Create a Manual Chinese Word Segmentation Dataset Using Crowdsourcing Method

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

Abstract

The manual Chinese word segmentation dataset *WordSegCHC 1.0* which was built by eight crowdsourcing tasks conducted on the Crowdflower platform contains the manual word segmentation data of 152 Chinese sentences whose length ranges from 20 to 46 characters without punctuations. All the sentences received 200 segmentation responses in their corresponding crowdsourcing tasks and the numbers of valid response of them range from 123 to 143 (each sentence was segmented by more than 120 subjects). We also proposed an evaluation method called *manual segmentation error rate* ($MSER$) to evaluate the dataset; the $MSER$ of the dataset is proved to be very low which indicates reliable data quality. In this work, we applied the crowdsourcing method to Chinese word segmentation task and the results confirmed again that the crowdsourcing method is a promising tool for linguistic data collection; the framework of crowdsourcing linguistic data collection used in this work can be reused in similar tasks; the resultant dataset filled a gap in Chinese language resources to the best of our knowledge, and it has potential applications in the research of word intuition of Chinese speakers and Chinese language processing.

1 Introduction

Chinese word segmentation which can be conducted by human or computer in the form of written or oral, is a hot topic receiving great interest from several branches of linguistics especially from theoretical, computational and psychological linguistics, simply because it relates to or perhaps is the key to several critical theoretical and applicational issues, for example word definition, word intuition and Chinese language processing.

However in the traditional laboratory setting, limited by budget and/or the difficulty of large scale subject recruitment, etc., it is very difficult or even impossible to build large manual Chinese word segmentation dataset (the defining feature of this kind of dataset is that each sentence must be segmented by a large group of people in order to measure word intuition of Chinese speakers) and this hinders the availability of such language resource. Fortunately, the crowdsourcing method perhaps can help us to solve this problem. Being aware of this background, the crowdsourced manual Chinese word segmentation dataset *WordSegCHC 1.0* was built with multiple purposes in our mind.

The first purpose is to further explore the application of crowdsourcing method in language resource building and linguistic studies in the context of the Chinese language. Crowdsourcing method is a promising tool to solve the linguistic data bottleneck problem which widely happens in various linguistic studies; it is efficient and economic and can help us realize much higher randomness and much larger scale in sampling; in annotation tasks we can also get much higher redundancy to help us make decisions on ambiguous cases with more confidence; although its signal-to-noise ratio (SNR) is usually lower than the traditional laboratory method, it can yield high quality data as good as or even better than the traditional method when combined with several data quality control measures including parameter optimization, screening questions, performance monitoring, data valida-

Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing (SIGHAN-8), pages 7–14,
Beijing, China, July 30-31, 2015.

tion, data cleansing, majority voting, peer review, spammer monitor, etc (Crump et al., 2013; Allahbakhsh et al., 2013; Mason and Suri, 2012; Behrend et al., 2011; Buhrmester et al., 2011; Callison-Burch and Dredze, 2010; Paolacci et al., 2010; Ipeirotis et al., 2010; Munro et al., 2010; Snow et al., 2008).

We have already successfully applied crowdsourcing method to the semantic transparency of compound rating task and built a semantic transparency dataset which contains the semantic transparency rating data of about 1,200 disyllabic Chinese nominal compounds (Wang et al., 2014a); we want to further extend the application of crowdsourcing method to Chinese word segmentation task to further evaluate the crowdsourcing method and to build new language resource.

The second purpose is to support the studies on word intuition of Chinese speakers in general and to examine the effect of semantic transparency on word intuition in particular. Word intuition is speakers' intuitive knowledge on wordhood, i.e., what a word is. Laymen's word segmentation behavior is not instructed by linguistic theories on word, but by their word intuition, hence reflects their word intuition; because of this, the word segmentation task has been used to measure and study word intuition (王立, 2003; Hoosain, 1992). The basic idea is like this: if a Chinese sentence is segmented by, for example, 100 subjects, we can then observe what slices of the sentence are consistently treated as words by these subjects, what slices are consistently treated as non-words, and what slices are not so consistent by being treated as words by some and non-words by others. This kind of segmentation consistency can be a convenient measurement of Chinese speakers' word intuition.

