO compounds

Type	Compound	#Resp	#AB	#A/B	#Other
OO-AN	黑手 *hēishǒu* 'black hand'	138	112	0	26
OO-AN	热线 *rèxiàn* 'hotline'	138	130	1	7
OO-AN	大麻 *dàmá* 'marijuana'	133	96	0	37
OO-NN	色狼 *sèláng* 'masher'	142	133	0	9
OO-NN	脾气 *píqi* 'temperament'	138	109	0	29
OO-NN	粉刺 *fěncì* 'acne'	135	133	0	2
OO-NN	手下 *shǒuxià* 'heeler'	135	128	1	6
OO-NN	龙眼 *lóngyǎn* 'longan'	133	129	0	4
OO-NN	风头 *fēngtóu* 'trend of events'	127	117	1	9
OO-VN	通货 *tōnghuò* 'currency'	143	140	0	3
OO-VN	起色 *qǐsè* 'improvement'	135	127	1	7
OO-VN	炒家 *chǎojiā* 'speculator'	127	119	1	7

with each other almost perfectly on what is a word. Measured by proportionate agreement, the word intuition agreement between Chinese speakers is about 0.9 on average; measured by Cohen's kappa, the word intuition agreement between Chinese speakers is about 0.8~0.9 on average; and measured by Fleiss' kappa, the word intuition agreement among Chinese speakers is about 0.8 which would be higher if we further filter out outliers. There are word intuition differences among Chinese speakers, but the differences are not as large as we thought. These statistics strongly support the psychological reality of Chinese word and suggest that the concept of word in Chinese linguistics has solid psychological foundation in Chinese speaking community. We also studied the role of semantic transparency in word intuition agreement; we found that at least in terms of the compounds we examined there is no evidence to support certain semantic transparency effect on word intuition agreement. Although there are some debates among linguists on the wordhood of semantically transparent forms which consist of free forms such as *jiāngshuǐ*, there is no intuitive divergence among Chinese speakers even to the least extent. Such high word intuition agreement also suggests that it is quite feasible to formulate a definition of Chinese word according to the collective word intuition of Chinese speakers. And such a definition of Chinese word will be quite different from the classic word definition (i.e.,"minimum free form"). In addition, the data collected in this study can be probably annotated lexical resource to support computational word segmentation task in the future. Last, but not the least, although kappa based agreement measure cannot be directly compared with *F*-scores, the fact that word intuition agreement is about 0.9 among native speakers suggests that perhaps the pursuit for 0.97+ *F*-score in current Chinese segmentation bakeoff competition could be a result of overfitting rather than real improvements in methodology.

Acknowledgements
We would like to thank the reviewers for their constructive comments and suggestions. This work is supported by General Research Fund (GRF) of the Research Grants Council of the Hong Kong SAR, China (Project No. 544011), Shandong University Youth Team Project (Project No. IFYT17005), and Shandong Social Science Planning Fund Program (Project No. 17DYYJ06).

Authors' contributions
All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]School of Literature, Shandong University, Jinan, China. [2]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hung Hom, Hong Kong.

Received: 24 April 2017 Accepted: 6 November 2017
Published online: 20 December 2017

References

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis* 20(3): 351–368.

Bloomfield, Leonard. 1933. *Language*. New York: Holt, Rinehart and Winston.

Buhrmester, Michael, 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.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1): 37–46.

Cohen, Jacob. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* 70(4): 213.

Crump, Matthew J. C., 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.

Enochson, Kelly, and Jennifer Culbertson. 2015. Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PLoS One* 10(3): e0116946, 03.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5): 378.

Hoosain, Rumjahn. 1992. Psychological reality of the word in Chinese. *Advances in Psychology* 90: 111–130.

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

Huang, Chang-ning, and Hai Zhao 黄昌宁, 赵海. 2007. Chinese word segmentation: A decade review 中文分词十年回顾. *Journal of Chinese Information Processing* 中文信息学报 21(3): 8–19.

Huang, Chu-Ren, Keh-jiann Chen, and Lili Chang. 1996. Segmentation standard for Chinese natural language processing. In *Proceedings of the 16th International Conference on Computational Linguistics*, 5–9. Copenhagen: Denmark.

Huang, Chu-Ren, Petr Šimon, Shu-Kai Hsieh, and Laurent Prévot. 2007. Rethinking Chinese word segmentation: Tokenization, character classification, or wordbreak identification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 69–72. Stroudsburg: Association for Computational Linguistics.

Huang, Chu-Ren, and Nianwen Xue. 2012. Words without boundaries: Computational approaches to Chinese word segmentation. *Language and Linguistics Compass* 6(8): 494–505.

Huang, Chu-Ren, and Nianwen Xue. 2015. Modeling word concepts without convention: Linguistic and computational issues in Chinese word identification. In *The Oxford handbook of Chinese linguistics*, ed. William S.-Y, Wang and Chaofen Sun, 348–361. New York: Oxford University Press.

Landis, J. Richard, and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33(1): 159–174.

Li, Shoushan, and Chu-Ren Huang. 2009. Word boundary decision with CRF for Chinese word segmentation. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, 726–732. Hong Kong.

Libben, Gary, Martha Gibson, Yeo Bom Yoon, and Dominiek Sandra. 2003. Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language* 84(1): 50–64.

Liu, Yuan, and Nanyuan Liang 刘源, 梁南元. 1986. Foundation of Chinese language processing: Modern word frequency statistics 汉语处理的基础工程——现代词频统计. *Journal of Chinese Information Processing* 中文信息学报 1: 17–25.

Liu, Yuan, Qiang Tan, and Xukun Shen 刘源, 谭强, 沈旭昆. 1994. *Contemporary Chinese language word segmentation specification for information processing and automatic word segmentation methods* 信息处理用现代汉语分词规范及自动分词方法. Beijing: Tsinghua University Press.

Mason, Winter, and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44(1): 1–23.

Munro, Robert, Steven Bethard, Victor Kuperman, Vicky T. 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*, 22–130. Stroudsburg: Association for Computational Linguistics.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5(5): 411–419.

Schnoebelen, Tyler, and Victor Kuperman. 2010. Using Amazon Mechanical Turk for linguistic research. *Psihologija* 43(4): 441–464.

Simcox, Travis, and Julie A. Fiez. 2014. Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behavior Research Methods* 46(1): 95–111.

Sproat, Richard, William Gale, Chilin Shih, and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics* 22(3): 377–404.

Sprouse, Jon. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43(1): 155–167.

Wang, Li 王立. 2003. *Socio-linguistic investigation of Chinese word* 汉语词的社会语言学研究. Beijing: The Commercial Press.

Wang, Shichang. 2016. Crowdsourcing method in empirical linguistic research: Chinese studies using Mechanical Turk-based experimentation. PhD thesis. Hong Kong: The Hong Kong Polytechnic University.

Wang, Shichang, Chu-Ren Huang, Yao Yao, and Angel Chan. 2014a. 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*, 147–156. Dublin: Association for Computational Linguistics and Dublin City University.

Wang, Shichang, Chu-Ren Huang, Yao Yao, and Angel Chan. 2014b. Exploring mental lexicon in an efficient and economic way: Crowdsourcing method for linguistic experiments. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, 105–113. Dublin: Association for Computational Linguistics and Dublin City University.

Wang, Shichang, Chu-Ren Huang, Yao Yao, and Angel Chan. 2015. Mechanical Turk-based experiment vs laboratory-based experiment: A case study on the comparison of semantic transparency rating data. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC-29)*, 53–62. Shanghai: China.