The effect of morphological structure on semantic transparency ratings

Shichang Wang[1], Chu-Ren Huang[2], Yao Yao[2] and Angel Chan[2]
[1] Shandong University / [2] The Hong Kong Polytechnic University

Semantic transparency deals with the interface between lexical semantics and morphology. It is an important linguistic phenomenon in Chinese in the context of prediction of meanings of compounds from their constituents. Given prominence of compounding in Chinese morpho-lexical processes, to date there is no semantic transparency dataset available to support verifiable and replicable quantitative analysis of semantic transparency in Mandarin Chinese. In addition, the relation between semantic transparency and morphological structure has not been systematically examined. This paper reports a crowdsourcing-based experiment designed for the construction of a large semantic transparency dataset of Chinese compounds which includes semantic transparency ratings of both the compound and each constituent root of the compound. We also present an analysis of the effects of morphological structure on semantic transparency using the constructed dataset. Our study found that in a transparent modifier-head compound, the head tends to get greater semantic transparency rating than the modifier. Interestingly, no such effect is observed in coordinative compounds. This result suggests that compounds of different morphological structures are processed differently and that the concept of head plays an important role in the word-formation process of compounding. We advocate that crowdsourcing can be a highly instrumental method to collect linguistic judgments and to construct language resources in Chinese language studies. In addition, the proposed methodology of comparing constituent transparency and word transparency sheds light on the relation between morpho-lexical structure and cognitive processing of lexical meanings.

Keywords: Mandarin Chinese, constituent semantic transparency, compound word-formation, headedness, crowdsourcing

https://doi.org/10.1075/lali.00035.wan
Language and Linguistics 20:2 (2019), **225–255**
ISSN 1606-822X / E-ISSN 2309-5067

1. Introduction

Semantic transparency deals with the interface between lexical semantics and morphology. It is an important linguistic phenomenon in Chinese in the context of prediction of meanings of compounds from their constituents. However, the meaning composition relation of compounds in Mandarin Chinese can range from completely transparent to completely opaque. The meaning of 馬虎 *mǎhu* 'careless' (lit. 'horse-tiger') has nothing to do with either 馬 *mǎ* 'horse' or 虎 *hǔ* 'tiger'. However the meaning of 道路 *dàolù* 'road' (lit. 'way-road') is basically equal to 道 *dào* 'way' or 路 *lù* 'road'. There are many more examples in which the semantic composition is neither completely transparent nor completely opaque. For instance, although the meaning of 江湖 *jiānghú* 'all corners of the country' (lit 'river-lake') is not equal to 江 *jiāng* 'river' plus 湖 *hú* 'lake', but a relatedness between the meaning of the compound and a concatenation of the two constituent meanings can be observed. This phenomenon involving the compositionality of meaning of compounds from its constituents is called semantic transparency of compounds in the literature. Theoretically, it involves the interface between morphology and lexical semantics. The relationship between morphology and lexical semantics is "clearly a rich, though underexplored, area of study" (Levin & Hovav 2001: 267). As a typical and major issue in this area, semantic transparency of compounds has not been sufficiently explored.

The concept of semantic transparency was discussed in the literature dating back to the 1970s (Aronoff 1976; Allen 1979). Yet the term did not become popular until the 1990s when psycholinguists started to investigate the relations between semantic transparency and the mental lexicon, especially the role of semantic transparency in the representation and processing of compounds (Marslen-Wilson et al. 1994; Zwitserlood 1994; Schreuder & Baayen 1995; Tsai 1996; Libben 1998; Feldman & Pastizzo 2003; Libben et al. 2003; Myers et al. 2004b; Pollatsek & Hyönä 2005; Frisson et al. 2008; Mok 2009; Han et al. 2014). There are also many articles dealing with semantic transparency in Chinese linguistic literature (Wang & Peng 1999; 2000; Gao & Gao 2005; Li & Li 2008; Li 2011; Ren 2012; Song 2013).

It is interesting to note that there is not a single unified, widely-accepted definition of semantic transparency in spite of the rich literature in both theoretical and psycholinguistics. Marslen-Wilson et al. (1994) defined the semantic transparency of a complex word as whether its meaning is "synchronically compositional". Zwitserlood (1994) defined semantic transparency of a compound as whether its meaning is "synchronically related to the meaning of its composite words". Schreuder & Baayen (1995) proposed to model semantic transparency based on the "overlap between the set of (semantic) representations of the complex word and the sets of representations of its constituents". Pollatsek & Hyönä (2005) thought

that the "approximate meaning" of a semantically transparent compound "can be derived from the constituent meanings by 'gluing' them together"; on the other hand, the meaning of a semantically opaque compound "cannot be computed by simply gluing together constituent meanings". Libben (1998) said that a compound is semantically transparent if its meaning is "predictable from the meaning of the constituents". Plag (2003) also used the word "predictable" but from a different angle: words are semantically transparent if "their meanings are predictable on the basis of the word-formation rule according to which they have been formed"; in the Chinese literature on semantic transparency, most scholars (Wang & Peng 1999; Li & Li 2008; Ren 2012; Song 2013) adopt the same range of variations of definition.

Libben et al. (2003) introduced an additional dimension to semantic transparency by asking the important question of "whether semantic transparency is best viewed as a property of the entire multimorphemic string or as a property of constituent morpheme". Hence they proposed to distinguish between the semantic transparency of compounds and the semantic transparency of their individual constituent morphemes. This distinction is not only necessary but also very important, for without the concept of semantic transparency of individual morphemes we cannot capture the internal structures of semantic transparency of compounds. An individual constituent morpheme is semantically transparent if its meaning is "transparently represented in the meaning of the compound as a whole" (Libben et al. 2003). We adopt their position and call the semantic transparency of the entire compound "overall semantic transparency" and the semantic transparency of its individual morphemes "constituent semantic transparency".

Zwitserlood (1994) claimed that "semantic transparency of compounds is defined by the semantic relation between a compound and its component morphemes". In contrast, Aronoff (1976: 32) argued that the semantic relationship between a word and its base "will seldom be one of neat compositionality" and there usually is "some sort of divergence" and "this divergence is not between the derivative and the base, but rather between the actual meaning of the derivative and the meaning we expect it to have." In this view then one may wonder whether semantic transparency can be viewed as the semantic relation between a compound and its constituent morphemes or should it be viewed as the semantic relation between the actual meaning of the compound and the 'expected' meaning? We think the latter is better. The meaning of a compound we expect it to have is actually its compositional meaning which is computed by the meanings of its constituent morphemes and the morphological and semantic structures between them. In a compound, constituent morphemes are combined together by grammatical and semantic structures and there is no reason to neglect these structures but only use the meanings of constituent morphemes. Therefore, we view the overall semantic transparency of a compound as some kind of semantic relation between its actual

meaning and its compositional meaning. In this way, the overall transparency of a compound can be clearly differentiated from the constituent semantic transparency, which can be calculated in terms of the distance between the constituent meaning and compound meaning for each constituent.

The word "transparency" can be more easily comprehended using a metaphor in optics, in which transparency is the physical property of a material which allows light to pass through. As a metaphor, transparency in semantics resembles transparency in optics. A compound word can be modeled to equip with an interpretation function $I(x)$ which computes its actual meaning from its compositional meaning. Compositional function $C(x)$ combines the meanings of constituents and the morphological and semantic structures between them to predict the meaning of the compound. Suppose *ab* is a compound and *S* is its actual meaning, then $S = I(C(ab))$. The semantic transparency of a compound is actually a property of its interpretation function or the relation between the actual meaning and the compositional meaning. A transparent interpretation function allows compositional meaning to pass through in full without distortion, while an opaque one does not allow it to pass through at all and instead project a totally unrelated meaning. In most cases, a semi-transparent interpretation function only allows compositional meaning to pass through partially (or to use the metaphor again, passing through the compositional meaning with distortion). So for a transparent compound, its actual meaning is (roughly) equal to its compositional meaning. Here 'roughly' is crucial as we do not have a direct way to verify the actual non-realized compositional meaning. Since the compositional meaning of a compound contains the meanings of its constituents, we can also analyze to what extent the meaning of a constituent morpheme passes through. This results in the analysis of constituent semantic transparency. We propose the following definitions of overall and constituent semantic transparency. The semantic transparency of a compound, i.e. the overall semantic transparency (OST), is the extent to which the actual meaning of the compound is similar to its compositional meaning. The semantic transparency of a constituent of a compound, i.e. the constituent semantic transparency (CST), is the extent to which the constituent retains its meaning in the actual meaning of the compound.

If we assign zero to "fully opaque" and one to "fully transparent", then semantic transparency can be quantified as a continuum from zero to one. Two kinds of measurement methods can be found in literatures, i.e. the experimental method and the computational method. Semantic computation is still far from maturity; this limits the reliability of the computational method. Experimental method includes the traditional laboratory experimental method and the emerging crowdsourcing experimental method (Wang et al. 2014; Huang & Wang 2016, among others).