Word intuition per se is an important issue awaiting more research which can contribute to the investigation of cognitive mechanism of humans' language competence and shed new light on the theoretical problem of word definition for the theoretical definition of word should generally accord with the speakers' word intuition (王洪君, 2006; 王立, 2003; 胡明扬, 1999; 陆志韦, 1964).

Semantic transparency/compositionality of a multi-morphemic form, simply speaking, is the extent to which the lexical meaning of the whole form can be derived from the lexical meanings of its constituents. More accurately speaking, this definition is merely the definition of overall semantic transparency (OST) of a multi-morphemic form; besides that, there is constituent semantic transparency (CST) too which means the extent to which the lexical meaning of each constituent as a independent lexical form retains itself in the lexical meaning of the whole form.

In the context of theoretical linguistics, semantic transparency is used as an empirical criterion of wordhood (Duanmu, 1998; 吕叔湘, 1979; Chao, 1968), but for Chinese disyllabic forms this criterion seems to be ignored to some extent by some linguists based on word intuition (王洪君, 2006; 冯胜利, 2004; 王立, 2003; 冯胜利, 2001; 胡明扬, 1999; 冯胜利, 1996; 吕叔湘, 1979); it is also treated as an indicator of lexicalization (Packard, 2000; 董秀芳, 2002; 李晋霞 and 李宇明, 2008). In the context of psycholinguistics, it is an "extremely important factor" (Libben, 1998) affecting the mechanism of mental lexicon, for example the representation, processing/recognition, and memorizing of multi-morphemic words (Han et al., 2014; Mok, 2009; 王春茂 and 彭聃龄, 2000; 王春茂 et al., 2000; 王春茂 and 彭聃龄, 1999; Libben, 1998; Tsai, 1994). Following this line of investigations, it is significant to examine the role semantic transparency plays in Chinese speakers' word intuition towards Chinese disyllabic forms. When we build the dataset, we carefully select sentence stimuli which contain word stimuli that cover all possible kinds of semantic transparency types to enable us to examine the role semantic transparency plays in word intuition of Chinese speakers.

The widely used Chinese segmented corpora, for example, the Sinica corpus (Chen et al., 1996), are usually segmented firstly by segmentation programs and then revised by experts according to certain word segmentation standard. From the inconsistent segmentation cases we can find plenty useful information to explore word intuition. But from the perspective of the measurement of Chinese speakers' word intuition, the data are biased by segmentation programs and word segmentation standards, so they are not so suitable and reliable for this purpose.

In order to better serve the studies of word intuition of Chinese speakers, we need manual word segmentation datasets. In such a dataset, each and every sentence is segmented manually by a large group of laymen, say 100, without the influence of any linguistic theory or any Chinese word seg-

mentation standard. This kind of dataset which is both large and publicly accessible, to the best of our knowledge, is still a gap in Chinese language resources.

And the third purpose is that the resultant manual Chinese word segmentation dataset may have potential applications in the studies of Chinese language processing especially in the studies of automatic Chinese word segmentation and cognitive models of Chinese language processing.

2 Construction

2.1 Materials

The stimuli 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 decide to use more than 150 long sentences (the crowdsourcing method makes this possible). Meanwhile, the resultant dataset must be able to support the examination of the effect of semantic transparency on word intuition; so these sentence stimuli should also contain the words which cover all the word stimuli to be used in the examination of semantic transparency effect. So the stimuli selection procedure consists of two steps: (1) word selection, i.e., to select an initial set of word which covers all the word stimuli 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.

Word Selection

We have already created a crowdsourced semantic transparency dataset *SimTransCNC 1.0* which contains the overall and constituent semantic transparency rating data of about 1,200 Chinese bimorphemic nominal compounds which have mid-range word frequencies (Wang et al., 2014a). Based on this dataset, 152 words are selected, for the distribution of these words, see Table 1.

