Accepted for Publication in ***Language and Linguistics 20.2. April 2019*** by Dr. *Shichang Wang* (PhD from CBS, 2016, currently at Shandong University), Prof. Huang, Dr. Yao, and Dr. Chan is a paper exploring the relation between lexical semantic processing and headedness based on semantic transparency. The study used the innovative method of crowdsourcing to build a semantic transparency dataset that included transparency rating for each compound and both of its constituent roots. Colleagues who are interested in reading and commenting on the pre-final version please contact the authors. The semantic transparency dataset will be available worldwide through LDC, UPenn and is open to our colleagues for research. Please see second page for description of the dataset.

Shichang Wang, Chu-Ren Huang,Yao Yao and *Wing Shan Angel Chan. 2019.* **The effect of morphological structure on semantic transparency ratings**. *Language and Linguistics 20.2.* April 2019.

Abstract

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 its 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 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 judgements 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: compound semantic transparency, constituent semantic transparency, semantic transparency dataset, headedness, crowdsourcing

SemTransCNC 1.0:
Semantic Transparency of Chinese Nominal Compounds 1.0

Authors: Shichang Wang, Chu-Ren Huang, Yao Yao, Angel Chan

SemTransCNC 1.0 is a semantic transparency dataset of Chinese nominal compound which was built using a series of Mechanical Turk-based experiments. It consists of the overall and the constituent semantic transparency (OST, CST respectively) data of 1,176 dimorphemic Chinese nominal compounds which consist of free morphemes and which have mid-range frequencies. The construction methodology of this dataset is described in details in Wang et al. (2014) and Huang and Wang (2016). Some of the important features of the datasets are further elaborated in Wang et al. (2019). The dataset is in CSV format which has 11 columns. The columns "WORD" and "WORDT" list Chinese nominal compounds in simplified and traditional Chinese characters respectively. The column "STRUCT" stores the morphological structure of the compounds. The columns "FREQ" and "RFREQ" show the absolute and relative frequencies respectively of the compounds according to the Sinica Corpus V4.0. The columns "NOST", "NC1CST", and "NC2CST" respectively store the overall semantic transparency value of each compound and the constituent semantic transparency values of both constituents of each compound; in these columns, 1 means completely transparent and 0 means completely opaque. The columns "OSD", "C1SD", and "C2SD" are the standard deviations of the overall and constituent semantic transparency rating data of each compound.

Wang, Shichang, Chu-Ren Huang, Yao Yao and Angel Chan. 2014. Building a Semantic Transparency Dataset of Chinese Nominal Compounds: A Practice of Crowdsourcing Methodology. Proceeding of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014), at the 25th COLING, pp. 147-156. Dublin. http://anthology.aclweb.org/W/W14/W14-58.pdf#page=155

Huang, Chu-Ren, and Shichag Wang. 2016. 众包策略在语言资源建设中的应用. The Application of Crowdsourcing Strategy in Utilizing Language Resources. Chinese Journal of Language Policy and Planning (语言战略研究). 2016.6.36.46.

Shichang Wang, Chu-Ren Huang,Yao Yao and *Wing Shan Angel Chan. 2019.* The effect of morphological structure on semantic transparency ratings. *Language and Linguistics 20.2*