The standard method to measure semantic transparency in psycholinguistics is laboratory-based rating experiment, as implemented by Libben et al. (2003). In

their study, 91 undergraduate students were asked to rate a list of compounds. They participated in two tasks which measured overall semantic transparency and constituent semantic transparency respectively. In the first task, they were asked to "rate each compound in terms of the extent to which its meaning was predictable from the meanings of its parts" on a four-point scale. In the second task, they were asked to rate "the extent to which the constituent retained its individual meaning in the whole word". Again, a four-point scale was employed. Based on the two kinds of transparency scores, they divided these compounds into four types: TT, OT, TO, and OO where 'T' means 'transparent' and 'O' means 'opaque'. Another typical paradigm was proposed by Wang & Peng (1999). Two hundred Chinese undergraduate students were asked to rate 1,500 two-character Chinese words. They were asked to rate the extent to which the meanings of the first character and the second character in a word were related to the meaning of the word respectively on a nine-point scale. Each word was rated by 20 participants. The average score of the 20 scores of each character was used as the final score of each character, and the average score of the two final scores of the two characters was used as the semantic transparency score of the word. These two studies have noticeable differences. Libben et al. (2003) measured both overall semantic transparency and constituent semantic transparency separately while Wang & Peng (1999) only directly measured constituent semantic transparency and took word transparency as the aggregation of constituent transparency. In addition, they used different elicitation questions in their study of constituent transparency. Libben et al. (2003) used a four-point scale to achieve discrete categories of T and O; while the latter used a nine-point scale. Subjects in both studies were undergraduate students, although the number of subjects differ (91 vs. 20). Neither subject pools are big enough to offer robust results of the long list of stimuli.

The quantitative analysis and modeling of semantic transparency must be supported by proper semantic transparency datasets in order to be reusable and verifiable. Some previous studies on the semantic transparency of Chinese compounds were based on datasets too small and restricted, either in terms of number of subjects or number of compounds, to be more reusable (e.g. Xu & Li (2001); Myers et al. (2004a); Gan (2008); Mok (2009)). Some datasets, although large enough and useable for other studies, are not publicly accessible, for example Wang & Peng (1999); Gao & Gao (2005). A large and publicly accessible semantic transparency dataset of Chinese compounds is still a gap in Chinese language resources. However, large linguistic dataset construction is very time- and resource-intensive, especially for the tasks requiring large amounts of human raters. Thus an alternative experimental paradigm is needed to allow construction of a large semantic transparency dataset of Chinese compounds.

Crowdsourcing, an emergent distributed problem-solving strategy designed to leverage the current highly connected population to resolve subject pool size

constraints as well as to eliminate experimenter bias, is adopted in our study (Wang et al. 2014; Huang & Wang 2016). Crowdsourcing experiments obtain and organize flexible human resources and realizes collaboration through Internet and by which, based on mutual benefits, requesters outsource their jobs to crowds of workers via open call on crowdsourcing platforms. The efforts of the crowds are combined to complete the jobs or the best solutions to the jobs will be selected and adopted. Although the crowdsourcing environment is not as controllable as the laboratory environment in many aspects as it contains noise by nature, it has some attractive merits. It is easier to access large crowds of diversified participants on the web beyond spatial and temporal limitations, and it is usually much cheaper than laboratory experiments. It also in general greatly reduces experimenter bias. Hence it is particularly appropriate for tasks which require large amounts of diversified participants and/or aim to process large amounts of data in an economical way both in expenditure and time. Laboratory and crowdsourcing experiments share the same basic principles and design, but quality control are more crucial in the crowdsourcing environment than the laboratory environment. Schnoebelen and Kuperman (2010) collected semantic transparency judgments of phrasal verbs using Amazon's Mechanical Turk (AMT), and they found that it can provide data "comparable in a number of parameters to the data obtained in the lab". Reddy et al. (2011) also used AMT to collect semantic transparency judgments of English compound nouns; they found that "the inter annotator agreement is high and the standard deviation of most tasks is low", so they believed the data were reliable.

Lastly, we need a reliable measurement of semantic transparency to establish and test linguistic model and hypothesis. Semantic transparency effect has been reported in a series of articles to be a crucial factor in the processing of compounds in the mental lexicon (Zwitserlood 1994; Libben 1998; Wang & Peng 1999; Xu & Li 2001; Gao & Gao 2005; Pollatsek & Hyönä 2005; Frisson et al. 2008; Gan 2008; Han et al. 2014). However, how semantic transparency is perceived or processed cognitively is still not clear. Semantic transparency affects the processing of compounds, but what affects the processing of semantic transparency? Since semantic transparency deals with semantic compositionality at the morpho-lexical level, that the morpho-lexical structure of compounds may play a role in the perception and processing of semantic transparency. Hence the working hypothesis in our paper is that there is a morphological structure effect on semantic transparency ratings. We further hypothesize that the different positions of the two constituents of a compound may play a role in how they contribute to the meaning of compound and hence play a role in semantic transparency. It is noted that among all types of Chinese compounds, the modifier-head is the most productive morpho-lexical structure and accounts for about 54.4% of the total. In addition, the coordinative structure is the second most productive one and consists of 21.2% of all Chinese

compounds (Yuan & Huang 1998). In the coordinative structure, the two constituents have equal status, but they have different status in the modifier-head pattern. In modifier-head compounds, one of the constituents is head and carries the semantic category information. This contrast allows us to test whether headedness plays a role in the constituent semantic transparency ratings of compounds. If proven, our study will also provide evidence to support the psychological reality of the linguistic concept of head.

2. Building a semantic transparency dataset

2.1 Method

2.1.1 *Materials*

The following criteria are adopted to select compounds for our study: (1) they must be disyllabic nominal compounds; (2) each of them has the structure of NN, AN, or VN; (3) they are composed of free morphemes; (4) they have mid-range word frequencies; and (5) they are used in both Mainland China and Taiwan. To meet the above five criteria, the following procedure is adopted in implementation:

1. Extract monosyllabic nouns, adjectives and verbs according to *The Contemporary Chinese Dictionary* (2012), *A Dictionary of Modern Chinese Grammar Information* (2017), and our linguistic intuition sometimes; thus we get three sets: (a) the set of monosyllabic nouns, N (n = 604); (b) the set of monosyllabic adjectives, A (n = 312); and (c) the set of monosyllabic verbs, V (n = 1,362). Monosyllabic words are typically free morphemes, and will be treated as such in our study as further analysis often does not yield any clearer picture.
2. Extract the words of the structure NN, AN, or VN (see Yuan & Huang (1998) and Huang (1998) for relevant statistics) from the *Lexicon of Common Words in Contemporary Chinese* (2008). In this step, NN means both morphemes of the word appear in the set N; AN means the first morpheme appears in the set A and the second appears in the set N; VN means the first morpheme appears in the set V and the second appears in the set N. After this step, we get "word list 1".
3. Extract the words which have mid-range frequencies from the Sinica Corpus 4.0 (Chen et al. 1996). We use cumulative frequency feature to determine mid-range frequency. Sort the word frequency list of Sinica Corpus 4.0 in descending order; then calculate cumulative frequency word by word until each word corresponds with a cumulative frequency value; finally, plot a curve on a coordinate plane whose x-axis represents the ranks of words in the sorted list, and the y-axis represents cumulative frequency values. Intuitively, this curve

can be divided into three successive phases; the words within each phase have similar word frequency features. According to this, we identify three word frequency categories, 5,163 high-frequency words (frequency range: [182, 581,823], cumulative frequency range: [0%, 80%]), 19,803 mid-range frequency words (frequency range: [23, 181], cumulative frequency range: (80%, 93%]), and 177,496 low-frequency words (frequency range: [1, 22], cumulative frequency range: (93%, 100%]); Sinica Corpus 4.0 contains about 11.2 million word tokens. The extracted words are represented in traditional Chinese characters. We convert them into simplified Chinese characters and only reserve the words which also appear in "word list 1". After this step, we get "word list 2".

4. Manually verify "word list 2" to generate the final list. Things needing to be verified include the following aspects. (a) Because in "word list 2" word structures are judged automatically, there are many errors, so we have to verify the correctness of the word structure judgments. (b) We have to make sure that the morphemes of each word are free morphemes (because a morpheme can be free in some meanings and bound in others). (c) We also need to delete some proper names.

The words we selected appear in both Sinica Corpus 4.0 and *Lexicon of Common Words in Contemporary Chinese* (2008). Since there is no completely reliable criterion to identify Chinese words, appearing in two lexicons ensures their word identity. This also ensures that they are used in both Mainland China and Taiwan, and further means they are quite possible to be shared in other Chinese language communities, for example Hong Kong, Macau, and Singapore.