These words are bimorphemic nominal compounds of the structure modifier-head, and cover three substructures: NN, AN, and VN. 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 scores and the most balanced CST scores, e.g., "江水"; OO words have the lowest OST scores and the most balanced CST scores, e.g., "脾气"; TO and OT words bear mid-range OST scores and the most imbalanced CST scores, e.g., "音色" (TO) and "贵人" (OT).

	Word Structure		
Transaprency Type	NN	AN	VN
TT	20	10	10
TO	20	6	10
OT	20	10	10
OO	20	10	6

Table 1: Distribution of types of selected words.

Sentence Selection

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

	Length of Sentence
Min	20
Max	46
Sum	4,946
Mean	32.54
SD	5.46

Table 2: Length distribution of selected sentences.

2.2 Crowdsourcing Task Design

These 152 sentence stimuli 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 *Amerzon Mechanical Turk* (MTurk), *Crowdflower* is a more feasible platform for Chinese linguistic data collection (Wang et al., 2014b; Wang et al., 2014a).

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 on-line subjects to provide their identity information on gender, age, level of education, 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 that character and type it in the text-box blow 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, asks the subject to identify the antonymous character of the given one from 10 different characters. The section 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 sentence stimuli 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 right or wrong, but just follow their intuition.

Parameters of Tasks

These eight crowdsourcing tasks are created with the following parameters: (1) each worker 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, United States, and New Zealand; (4) we pay 0.25USD for one response.

Quality Control Measures

The following quality control measures are used: (1) the section 4, screening questions, is used to discriminate Chinese speakers from non-Chinese speakers and to block bots; (2) the section 5, the segmentation task, will keep invisible unless the first two screening questions are correctly answered; (3) the answers to the segmentation questions in section 5 must comply with 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 will be blocked unless all the required questions are answered and the answers satisfy the above conditions; (5) data cleansing will be 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 two 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 three to six days; then the speed became slower and slower, it eventually took us about 1.3 months to reach our aim; after all, *Crowdflower* is not a Chinese native crowdsourcing platform, this kind of speed is understandable.

2.4 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 seriously 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 seg-

ments of its 19 sentences; (4) the completion time of the response is shorter than five minutes; (5) the completion time of the response is longer than one hour. Invalid responses were ruled out; the numbers of valid response of the eight tasks are listed in Table 3.

2.5 Results

The resultant dataset contains the manual Chinese word segmentation data of 152 sentences whose length 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 ($M = 133.5, SD = 7.37$).

Task	Valid Response	%
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

Table 3: Numbers of valid response of the tasks.

3 Evaluation

Although Fleiss' kappa can be used to measure the agreement between raters, high agreement does not necessarily means 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, a evaluation method called manual segmentation error rate (MSER) is proposed to evaluate the resultant dataset.

3.1 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", the segment "好东" cannot be treated as one unit/word, because it is incomprehensible. An indivisible segment usually cannot be divided, because it is an fixed unit and its lexical meaning cannot be derived easily from the lexical meanings of its constituents (or semantically opaque); it will become incomprehensible if it is divided; for example, in the phrase example, the segment "东西" 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" 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; so 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 Manual Segmentation Error Rate

A subject divides the phrase/sentence S into n ($n \in \mathbb{N}^+$) 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"), or no error. Suppose $e^{'}$ ($e^{'} \in \mathbb{N}$) is the number of times the type I error occurred during the segmentation process, and $e^{''}$ ($e^{''} \in \mathbb{N}$), 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/", $e^{'} = 2$, $e^{''} = 1$, $n = 2$, 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's consider its collective form; if S is segmented by m ($m \in \mathbb{N}^{+}$) subjects, and the ith ($1 \leqslant i \leqslant m$) 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.

