Universal or variation? Semantic networks in English and Chinese

Understanding the structures of semantic networks can provide great insights into lexico-semantic knowledge representation. Previous work reveals small-world structure in English, the structure that has the following properties: short average path lengths between words and strong local clustering, with a scale-free distribution in which most nodes have few connections while a small number of nodes have many connections[1]. However, it is not clear whether such semantic network properties hold across human languages. In this study, we investigate the universal structures and cross-linguistic variations by comparing the semantic networks in English and Chinese.

Network description To construct the Chinese and the English semantic networks, we used Chinese Open Wordnet[2,3] and English WordNet[4]. The two wordnets have different word forms in Chinese and English but common word meanings. Word meanings are connected not only to word forms, but also to other word meanings if they form relations such as hypernyms and meronyms (Figure 1).

1. Cross-linguistic comparisons

Analysis The large-scale structures of the Chinese and the English networks were measured with two key network metrics, small-worldness[5] and scale-free distribution[6].

Results The two networks have similar size and both exhibit small-worldness (Table 1). However, the small-worldness is much greater in the Chinese network (σ = 213.35) than in the English network (σ = 83.15); this difference is primarily due to the higher average clustering coefficient (ACC) of the Chinese network. The scale-free distributions are similar across the two networks, as indicated by ANCOVA, F (1, 48) = 0.84, p = .37. The results suggest both universal structure and cross-linguistic variation between the two languages.

2. Cross-linguistic variation

Analysis To explore the differences based on the properties of the ACC, we counted the number of words having semantic meanings that are mutually related. To further investigate whether the network clustering varies across different word categories, we extracted four subgraphs each for a different word category. We then calculated the ACC of each subgraph.

Results Compared to the English network, the Chinese network contains more words having meanings that are mutually connected (Figure 2b and 2c). This pattern may suggest the meanings of a Chinese word are more likely to be mutually related than the meanings of an English word.

A within-language rather than a between-language variation was also observed (Figure 2). Although the words in the Chinese network have a higher ACC than the ones in the English network across all the word categories, the nouns in the Chinese network have the highest ACC (C = 0.24), whereas in the English network, the verbs are the ones with the highest ACC (C = 0.05). The results suggest that, within each language, Chinese nouns and English verbs are the categories with more connected semantic meanings.

Discussion Despite similar scale-free distributions, words in the Chinese network have more connected semantic meanings than words in the English network. The relative patterns of connectedness for nouns and verbs differ between the two languages. Such differences could arise from many cross-linguistic properties. For example, Chinese radicals convey more specific semantic information[7] and therefore could act to connect the semantic meanings of words during the learning and processing of Chinese words. The radical 氵*(water)* directly provides the substance of 河*(river)*, whereas the radical 扌*(hand)* indicates hand-related actions in words such as 打*(hit),* and such examples are pervasive in Chinese. Finally, Chinese radicals in nouns may play a more concrete role than those in verbs[8].

References
1. Steyvers, M., & Tenenbaum, J. B. (2005). Cognitive Science.
2. Wang, S., & Bond, F. (2013). International Joint Conference on Natural Language Processing, October.
3. Bond, F., & Foster, R. (2013). In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics
4. Christiane Fellbaum (1998, ed.) Cambridge, MA: MIT Press.
5. Humphries, M. D., & Gurney, K. (2008). PloS one.
6. Barabási, A. L., & Albert, R. (1999). science.
7. Huang, C. R., & Hsieh, S. K. (2015). The Oxford handbook of Chinese linguistics (pp. 290-305).
8. Zhang, J., Fang, Y., Chen, X. (2006). Acta Psychologica Sinica.

Table 1
Summary statistics for semantic networks

Network	Type	*n*	<*k*>	*C*	*L*	C_r	L_r	σ	γ
Chinese	Forms	48094	1.36						
	Meanings	33788	4.06	.037	12.66	$1.22e^{-4}$	8.99	213.35	-3.22
English	Forms	51705	1.40						
	Meanings	35220	4.12	.008	11.23	$7.50e^{-5}$	8.87	83.15	-3.19

Note. *n* = the number of nodes; <k> = Degree (i.e. the average number of connections); C/C_r = the average clustering coefficient of the network or its random network; L/L_r = the average shortest path length of the network or its random network. σ = the small-world-ness coefficient. The bigger the coefficient is, the greater the small-worldness. γ = the power-law exponential for the degree distribution. Word forms and word meanings were measured separately for *n* and <*k*> but together for other coefficients.