According to the above criteria and procedure, we selected a total of 1,176 words. 664 (56.46%) of them have the structure NN; 322 (27.38%) have the structure AN; and 190 (16.16%) have the structure VN. According to our analysis, 1,053 (89.54%) words of them have the structure "modifier-head", for example 美人 *měirén* 'beauty' (lit. 'beautiful-person'), 野花 *yěhuā* 'wild flower' (lit 'wild-flower'); 107 (9.1%) words have the structure "coordination", for example 歌舞 *gēwǔ* 'song and dance' (lit. 'song-dance'), 山河 *shānhé* 'mountains and rivers' (lit. 'mountain-river'); and only 16 (1.36%) words have other structures, for example the words of "verb-object" structure, 指南 *zhǐnán* 'guide' (lit. 'indicate-south'), 知音 *zhīyīn* 'bosom friend' (lit. 'know-sound).

Normally, a crowdsourcing experiment should be reasonably small in size. We randomly divide these 1,176 words into twenty-one 56-word groups G_i (i = 1,2,…,21).

2.1.2 *Questionnaires and data quality assurance*

We collect overall semantic transparency (OST) and constituent semantic transparency (CST) data of all test words. Two sets of questionnaires are designed to collect OST data and CST data respectively. Hence each group of test words G_i lead to two sets of data, one on OST and one on CST. After title and instruction, each questionnaire contains three sections. Section 1 is used to collect metadata information including gender, age, education, and location. Section 2 contains four very simple questions about the Chinese language; the first two questions involve open-ended Chinese character identification, the third question is a close-ended homophonic character identification question, and the fourth one is a close-ended antonymous character identification. Each questionnaire contains different questions. Section 3 is the main body of semantic transparency rating task. For a disyllabic nominal compound *AB* consists of constituents *A* and *B*, we use the following question to collect its OST rating scores: "What is the degree of similarity between the sum of the meanings of *A* and *B* and the meaning of *AB*?" (A 和 B 的意思加起來與 AB 的意思在多大程度上相似?) And the following two questions are used to collect CST rating scores of the two constituents: "What is the degree of similarity between *A* when it is used alone and its meaning in *AB*?" (A 單獨使用時的意思和它在 AB 中的意思在多大程度上相似?) and "What is the degree of similarity between *B* when it is used alone and its meaning in *AB*?" (B 單獨使用時的意思和它在 AB 中的意思在多大程度上相似?). Seven-point scales are used in § 3, ranging from 1: "not similar at all" (毫不相似) to 7: "almost the same" (幾乎相同).

In order to control the data quality received in the experiments, we embedded some overlapping test words in the questionnaires. Two compounds are repeated in each and every group: w_1 地步 *dìbù* 'situation' (lit. 'ground-step'), w_2 高山 *gāoshān* 'high mountain' (lit. 'high-mountain'). These two compounds provide inter-group consistency data. In addition, for each group, two compounds appear twice and are called intra-group repeated words. They are used to evaluate intra-group consistency. A dataset is considered to have good quality when it has both good intra-group consistency and good inter-group consistency.

2.1.3 *Quality control in crowdsourcing*

We choose CrowdFlower[1] as our crowdsourcing platform because of its versatility and because the other dominant platform Amazon Mechanical Turk does not have access to participants in China. It is important to note that since the participants remain anonymous and can sometimes be robots (i.e. programs written to fool the

1. CrowdFlower platform has since been updated and become a more powerful platform for machine learning called Figure Eight https://www.figure-eight.com/.

platform), we need additional measures to ensure the validity of the data and to rule out unqualified or untruthful participants. We also need to stop spammers from continuously submitting invalid data at very high speed in order to gain subject money automatically. In order to ensure that the participants are native Chinese speakers and provide good data quality, we implemented the following measures, following Wang et al. (2014); Huang & Wang (2016): (1) Section 2 contains four qualification assurance questions. A participant must correctly answer the first two Chinese character identification questions in addition to one of the last two questions in order to proceed to § 3 to perform rating tasks; (2) each word stimulus in § 3 has an opt-out 'skip' option. Except for questions where the participant opts out and skips explicitly, all the questions in the questionnaires must be answered for the data to be considered valid; (3) a monitor program is installed to detect and terminate spammers automatically; and (4) after the experiment is finished, data are analyzed to filter out data with poor quality, which will be discussed in more details in § 2.2.1.

2.1.4 *Platform and procedure*

CrowdFlower is chosen as our experimental platform as it is shown to be a feasible and effective crowdsourcing platform to collect Chinese language data over other alternatives (Wang et al. 2014, 2015). Each questionnaire is treated as one task on the platform. Since there are 21 groups of words and two questionnaires (for OST and CST respectively) for each group, 42 tasks in total are created T_i^{ost}, T_i^{cst} $(i = 1, 2, \ldots, 21)$. These 42 tasks are published successively with the following parameters: (1) each task will collect 90 responses; (2) 0.15USD paid for each OST questionnaire and 0.25USD for each 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) participants are accepted from the following regions only (controlled by IP addresses): Mainland China, Hong Kong, Macau, Taiwan, Singapore, Malaysia, USA, UK, Canada, Australia, Germany, France, Italy, New Zealand, and Indonesia. In addition, we retain the ability to dynamically disable or enable certain regions on demand in order to ensure both data quality and quantity.

2.2 Results

2.2.1 *Data cleaning and result calculation*

The OST dataset produced by the task T_i^{ost} is $D_i^{ost}(i = 1, 2, \ldots, 21)$. The CST dataset produced by the task T_i^{cst} is $D_i^{cst}(i = 1, 2, \ldots, 21)$. Each dataset contains 90 responses. Because of the nature of crowdsourcing environment, each dataset contains some invalid responses originally that needed to be filtered out. Our filter is based on the following parameters: a response is invalid if (1) its completion time is less than 135 seconds (for OST responses) or less than 250 seconds (for CST responses);[2] or (2) it fails to correctly answer either of the first two questions of § 2; or (3) it fails to correctly answer both last two questions of § 2; or (4) it skipped more than six words in § 3; or (5) it used less than three numbers on the seven-point scales in § 3. We also filtered out the responses from the participants who appeared in more than one country/region according to their IP addresses. The statistics of valid response are shown in Table 1.

The OST dataset D_i^{ost} contains n_i valid responses; it means word *w* in the OST dataset of the *i*th 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.

2.2.2 *Evaluation*

Three evaluation measures are used for data quality: (1) the intra-group consistency of the OST and CST results, (2) the inter-group consistency of the OST and CST results, and (3) the correlation between the OST and CST results.

2.2.2.1 *Intra-group consistency*

In each group G_i we repeat two selected words $w_{i,1}$, $w_{i,2}$ (intra-group repeated words) with enough distance. The difference in ratings of the two repeated appearances of these words is recorded and calculated.

---

2. Each OST questionnaire has about 70 questions, and each CST questionnaire has about 130; in an OST or CST questionnaire, almost all the questions are the same except the stimuli words and can be instantly answered by intuition; note that a participant can take part in as many as 42 tasks; according to our test, if a participant is familiar with the tasks, he/she can answer each question in less than 2 seconds (less than 1 second to identify the stimulus word and another less than 1 second to rate it) without difficulty. Since $70 \times 2 = 140$ seconds, the expected time should be less than this, so we use 135 seconds as the temporal threshold for valid OST responses. The calculation of the temporal threshold for valid CST responses is similar. Since $130 \times 2 = 260$ seconds, the expected time should be less than this, so we use 250 seconds.

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

G_i	OST		CST	
	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

Intra-group consistency of OST scores

There are 21 groups and in each group there are two intra-group repeated words, so there are a total of 42 such words. Each intra-group repeated word appears twice, so we can obtain two OST results r_1, r_2. The difference value between the two results, $d = r_1 - r_2$, of each intra-group repeated word is calculated, so there are 42 difference values. Among them, the maximum value is 0.23; the minimum value is −0.39; the median is 0.02; their mean is −0.01; and their standard deviation is 0.15; most differences are within ±0.2; all of these values are low and indicate that these OST datasets have good intra-group consistency, see Table 2 for details. A paired *t* test was calculated to compare the first and the second OST rating scores (r_1 and

r_2) of the intra-group repeated words. The analysis produced an insignificant t value ($t_{(41)} = -0.32$, $p > 0.05$) which indicates that there is no significant difference between r_1 and r_2.