3.3 Evaluation Procedure and Results

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

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	68375	522	353	
	Mean	2848.96	21.75	14.71	.013
	SD	989.76	8.51	6.3	.004

Table 4: Segmentation error rates ($MSER$) of the segmentation results of the eight tasks.

4 Conclusion

We created the manual Chinese word segmentation dataset *WordSegCHC 1.0* using the crowdsourcing method; to the best of our knowledge, there is no publicly available resources of this kind; it can support the studies of word intuition especially the effect of semantic transparency on word intuition and has potential applications in Chinese language processing.

We also proposed an evaluation method called manual segmentation error rate ($MSER$) to evaluate manual word segmentation dataset. The error rate of the dataset is proved to be very low, and this indicates that its data quality is reliable.

This work also confirmed again that the crowdsourcing method is a feasible, convenient, and reliable tool to collect linguistic data. And through this work, a reusable general framework of crowdsourcing linguistic data collection is also presented. Following this framework, larger similar Chinese language resources can be constructed.

We will use this dataset to examine the role of semantic transparency in word intuition of Chinese speakers and to induce the factors affecting word intuition. The consequent discoveries will deepen our understanding of the word definition problem in the Chinese language which has both theoretical and applicational significance.

In the future, once the factors modulating Chinese Speakers' word intuition are clear, perhaps a computational cognitive model of Chinese word segmentation (Wu, 2011) can be proposed and we believe that this could be an interesting new direction of Chinese word segmentation research.

Acknowledgments

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

References

M Allahbakhsh, B Benatallah, A Ignjatovic, HR Motahari-Nezhad, E Bertino, and S Dustdar. 2013. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, 17(2):76–81.

Tara S Behrend, David J Sharek, Adam W Meade, and Eric N Wiebe. 2011. The viability of crowdsourcing for survey research. *Behavior research methods*, 43(3):800–813.

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

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

Yuen Ren Chao. 1968. *A grammar of spoken Chinese*. University of California Pr.

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

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

San Duanmu. 1998. Wordhood in chinese. *New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese*, pages 135–196.

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

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

Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67. ACM.

Gary Libben, 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.

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

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

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

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

Jerome L Packard. 2000. *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge University Press.

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

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

Chih-Hao Tsai. 1994. Effects of semantic transparency on the recognition of chinese two-character words: Evidence for a dual-process model. Master's thesis, Graduate Institute of Psychology, National Chung Cheng University, Chia-Yi, Taiwan.

Shichang Wang, 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*, pages 147–156, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Shichang Wang, 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)*, pages 105–113, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Zhijie Wu. 2011. A cognitive model of chinese word segmentation for machine translation. *Meta : journal des traducteurs / Meta: Translators' Journal*, 56(3):631–644, 9.

冯胜利. 1996. 论汉语的“韵律词”. *中国社会科学*, (1):161–176.

冯胜利. 2001. 从韵律看汉语“词”“语”分流之大界. *中国语文*, (1):27–37.

冯胜利. 2004. 论汉语“词”的多维性. *当代语言学*, 3(3):161–174.

吕叔湘. 1979. *汉语语法分析问题*. 商务印书馆.

李晋霞 and 李宇明. 2008. 论词义的透明度. *语言研究*, (3):60–65.

王春茂 and 彭聃龄. 1999. 合成词加工中的词频, 词素频率及语义透明度. *心理学报*, 31(3):266–273.

王春茂 and 彭聃龄. 2000. 多词素词的通达表征: 分解还是整体. *心理科学*, 23(4):395–398.

王春茂, 彭聃龄, et al. 2000. 重复启动作业中词的语义透明度的作用. *心理学报*, 32(2):127–132.

王洪君. 2006. 从本族人语感看汉语的“词”. *语言科学*.

王立. 2003. *汉语词的社会语言学研究*. 商务印书馆.

胡明扬. 1999. 说“词语”. *语言文字应用*, 3.

董秀芳. 2002. *词汇化: 汉语双音词的衍生和发展*. 四川民族出版社.

陆志韦. 1964. *汉语的构词法*. 科学出版社.