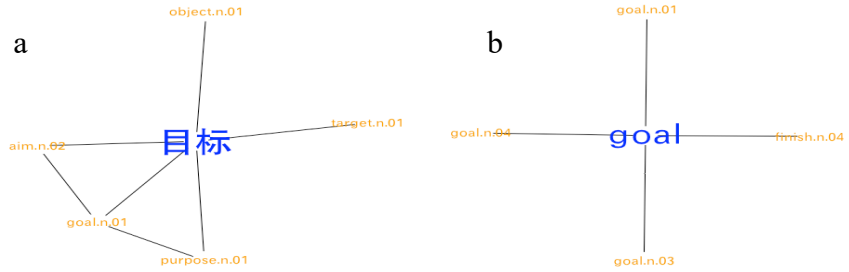

Figure 1. The examples of the Chinese (a) and the English (b) networks. A node could be a word form (in blue) or a word meaning (in yellow). The Chinese word form *目标* in (a) and the English word *goal* in (b) have overlapping semantic meaning *goal.n.01* (meaning 'the state of affairs that a plan is intended to achieve and that (when achieved) terminates behavior intended to achieve it'[4]). But some of the word meanings are unique in *目标* such as *aim.n.02* (meaning 'the goal intended to be attained (and which is believed to be attainable)'[4]), whereas others are unique in *goal* such as *finish.n.04* (meaning 'the place designated as the end (as of a race or journey)'[4]).

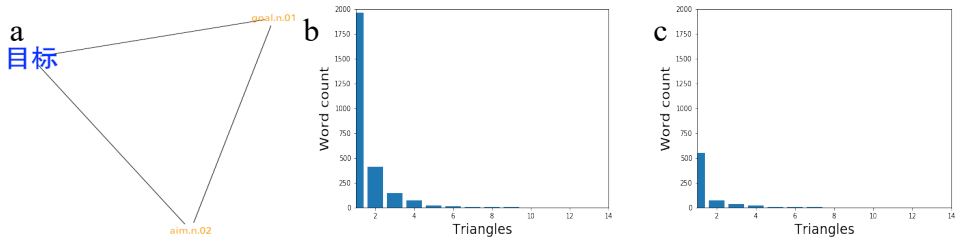

Figure 2. An example of a word form and its two neighboring word meanings that form a triangle relation (a) and the word count as the function of the number of triangles each word has in the Chinese (b) and the English (c) network. Compared to English, there are a greater number of words in Chinese having meanings that are mutually connected.

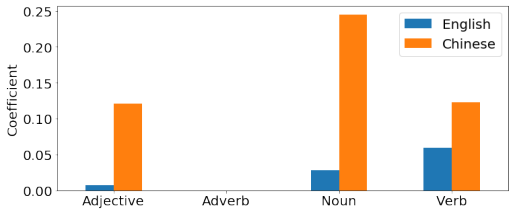

Figure 3. The average clustering coefficient of the subgraphs in each word category. The average clustering coefficients for the adverbs are zero in both networks.

Supplementary information

Network The semantic networks in this study consist of nodes and undirected edges that connect pairs of nodes.

Measurements

Degree. The average number of edges each node has. Degree was measured separately for nodes as word forms and nodes as word meanings.

Average clustering coefficient (ACC). Clustering coefficient (*C*) represents the probability that the two neighbors of a node are themselves neighbors. *C* for each node is calculated by equation 1, where *T(u)* represents the number of connections between the neighbors of node *u*, *k(u)* is the degree of node *u*[5]. We took the average of *C*s over all nodes to calculate ACC.

$$C = \frac{2T(u)}{k(u)(k(u)-1)} \quad (1)$$

Average shortest path length (L). Shortest path length represents the minimum number of edges that link two nodes. *L* was calculated by taking the average of shortest path length over all pairs of nodes.

Small-worldness (σ). σ is computed as: $\sigma = C/C_r \,/\, L/L_r$ where *C* and C_r refer to the average clustering coefficient of the network and a random refence, and *L* and L_r are the average shortest path length of the network and its random reference respectively[5]. The random reference is a random network with equal size and density as the network being analyzed. Size is defined by the number of nodes and density is defined by the probability that the two nodes are connected[5].

Scale-free distribution. Power-law regression was fit[6]: $P(k) \sim k^{-\gamma}$ where *k* is the degree, *P(k)* is the probability that a node has degree *k*, which is degree of the node divided by the overall number of nodes. γ is the power-law exponential for the distribution.

Additional analyses and results

The subgraph for a word category. Word categories were labeled in word meanings (e.g. *goal.n.01* denotes meaning in noun). To construct a subgraph given a particular word category, we extracted all the word meanings belonging to this category and all the word forms with at least two meanings of this category. For example, *goal* will be included in the noun subgraph because it contains more than two noun meanings.

Similar size between two networks. The Chinese and the English networks have similar size in terms of the number and the degree of word forms and word meanings(Table 1). This rules out the alternative interpretation that the difference of small-worldness between two networks is caused by different basic structures between them.

Scale-free distribution. P(k) and *k* were log-transformed into $\log P(k)$ and $\log k$ before fitting into a linear regression model. The models fitting the Chinese and the English data both show that $\log \mathrm{k}$ significantly predicted $\log P(k)$: for the Chinese model, $F(1, 20) = 2025$, $p < .001$, $R^2 = .99$ and for the English model, $F(1, 28) = 1520$, $p < .001$, $R^2 = .98$. The results suggest that it is more likely to have a node with smaller degree than a node with greater degree. The degree distribution follows the power-law, indicating that most nodes have a small degree while only a few nodes have a large degree.