Table 2. Intra-group consistency of the OST results of each group

G_i	$w_{i,1/2}$	r_1	r_2	d
G_1	野狗 (yěgǒu, wild-dog, 'wild dog')	5.47	5.4	0.07
	關節 (guānjié, pass-knot, 'joint/critical points')	3.56	3.65	−0.09
G_2	火災 (huǒzāi, fire-disaster, 'fire disaster')	5.65	5.78	−0.13
	耳光 (ěrguāng, ear-light, 'a slap on the face')	2.55	2.73	−0.18
G_3	笑臉 (xiàoliǎn, smile-face, 'smiling face')	5.48	5.49	−0.01
	神氣 (shénqì, god-air, 'air (of arrogance)')	3.61	3.64	−0.03
G_4	雜草 (zácǎo, mixed-grass, 'weeds')	5.46	5.49	−0.03
	死黨 (sǐdǎng, dead-party, 'sworn followers')	3.23	3.02	0.21
G_5	毒癮 (dúyǐn, drug-addiction, 'drug addiction')	5.29	5.22	0.07
	水貨 (shuǐhuò, water-goods, 'smuggled goods')	2.94	3.16	−0.22
G_6	手掌 (shǒuzhǎng, hand-palm, 'palm')	5.64	5.56	0.08
	火燒 (huǒshāo, fire-burn, 'burnt/baked wheaten cake')	5.2	5.13	0.07
G_7	低價 (dījià, low-price, 'low price')	5.26	5.31	−0.05
	黑洞 (hēidòng, black-hole, 'black hole')	4.22	4.17	0.05
G_8	涼風 (liángfēng, cool-wind, 'cool breeze')	5.45	5.37	0.08
	風水 (fēngshuǐ, wind-water, 'geomancy')	3.17	3.2	−0.03
G_9	琴聲 (qínshēng, instrument-sound, 'sound of piano etc.')	5.35	5.25	0.1
	手筆 (shǒubǐ, hand-pen, 'literary skill')	3.81	3.79	0.02
G_{10}	白雲 (báiyún, white-cloud, 'white cloud')	5.64	5.71	−0.07
	風土 (fēngtǔ, wind-soil, 'folk customs')	3.26	3.4	−0.14
G_{11}	雨傘 (yǔsǎn, rain-umbrella, 'umbrella')	5.5	5.46	0.04
	背心 (bèixīn, back-heart, 'vest')	3.15	3.13	0.02
G_{12}	燈塔 (dēngtǎ, lamp-tower, 'lighthouse')	5.09	5.29	−0.2
	脾氣 (píqì, spleen-air, 'temperament')	3.51	3.44	0.07
G_{13}	狂風 (kuángfēng, mad-wind, 'gale')	5.37	5.15	0.22
	藍本 (lánběn, blue-book, 'blueprint')	3.13	3.27	−0.14
G_{14}	高樓 (gāolóu, high-building, 'high-rise')	5.54	5.55	−0.01
	口角 (kǒujiǎo, mouth-horn, 'quarrel')	3.34	3.46	−0.12
G_{15}	泥土 (nítǔ, mud-soil, 'soil')	5.57	5.35	0.22
	苦心 (kǔxīn, bitter-heart, 'trouble taken')	3.26	3.65	−0.39
G_{16}	鮮花 (xiānhuā, fresh-flower, 'fresh flower')	5.57	5.55	0.02
	本分 (běnfèn, original-duty, 'obligation')	3.86	4.1	−0.24
G_{17}	店主 (diànzhǔ, shop-host, 'shopkeeper')	5.15	5.38	−0.23
	香火 (xiānghuǒ, fragrant-fire, 'incense/family lineage')	3.62	3.6	0.02

Table 2. (*continued*)

G_i	$w_{i,1/2}$	r_1	r_2	d
G_{18}	桃花 (táohuā, peach-flower, 'peachblossom/romantic liaison')	5.45	5.26	0.19
	色狼 (sèláng, color-wolf, 'sexual predator')	3.42	3.23	0.19
G_{19}	錢包 (qiánbāo, money-bag, 'wallet')	5.42	5.47	−0.05
	火氣 (huǒqì, fire-air, 'temper')	4.11	3.98	0.13
G_{20}	河岸 (héàn, river-bank, 'river bank')	5.34	5.26	0.08
	毛病 (máobìng, hair-illness, 'defect/bad habit')	4.08	3.85	0.23
G_{21}	古城 (gǔchéng, ancient-city, 'ancient city')	5.23	5.12	0.11
	溫床 (wēnchuáng, warm-bed, 'hotbed/catalyst')	3.75	3.98	−0.23
	Min			−0.39
	Max			0.23
	Median			0.02
	Mean			−0.01
	SD			0.15

Intra-group consistency of CST scores

Each intra-group repeated word has two constituents, c_1 and c_2, so each constituent gets two CST results, i.e. $r_{c1,1}$, $r_{c1,2}$ and $r_{c2,1}$, $r_{c2,2}$. We calculate the difference values for the two constituents, $d_1 = r_{c1,1} - r_{c1,2}$ and $d_2 = r_{c2,1} - r_{c2,2}$, and get 42 difference values of the first constituents and 42 difference values of the second constituents. Among the difference values of the first constituents, the maximum value is 0.23; the minimum value is −0.19; the median is 0; their mean is 0, and their standard deviation is 0.11; most differences are within ±0.2; all of these values are low; this indicates that the CST results of the first constituents in the CST datasets of the 21 groups have good intra-group consistency. Among the difference values of the second constituents, the maximum value is 0.28; the minimum value is −0.4; the median is −0.04; their mean is −0.02, and their standard deviation is 0.12; most differences are within ±0.2; all of these values are low; this indicates that the CST results of the second constituents in the CST datasets of the 21 groups have good intra-group consistency, see Table 3 for details. So these 21 CST datasets have good intra-group consistency.

A paired *t* test was calculated to compare the CST rating scores of the first constituents ($r_{c1,1}$ and $r_{c1,2}$) of the intra-group repeated words. The analysis produced an insignificant *t* value ($t_{(41)} = 0.03, p > 0.05$) which indicates that there is no significant difference between $r_{c1,1}$ and $r_{c1,2}$. The same analysis was applied to $r_{c2,1}$ and $r_{c2,2}$ and no significant difference was found either ($t_{(41)} = -1.26, p > 0.05$). Note that the same compounds are used in both studies.

Table 3. Intra-group consistency of the CST results of each group

G_i	$w_{i,1/2}$	c_1			c_2		
		$r_{c1,1}$	$r_{c1,2}$	d_1	$r_{c2,1}$	$r_{c2,2}$	d_2
G_1	野狗	3.94	4.11	−0.17	5.49	5.53	−0.04
	關節	2.86	2.87	−0.01	3.8	3.74	0.06
G_2	火災	5.08	5.27	−0.19	5.09	5.12	−0.03
	耳光	4.17	4.22	−0.05	2.22	2.42	−0.2
G_3	笑臉	5.12	5.12	0	5.34	5.41	−0.07
	神氣	2.86	2.81	0.05	3.14	3.29	−0.15
G_4	雜草	4.71	4.48	0.23	5.64	5.5	0.14
	死黨	2.28	2.4	−0.12	4.36	4.21	0.15
G_5	毒癮	4.88	4.8	0.08	5.19	5.19	0
	水貨	2.08	2.19	−0.11	4.73	4.81	−0.08
G_6	手掌	5.56	5.33	0.23	5.44	5.52	−0.08
	火燒	5.22	5.17	0.05	5.35	5.43	−0.08
G_7	低價	4.75	4.91	−0.16	5.16	5.15	0.01
	黑洞	3.8	3.89	−0.09	4.42	4.56	−0.14
G_8	涼風	5.1	4.92	0.18	5.44	5.42	0.02
	風水	3.15	2.96	0.19	3.17	2.96	0.21
G_9	琴聲	5.04	4.95	0.09	4.91	4.89	0.02
	手筆	3.53	3.62	−0.09	3.64	3.82	−0.18
G_{10}	白雲	4.44	4.47	−0.03	5.37	5.36	0.01
	風土	2.97	2.98	−0.01	3	3	0
G_{11}	雨傘	4.64	4.71	−0.07	5.62	5.71	−0.09
	背心	3.75	3.73	0.02	2.39	2.79	−0.4
G_{12}	燈塔	4.4	4.35	0.05	4.73	4.71	0.02
	脾氣	2.87	2.97	−0.1	3.05	3.11	−0.06
G_{13}	狂風	4.16	4.23	−0.07	5.14	5.26	−0.12
	藍本	2.49	2.68	−0.19	3.51	3.37	0.14
G_{14}	高樓	4.63	4.63	0	5.3	5.39	−0.09
	口角	3.52	3.54	−0.02	2.69	2.78	−0.09
G_{15}	泥土	5.21	5.15	0.06	5.11	5.21	−0.1
	苦心	3.08	3.06	0.02	3.62	3.58	0.04
G_{16}	鮮花	4.32	4.34	−0.02	5.3	5.21	0.09
	本分	3.64	3.5	0.14	3.2	3.25	−0.05
G_{17}	店主	4.62	4.7	−0.08	4.74	4.84	−0.1
	香火	3.88	3.84	0.04	3.68	3.7	−0.02
G_{18}	桃花	4.12	4.2	−0.08	4.78	4.69	0.09
	色狼	3.22	3.22	0	2.57	2.49	0.08
G_{19}	錢包	4.66	4.66	0	4.68	4.6	0.08
	火氣	3.54	3.32	0.22	3.4	3.46	−0.06

Table 3. (*continued*)

G_i	$w_{i,1/2}$	c_1			c_2		
		$r_{c1,1}$	$r_{c1,2}$	d_1	$r_{c2,1}$	$r_{c2,2}$	d_2
G_{20}	河岸	5	4.94	0.06	5.06	5.12	−0.06
	毛病	2.82	2.82	0	4.73	4.45	0.28
G_{21}	古城	4.69	4.55	0.14	5	5.1	−0.1
	溫床	3.69	3.86	−0.17	3.63	3.65	−0.02
	Min			−0.19			−0.4
	Max			0.23			0.28
	Median			0			−0.04
	Mean			0			−0.02
	SD			0.11			0.12

2.2.2.2 *Inter-group consistency*

We inserted two inter-group repeated words, w_1 地步 and w_2 高山, into all of these 21 groups in order to evaluate the inter-group consistency by comparing their semantic transparency rating results in different groups. Since w_1, w_2 appear in all OST and CST questionnaires of 21 groups, we can obtain (1) 21 OST results of w_1, (2) 21 OST results of w_2, (3) 21 CST results of each of the two constituents $w_{1,c1}$, $w_{1,c2}$ of w_1, and (4) 21 CST results of each of the two constituents $w_{2,c1}$, $w_{2,c2}$ of w_2. Standard deviation can be used to measure difference, for example, the standard deviation of the 21 OST results of w_1 is 0.21; this value is small and indicates high consistency; because these 21 results are from the OST datasets of 21 groups respectively, so we can say that these 21 OST datasets have good inter-group consistency. The standard deviation of the 21 OST results of w_2 is 0.12; the standard deviation of 21 CST results of the first constituent of w_1 is 0.21, and that of the second is 0.18; the standard deviation of 21 CST results of the first constituent of w_2 is 0.14, and that of the second is 0.16; all of these values are small and all of them indicate good inter-group consistency (see Table 4).

Table 4. Inter-group consistency of the OST and CST results

G_i	OST			CST		
	w_1	w_2	$w_{1,c1}$	$w_{1,c2}$	$w_{2,c1}$	$w_{2,c2}$
G_1	2.82	5.66	2.8	2.94	4.66	5.63
G_2	3.58	5.55	3.11	3.17	4.77	5.67
G_3	3.54	5.59	3.21	3.22	4.72	5.62
G_4	3.86	5.77	3.64	3.59	4.64	5.4
G_5	3.71	5.53	3.42	3.58	4.71	5.56
G_6	3.69	5.65	3.69	3.61	4.89	5.72

(*continued*)

Table 4. (*continued*)

G_i	OST			CST		
	w_1	w_2	$w_{1,c1}$	$w_{1,c2}$	$w_{2,c1}$	$w_{2,c2}$
G_7	3.61	5.65	3.44	3.55	4.75	5.45
G_8	3.32	5.6	3.33	3.25	4.81	5.52
G_9	3.44	5.33	3.35	3.4	4.73	5.58
G_{10}	3.6	5.57	3.32	3.25	4.58	5.36
G_{11}	3.73	5.79	3.52	3.45	4.75	5.52
G_{12}	3.4	5.71	3.19	3.19	4.29	5.27
G_{13}	3.44	5.62	3.44	3.42	4.75	5.49
G_{14}	3.61	5.68	3.07	3.02	4.7	5.56
G_{15}	3.5	5.44	3.57	3.42	4.83	5.25
G_{16}	3.6	5.64	3.45	3.34	4.71	5.38
G_{17}	3.5	5.56	3.36	3.34	4.68	5.34
G_{18}	3.6	5.77	3.18	3.12	4.47	5.08
G_{19}	3.58	5.75	3.24	3.26	4.56	5.38
G_{20}	3.66	5.55	3.22	3.33	4.84	5.45
G_{21}	3.46	5.42	3.22	3.18	4.65	5.43
Min	2.82	5.33	2.8	2.94	4.29	5.08
Max	3.86	5.79	3.69	3.61	4.89	5.72
Median	3.58	5.62	3.33	3.33	4.71	5.45
Mean	3.54	5.61	3.32	3.32	4.69	5.46
SD	0.21	0.12	0.21	0.18	0.14	0.16

2.2.2.3 *Correlation between OST and CST results*

Each compound in the datasets has two constituents; both constituents affect the OST of the compound, but neither of them can solely determine the OST of the compound. So the mean of the two CST values of a compound is a fairly good estimation of its OST value. Therefore, if the datasets are reliable, in each group, we should observe strong correlation between the OST results and their corresponding means of the CST results. For each group, we calculate three Pearson correlation coefficients (r); r_1 is the r between the OST results and their corresponding CST results of the first constituents; r_2 is the r between the OST results and their corresponding CST results of the second constituents; and r_3 is the r between the OST results and their corresponding means of the CST results. The r_3 values of the 21 groups are all greater than 0.9 which indicates very strong correlation; the r_1 and r_2 values are also reasonably high, see Table 5 for details. After merging and normalization (see § 2.2.3), we calculated these three correlation coefficients on the merged datasets, the results are $r_1 = 0.7$, $r_2 = 0.69$, $r_3 = 0.88$ (see Figure 1). These results support the reliability of these datasets.

Table 5. Correlation coefficients between the OST and CST results

G_i	r_1	r_2	r_3
G_1	0.71	0.74	0.93
G_2	0.73	0.70	0.93
G_3	0.76	0.78	0.96
G_4	0.79	0.77	0.96
G_5	0.79	0.57	0.95
G_6	0.65	0.74	0.91
G_7	0.83	0.77	0.94
G_8	0.75	0.78	0.95
G_9	0.72	0.81	0.96
G_{10}	0.84	0.83	0.96
G_{11}	0.82	0.73	0.94
G_{12}	0.70	0.78	0.94
G_{13}	0.86	0.85	0.96
G_{14}	0.71	0.85	0.96
G_{15}	0.71	0.80	0.96
G_{16}	0.82	0.83	0.95
G_{17}	0.80	0.85	0.95
G_{18}	0.80	0.85	0.95
G_{19}	0.76	0.80	0.95
G_{20}	0.76	0.76	0.94
G_{21}	0.75	0.85	0.96
Min	0.65	0.57	0.91
Max	0.86	0.85	0.96
Median	0.76	0.78	0.95
Mean	0.77	0.78	0.95
SD	0.06	0.07	0.01

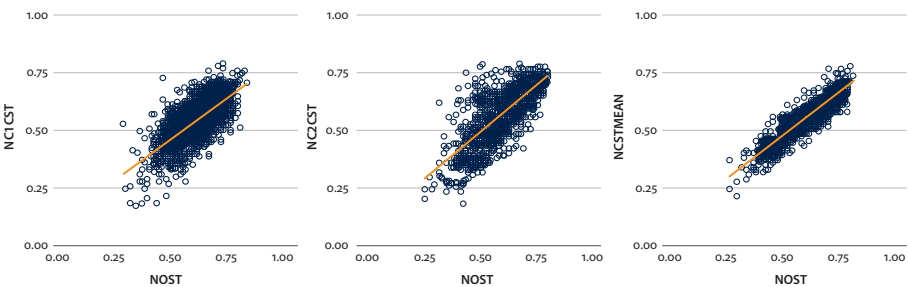

Figure 1. Scatter plot: Normalized OST results against normalized C1CST/C2CST results and the mean of normalized C1CST and C2CST in the merged dataset

2.2.3 *Merging and normalization*

The evaluation results show that the collected data are generally reliable and have relatively high intra-group and inter-group consistency which further indicate that these datasets share similar scale and are basically comparable, so we can merge the 21 OST datasets into one big OST dataset D_{ost} and merge the 21 CST datasets into one big CST dataset D_{cst}. When we merged these datasets, we deleted all the extra words which were used to evaluate the inter-group consistency; for the repeated words which are used to evaluate the intra-group consistency, the final result of each of them is the mean of its two results. According to our definition, the range of semantic transparency value is [0, 1], but the experimental results are obtained using seven-point scales, so we need to rescale these results in order to map them to the range [0, 1]. The normalized OST and CST results will be merged into D_{ost} and D_{cst} respectively. Assume that, in the dataset D_{ost}, the OST result of the *ith*(i = 1,2,...,1176) word is S_i^w, and the normalized result is S'^w_i, then,

$$S'^w_i = \frac{S_i^w - 1}{6}$$

Assuming that, in the dataset D_{cst}, the CST result of the jth (j = 1, 2) constituent of the ith word is $S'^c_{i,j}$, and the normalized result is $S'^c_{i,j}$, then,

$$S'^c_{i,j} = \frac{S^c_{i,j} - 1}{6}$$

2.2.4 *Distribution and classification*

The OST and CST results cannot cover the whole range of the scale [0, 1]; both ends shrink towards the center, and the shrinkage of each end is about 0.2; nevertheless, the results can still assign proper ranks of semantic transparency to the compounds and their constituents which are generally consistent with our intuitions. Among the normalized OST results, the maximum is 0.82; the minimum is 0.26; the median is 0.64; and their mean is 0.62 (SD = 0.1). Among the normalized CST results of the first constituents (C1CST results), the maximum is 0.79; the minimum is 0.18; the median is 0.57; and their mean is 0.56 (SD = 0.1). And among the normalized CST results of the second constituents (C2CST results), the maximum is 0.8; the minimum is 0.18; the median is 0.59; and their mean is 0.58 (SD = 0.11). The distributions of OST, C1CST, and C2CST results are similar; all of them are slightly negatively skewed (see Figure 2). These distributions exhibit that more compounds and their constituents in our datasets have relatively high semantic transparency values.

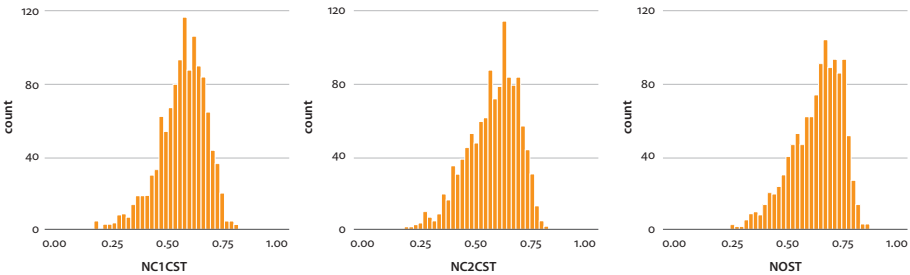

Figure 2. Distributions of normalized OST and CST results

When analyzing the semantic transparency of compounds, scholars usually classify compounds into three categories: transparent compounds, semi-transparent compounds, and opaque compounds (Zwitserlood 1994; Libben 1998; Libben et al. 2003; Mok 2009; Han et al. 2014). Since the range of semantic transparency is [0, 1], we can technically divide this interval into three equal segments, each corresponds with one category. As mentioned above, the semantic transparency results cannot cover the whole range, both ends shrink towards the center, and the actual range is [0.26, 0.82]. We choose to divide this range into three equal segments, for it actually covers the words from fully transparent to fully opaque according to our observation; and the length of each segment is about 0.18. In this way, we classify the compounds in the dataset into three categories: (1) transparent compounds (range = [0.63, 0.82], n = 627, 53.32%), (2) semi-transparent compounds (range = [0.44, 0.62], n = 472, 40.14%), and (3) opaque compounds (range = [0.26, 0.43], n = 77, 6.55%).

2.3 Correlations with laboratory experimental results

We can further evaluate the semantic transparency rating results from the crowdsourcing-based experiment by examining to what extent they correlate with the results from laboratory-based experiment. A sample of 152 compounds was rated in a laboratory experiment, as reported in Wang et al. (2015). We calculated three Pearson 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: 0.95, and (3) the correlation coefficient between the normalized CST results of the second morphemes of the compounds from the two experiments: 0.94. All of the correlation coefficients are greater than 0.9 which indicates that the results from the crowdsourcing-based experiment correlate strongly with the results from the laboratory-based experiment (see Figure 3) further confirming the reliability of the crowdsourced dataset.

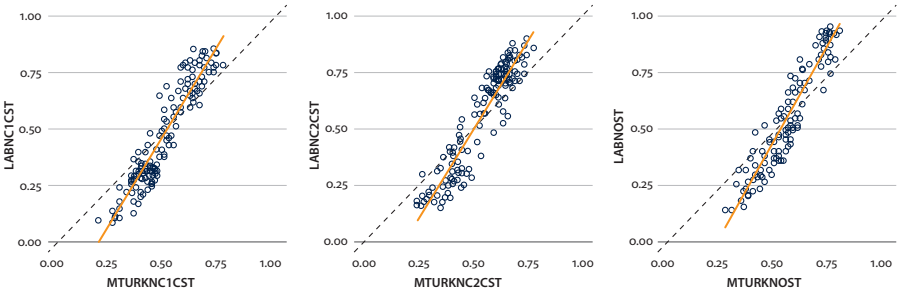

Figure 3. Correlations between normalized OST and CST results from the crowdsourcing-based and the laboratory-based experiments

3. Testing morphological structure effect on semantic transparency ratings

3.1 Method

A standard way to establish causal relation between two variables is to experiment directly on their correlations. For instance, we could take morphological structure as the independent variable, and the CST rating scores as the dependent variable, and all other variables must be kept constant at the same time. A causal relation can be established if we observed that there was a relation between morphological structure and CST rating scores. However, this classical method is untenable because it is not possible to keep the semantic contribution of the two constituents in a compound constant. However, with the OST and CST dataset, we now have an alternative method to establish the correlation between morphological structure and CST. That is, instead of directly measure the correlation between morphological structure and CST, we can now test the predictions of CST based on different morphological structures.

That is, a set of linguistic predictions can be made based on the hypothesis that there is morphological structure effect on the semantic transparency of the compounds and the semantic transparency dataset can be used to test these predictions. The morphological structure effect hypothesis will be supported if and only if these predictions are borne out.

By definition, the meanings of both constituents are retained in the meaning of dimorphemic transparent compounds. That is, regardless of the morphological structure, there should be high correlation between CST scores and between OST and CST scores. However, if the human subjects are aware of the morphological structure and this awareness could have effect on their semantic transparency rating behavior, then the following predictions could be made:

Prediction 1: For transparent compounds with headed morphological structures, such as modifier-head structure, the rating scores of the CST of their constituents will show significant difference.
Prediction 2: For transparent compounds with non-headed or double-headed morphological structures, such as the coordinative structure, there is no significant difference between the rating scores of the CST of their constituents.

In our semantic transparency dataset, the transparent modifier-head compounds can be used to test the first prediction and coordinative compounds can be used to test the second prediction. If these two predictions are borne out by the data, we can show that the position of the constituent has an effect on transparency rating and that the effect is determined by morphological structure instead of constituent order (as constituent order will predict that both set of data will have parallel effects). Of course the prediction verification study is not sufficient to prove direct causal relation between morphological structure and CST rating behavior. However, given constraints of the nature of the data, as discussed above, direct experiment controlling all other variables is not possible. Hence we will further strengthen our argument by additional supporting data in the form of a resampling simulation analysis.

3.2 Results

3.2.1 *Testing prediction 1*

There are 1,053 modifier-head compounds and 563 (53.5%) of them are transparent according to the criterion discussed in § 2.2.4. We then calculated the CST differences between the first and the second constituents (C1 and C2) of the transparent modifier-head compounds (*CST.DIFF* = *C2CST*−*C1CST*); the distribution of these differences is shown in Figure 4. Among these differences, 65% (n = 366)

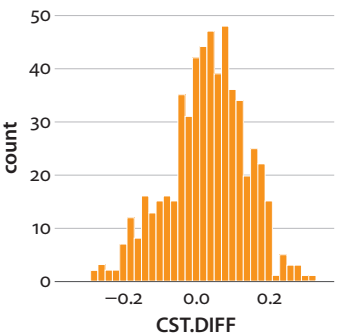

Figure 4. Distribution of constituent semantic transparency differences of transparent modifier-head compounds

are positive and only 35% (n = 197) are negative or zero. The tendency that the CST rating results of the heads are higher than those of the modifiers is very obvious. But is this tendency statistically significant? We conducted a two sample t test, $t_{(1096.6)} = -7.2$, $p < 0.01$ (one-tailed), which confirmed that the mean of constituent semantic rating results of the second constituents is significantly greater than that of the first constituents.

3.2.2 *Testing prediction 2*

There are 107 coordinative compounds in the dataset and 65 (60.75%) of them are transparent. We then calculated the CST differences between the first and the second constituents of these compounds, the distribution is shown in Figure 5. 50.8% (n = 33) of them is positive, and 49.2% (n = 32) is negative or zero. At first glance, compared with the modifier-head compounds, the differences are relatively small. The mean of the CST rating scores of the second constituents is 0.64, and that of the first constituents is 0.637, and the difference between them is 0.003. A two sample t test was conducted, $t_{(127.98)} = -0.3$, $p > 0.05$ (one-tailed), and found no significant difference.

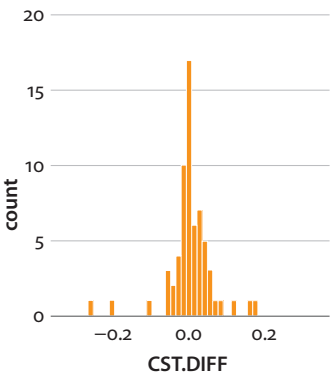

Figure 5. Distribution of constituent semantic transparency differences of transparent coordinative compounds

3.2.3 *Resampling simulation analysis*

Statistical tests, such as t test, always have the probability to yield false positive results. It is possible that our findings above are just coincidences. To eliminate such possibility, we use the technique of resampling simulation. The 563 transparent modifier-head compounds used in Test 1 and the 65 transparent coordinative compounds used in Test 2 will continue to be used in this analysis. We call the former the modifier-head set and the later the coordinative set. The procedure of this analysis adapted from the design of Sprouse (2011) runs as follows:

1. Choose a set *s* from the modifier-head set and the coordinative set.
2. Choose a sample size *m* (for example, 10).
3. Conduct a random sampling without replacement on set *s* to produce a sample of compounds of size *m*.
4. Run a two sample *t* test on the sample, H_1:*mean*(*C2CST*) ≠ *mean*(*C1CST*).
5. Repeat Steps 3 to 4 a total of 10,000 times.
6. Calculate detectability: the percentage of significant results ($p < 0.05$) out of those 10,000 tests.
7. Calculate directions: the percentages of directions of significant results (*C1CST*<*C2CST* or *C1CST*>*C2CST*).
8. Repeat Steps 2 to 5 for other sample sizes.
9. Repeat Steps 2 to 8 for the other set.

The results in Table 6 show that when sample size is small the chance to detect significant difference in CST rating scores between head and modifier in modifier-head compounds is quite low. For example when sample size is 10, the chance is only 16.95%. But as sample size increases, the detectability becomes greater and greater, when sample size reaches 140, the chance is already above 96% and when sample size reaches 180 or so, the chance becomes very close to 100%. And the difference in CST rating scores has a clear direction, as shown in the Direction columns, head tends to get a higher CST rating score than modifier.

In contrast, we basically cannot detect such a difference (detectability < 2%) in coordinative compounds. It also shows opposite trend from modifier-head compounds. When the sample size is small we can detect significant differences very occasionally, but as sample size increases less and less differences can be detected and the detectability has already reduced to zero when the sample size reaches 30. No matter the sample size is small or large, the mean of C2CST basically equals that of C1CST. However, the lack of difference does not necessarily mean that there is no morphological structure effect in coordinative compounds, since no difference can also be an effect. Since coordination is a non-headed (or double-headed) structure, a headedness effect would predict that no difference in their CSTs should be found.

The resampling results show convincingly that our findings in previous two tests are not coincidences. This analysis also excludes to some extent the possibility that the observed effect of morphological structure on semantic transparency ratings is caused by variables other than morphological structure (such as character frequency, family size of morpheme, word frequency, number of strokes, orthographic neighborhood, etc., just to name a few). Since random resampling was employed, effect of these other variables would be randomized and be presented in many different combinations without a constant pattern and cannot show consistent effects. The fact that headedness effect shows reliably when sample size

is large enough strongly support the correlation. It is worth noting that, of course, our study would not be able to differentiate the unlikely scenario where the original sample contains systematic biases cloaked as headedness.

Table 6. Detectability and direction statistics of various sample sizes for the modifier-head set and the coordinative set

Sample size	Modifier-head compound			Coordinative compound		
	Detect.	Direction		Detect.	Direction	
		C1 < C2	*C1 > C2*		*C1 < C2*	*C1 > C2*
10	16.95	98.47	1.53	0.25	96	4
20	27.89	99.89	0.11	0.11	100	0
30	37.81	99.97	0.03	0	N/A	N/A
40	48.25	99.98	0.02	0	N/A	N/A
50	56.9	100	0	0	N/A	N/A
60	63.77	100	0	0	N/A	N/A
70	71.24	100	0	N/A	N/A	N/A
80	77.03	100	0	N/A	N/A	N/A
90	81.39	100	0	N/A	N/A	N/A
100	85.5	100	0	N/A	N/A	N/A
110	89.43	100	0	N/A	N/A	N/A
120	92.26	100	0	N/A	N/A	N/A
130	93.97	100	0	N/A	N/A	N/A
140	96.25	100	0	N/A	N/A	N/A
150	97.27	100	0	N/A	N/A	N/A
160	98.14	100	0	N/A	N/A	N/A
170	98.48	100	0	N/A	N/A	N/A
180	99.09	100	0	N/A	N/A	N/A
190	99.53	100	0	N/A	N/A	N/A
200	99.63	100	0	N/A	N/A	N/A
210	99.82	100	0	N/A	N/A	N/A
220	99.89	100	0	N/A	N/A	N/A
230	99.92	100	0	N/A	N/A	N/A
240	99.97	100	0	N/A	N/A	N/A
250	100	100	0	N/A	N/A	N/A
260	100	100	0	N/A	N/A	N/A
270	100	100	0	N/A	N/A	N/A
280	100	100	0	N/A	N/A	N/A
290	100	100	0	N/A	N/A	N/A
300	100	100	0	N/A	N/A	N/A

4. General discussion and conclusion

We advocate in this paper that crowdsourcing can be a highly instrumental method to collect linguistic judgments and to construct language resources in Chinese language studies. In addition, the proposed methodology of comparing constituent transparency and word transparency sheds light on the relation between morpho-lexical structure and cognitive processing of lexical meanings. In particular, we created a semantic transparency dataset of Chinese compounds consisting of the overall and constituent semantic transparency rating scores (OST and CST) of 1,176 Chinese disyllabic nominal compounds. The data collected have good intra-group and inter-group consistency, the OST and CST data highly correlate with each other, and the results are consistent with our intuitions. We compared the rating data obtained from crowdsourcing and laboratory experiments and observed very strong correlation for both OST and CST rating data. This shows the two experiments yield comparable data and crowdsourcing experiments are a feasible and effective alternative to the laboratory experiment in linguistic studies especially when large sampling size is required. The crowdsourced and laboratory semantic transparency datasets can be found in Appendixes E and F of Wang (2016) respectively. And the full dataset will be available from Linguistic Data Consortium (LDC) of University of Pennsylvania as Wang et al. (2019).

The semantic transparency dataset was then used to explore the research question of whether morphological structure, headedness in particular, affects semantic transparency ratings. We found that in transparent modifier-head compounds, the two constituents tend to get significantly different CST rating scores biased towards the head. While in transparent coordinative compounds, the two constituents tend to get equal CST rating scores. The directional morphological structure effect on semantic transparency ratings for modifier-head compounds and the lack of such effects in coordinative compounds can be predicted with a headedness account. That is, subject favors heads in compounds and rated it higher in transparent conditions. Resampling simulation analysis shows that these findings are not likely to be coincidences by using random samples of various sample sizes and it also excludes to some extent the possibility that the observed effect is caused by incidental non-controlled variables. What our study could not show, however, is whether the effect resulted from the concept of morpho-lexical head or from the identity of the grammatical category of a compound and a specific constituent. This question will have to be answered by future studies.

It is observed that our proposed hypothesis of headedness account of semantic transparency is supported but cannot be considered conclusive due to restriction of sample size and impossibility to control all variables (such as variation of categorical matches). The impossibility of doing representative factorial experiments

in lexical research motivates the increasing use of megastudies (Keuleers & Balota 2015; Balota et al. 2012). A megastudy perhaps is a more appropriate for our research question. We only investigated the first two most productive structures in Chinese compounding (modifier-head structure and coordinative structure), future work should expand to cover other structures. Our study does show the new possibilities offered by crowdsourced experiments, which has been adopted by additional studies on word intuition (Wang et al. 2017) and on transparency of Chinese character components (Yang et al. 2018). We expect crowdsourcing linguistic experiment for Chinese will continue to develop and open up possibilities to tackle a wider range of issues.

Acknowledgements

We would like to thank the anonymous 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).

Abbreviations

A	adjective
AMT	Amazon's Mechanical Turk
AN	adjective-noun
CST	constituent semantic transparency
C1	constituent 1
C1CST	constituent semantic transparency of constituent 1
C2	constituent 2
C2CST	constituent semantic transparency of constituent 2
IP	Internet Protocol
LDC	Linguistic Data Consortium
N	noun
NN	noun-noun
O	opaque
OO	opaque-opaque
OST	overall semantic transparency
OT	opaque-transparent
T	transparent
TO	transparent-opaque
TT	transparent-transparent
V	verb
VN	verb-noun

References

Allen, Margaret R. 1979. Morphological investigations. Storrs: University of Connecticut. (Doctoral dissertation.)

Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge: The MIT Press.

Balota, David A. & Yap, Melvin J. & Hutchison, Keith A. & Cortese, Michael J. 2012. Megastudies: What do millions (or so) of trials tell us about lexical processing? In Adelman, James S. (ed.), *Visual word recognition volume 1: Models and methods, orthography and phonology*, 90–115. London: Psychology Press.

Chen, Keh-Jiann & Huang, Chu-Ren & Chang, Li-Ping & Hsu, Hui-Li. 1996. Sinica corpus: Design methodology for balanced corpora. In Park, Byung-Soo & Kim, Jong-Bok (eds.), *Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation*, 167–176. Seoul: Kyung Hee University.

Dictionary Editing Section Institude of Linguistics Chinese Academy of Social Sciences (ed.). 2012. *Xiandai Hanyu cidian* [The Contemporary Chinese dictionary]. 6th edn. Beijing: The Commercial Press.

Feldman, Laurie Beth & Pastizzo, Matthew John. 2003. Morphological facilitation: The role of semantic transparency and family size. In Baayen, R. Harald & Schreuder, Robert (eds.), *Morphological structure in language processing*, 233–258. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110910186.233

Frisson, Steven & Niswander-Klement, Elizabeth & Pollatsek, Alexander. 2008. The role of semantic transparency in the processing of English compound words. *British Journal of Psychology* 99(1). 87–107. https://doi.org/10.1348/000712607X181304

Gan, Hongmei. 2008. Yuyi toumingdu dui zhongji Hanyu yuedu zhong cihui xuexi de yingxiang [The effect of semantic transparency on vocabulary learning in intermediate Chinese reading]. *Yuyan Wenzi Yingyong* [Applied Linguistics] 2008(1). 82–90.

Gao, Bing & Gao, Fengqiang. 2005. Hanyu zici shibie zhong cipin he yuyi toumingdu de jiaohu zuoyong [The interaction between word frequency and semantic transparency in the recognition of Chinese words]. *Xinli Kexue* [Journal of Psychological Science] 28(6). 1358–1360.

Han, Yi-Jhong & Huang, Shuo-Chieh & Lee, Chia-Ying & Kuo, Wen-Jui & Cheng, Shih-Kuen. 2014. The modulation of semantic transparency on the recognition memory for two-character Chinese words. *Memory & Cognition* 42(8). 1315–1324. https://doi.org/10.3758/s13421-014-0430-1

Huang, Chu-Ren & Wang, Shichang. 2016. Zhongbao celüe zai yuyan ziyuan jianshe zhong de yingyong [The application of crowdsourcing strategy in utilizing language resources]. *Yuyan Zhanlüe Yanjiu* [Chinese Journal of Language Policy and Planning] 1(6). 36–46.

Huang, Shuanfan. 1998. Chinese as a headless language in compounding morphology. In Packard, Jerome L. (ed.), *New approaches to Chinese word formation: morphology, phonology and the lexicon in Modern and Ancient Chinese*, 261–284. Berlin: Walter de Gruyter. https://doi.org/10.1515/9783110809084.261

Keuleers, Emmanuel & Balota, David A. 2015. Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *Quarterly Journal of Experimental Psychology* 68(8). 1457–1468. https://doi.org/10.1080/17470218.2015.1051065

Levin, Beth & Hovav, Malka R. 2001. Morphology and lexical semantics. In Spencer, Andrew & Zwicky, Arnold M. (eds.), *The handbook of morphology*, 248–271. Hoboken: JohnWiley & Sons.

Lexicon of Common Words in Contemporary Chinese Project Group (ed.). 2008. *Xiandai Hanyu changyongcibiao* [Lexicon of common words in Contemporary Chinese]. Beijing: The Commercial Press.

Li, Jinxia. 2011. Xiandai Hanyu cidian de ciyi toumingdu kaocha [A quantification analysis of the transparency of lexical meaning of Modern Chinese dictionary]. *Hanyu Xuebao* [Chinese Linguistics] 2011(3). 54–62.

Li, Jinxia & Li, Yuming. 2008. Lun ciyi de toumingdu [On the transparency of lexical meaning]. *Yuyan Yanjiu* [Studies in Language and Linguistics] 28(3). 60–65.

Libben, Gary. 1998. Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. *Brain and Language* 61(1). 30–44. https://doi.org/10.1006/brln.1997.1876

Libben, Gary & Gibson, Martha & Yoon, Yeo B. & Sandra, Dominiek. 2003. Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language* 84(1). 50–64. https://www.sciencedirect.com/science/article/pii/S0093934X02005205?via%3Dihub

Marslen-Wilson, William & Tyler, Lorraine K. & Waksler, Rachelle & Older, Lianne. 1994. Morphology and meaning in the English mental lexicon. *Psychological Review* 101(1). 3–33. https://doi.org/10.1037/0033-295X.101.1.3

Mok, Leh Woon. 2009. Word-superiority effect as a function of semantic transparency of Chinese bimorphemic compound words. *Language and Cognitive Processes* 24(7–8). 1039–1081. https://doi.org/10.1080/01690960902831195

Myers, James & Derwing, Bruce & Libben, Gary. 2004a. The effect of priming direction on reading Chinese compounds. *Mental Lexicon Working Papers* 1(1). 69–86.

Myers, James & Libben, Gary & Derwing, Bruce. 2004b. The nature of transparency effects in Chinese compound processing. (Poster presented at the Fourth International Conference on the Mental Lexicon, Windsor, 30 June-3 July, 2004.

Plag, Ingo. 2003. *Word-formation in English*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511841323

Pollatsek, Alexander & Hyönä, Jukka. 2005. The role of semantic transparency in the processing of Finnish compound words. *Language and Cognitive Processes* 20(1-2). 261–290.

Reddy, Siva & McCarthy, Diana & Manandhar, Suresh. 2011. An empirical study on compositionality in compound nouns. In Wang, Haifeng & Yarowsky, David (eds.), *Proceedings of 5th International Joint Conference on Natural Language Processing*, 210–218. Chiang Mai: Asian Federation of Natural Language Processing.

Ren, Min. 2012. Yingxiang Xiandai Hanyu shuangyin fuheci yuyi toumingdu de jizhi yanjiu [Semantic transparency of the Modern Chinese compounds]. *Heibei Shifan Daxue Xuebao (Zhexue Shehui Kexue Ban)* [Journal of Hebei Normal University (Philosophy and Social Sciences Edition)] 2012(4). 85–91.

Schnoebelen, Tyler & Kuperman, Victor. 2010. Using Amazon Mechanical Turk for linguistic research. *Psihologija* 43(4). 441–464. https://doi.org/10.2298/PSI1004441S

Schreuder, Robert & Baayen, R. Harald. 1995. Modeling morphological processing. In Feldman, Laurie B. (ed.), *Morphological aspects of language processing*, 131–154. Hillsdale: Lawrence Erlbaum Associates.

Song, Xuan. 2013. Hanyu fuheci yuyi toumingdu de shiyi moshi fenxi [A paraphrastic analysis of the semantic transparency of Chinese compounds]. *Yunnan Shifan Daxue Xuebao (Duiwai Hanyu Jiaoxue Yu Yanjiu Ban)* [Journal of Yunnan Normal University (Teaching and Research on Chinese As a Foreign Language)] 11(3). 48–52.

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. https://doi.org/10.3758/s13428-010-0039-7

Tsai, Chih-Hao. 1996. Effects of semantic transparency and morphological structure on the representation and recognition of Chinese disyllabic words. In Cheng, Tsai-Fa & Li, Yafei & Zhang, Hongming (eds.), *Proceedings of the Joint Meeting of the Fourth International Conference on Chinese Linguistics and the Seventh North American Conference on Chinese Linguistics*, vol. 2, 326–343. Los Angeles: University of Southern California.

Wang, Chunmao & Peng, Danling. 1999. Hechengci jiagong zhong de cipin cisu pinlü ji yuyi toumingdu [The roles of surface frequencies, cumulative morpheme frequencies, and semantic transparencies in the processing of compound words]. *Xinli Xuebao* [Acta Psychologica Sinica] 31(3). 266–273.

Wang, Chunmao & Peng, Danling. 2000. Chongfu qidong zuoye zhong ci de yuyi toumingdu de zuoyong [The role of semantic transparencies in the processing of compound words]. *Xinli Xuebao* [Acta Psychologica Sinica] 32(2). 127–132.

Wang, Shichang. 2016. *Crowdsourcing method in empirical linguistic research: Chinese studies using Mechanical Turk-based experimentation*. Hong Kong: The Hong Kong Polytechnic University. (Doctoral dissertation.)

Wang, Shichang & Huang, Chu-Ren & Yao, Yao & Chan, Angel. 2014. Building a semantic transparency dataset of Chinese nominal compounds: A practice of crowdsourcing methodology. In Baptista, Jorge & Bhattacharyya, Pushpak & Fellbaum, Christiane & Forcada, Mikel & Huang, Chu-Ren & Koeva, Svetla & Krstev, Cvetana & Laporte Eric (eds.), *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014 at COLING-2014)*, 147–156. Dublin: Dublin City University.

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

Wang, Shichang & Huang, Chu-Ren & Yao, Yao & Chan, Angel. 2019 (to appear). *SemTransCNC 1.0: Semantic transparency of Chinese nominal compounds 1.0*. Philadelphia: Language Data Consortium, University of Pennsylvania.

Xu, Caihua & Li, Tang. 2001. Yuyi toumingdu yingxiang ertong cihui xuexi de shiyan yanjiu [The role of semantic transparency on word recognition and reading comprehension: An experimental study on children]. *Yuyan Wenzi Yingyong* [Applied Linguistics] 2001(1). 53–59.

Yuan, Chunfa & Huang, Changning. 1998. Jiyu yusu shujuku de Hanyu yusu ji gouci yanjiu [Studies on Chinese morpheme and word formation based on morpheme database]. *Yuyan Wenzi Yingyong* [Applied Linguistics] 1998(3). 83–88.

Zwitserlood, Pienie. 1994. The role of semantic transparency in the processing and representation of Dutch compounds. *Language and Cognitive Processes* 9(3). 341–368. https://doi.org/10.1080/01690969408402123

Authors' addresses

Shichang Wang (corresponding author)
School of Literature
Shandong University
27, Shanda Nan Road
Jinan, Shandong 250100
China
wangshichang@sdu.edu.cn

Publication history

Date received: 4 December 2017
Date accepted: 16 October 2018