ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 4, No. 6, pp. 1138-1143, June 2014

doi:10.4304/tpls.4.6.1138-1143

Corpus-based Analysis of Semantic Transparency between High Frequent English and Chinese Compounds

Wenyan Ma
School of Foreign Languages, Beijing Institute of Technology, Beijing, China

***Abstract*—From psycholinguistic and lexical semantic aspect, the semantic transparency of 2000 nominal English and Chinese high frequent compounds in the corpus have been analyzed, and related with word frequency. The result showed that in both languages, the number of Transparent-Transparent and Partially-Transparent compounds is larger than that of Opaque-Opaque compounds. Moreover, the relationship between compound frequency and the degree of semantic transparency is different between English and Chinese. Both of these results reflect the common features of mental lexicon process and differences in lexical structures in English and Chinese.**

***Index Terms*—corpus, high frequent English compound, high frequent Chinese compound, semantic transparency, word frequency**

I. Introduction

A considerable amount of research has been carried out on the mental representation and processing of compound words, consisting of more than one morpheme, e. g storybook. One of the main questions in this field is semantic transparency. It is normally believed that a semantically transparent compound can be understood by those who have never heard the word before, e. g storybook and 课桌 (desk) can be understood as a combination of the meanings of story and book, or 课 (lesson) and 桌 (desk). Whereas an opaque compound like humbug, 旗下 (subordinate), only knowing the constituent morphemes hum and bug or 旗 (flag) and 下 (under or below) cannot help understand the meaning of the word.

The notions of 'transparent' and 'opaque' refer to the degree of semantic transparency. It varies along in a single continuum within the same processing system from fully transparent to fully opaque. For example, one transparent type like lunchtime, 车主 (car owner), both of their word meanings can be completely identified from the constituent morphemes meanings. So the meanings of morphemes are apparent to the meanings of the words. But in another opaque type like black sheep, 旗下 (subordinate), their word meanings cannot be speculated or inferred by the constituents morphemes. These two words are extremely different in terms of degrees of semantic transparency. Therefore, semantic transparency is supposed to be a continuum process in which there are more than two clearly cut degrees. For example, shot gun, 抱歉 (be sorry/regret), fire engine, etc. Their degrees of semantic transparency are not the same as the two extremes, because only one instead of two in these words is efficient in meaning computations. In this aspect, semantic transparency reflects the relationship between compound word and its constituent morphemes. For clarity, a morpheme is defined as a minimal form/unit (orthographic and /or phonological) that carries meaning in a multimorphemic string (Li & Thompson, 1981).

There are many studies proved the centrality of semantic transparency in the processing of multimorphemic words. Laudana and Burani (1995) claimed that semantic transparency determines the presentation route of a multimorphemic word-whether in a whole-word recognition or through a morphological decomposition. The research conducted by Marslen-Wilson, Tyler, Waksler, and Older (1994) also supports the importance of semantic transparency. It is found that semantic transparent form has more significant whole-word constituent priming effects than semantic opaque. In the investigation of Schreuder and Bayyen (1995), a meta-model of morphological processing has been presented in which semantic transparency determines whether a multimorphemic form has its own representation or in terms of the constituents.

Though above studies are diverse in underlying assumptions and theories, they agree on the point that any discussion of multimorphemic words related to mind processing would have to include an account of how semantic transparency works. In this paper, we analyze and compare semantic processing of multimorphemic words-compounds in English and Chinese from the aspect of semantic transparency itself. It is claimed that semantic transparency is an important aspect to the study of compound processing from both psycholinguistics and semantics.

II. Semantic Transparency and Word Frequency

Semantic transparency is often considered to reflect the relationship between a multimorphemic word and its constituent morphemes, it is therefore preferably used in semantics studies, especially in Chinese lexical research. The first one to analyze Chinese words from the term of semantic transparency is in Li & Li's research (2008). Because Chinese is fruitful in compounds, mostly the words in the analysis are compounds. These words have been categorized into 4 degrees according to the representation of the constituent morphemes meanings in the whole word meaning, which include fully transparent, partially transparent, partially opaque and fully opaque. It is concluded that 4 degrees of semantic transparency from fully transparent to fully opaque reflects the diachronic process of lexicalization and structuralization.

In compound studies, semantic transparency often works together with word frequency. In psycholinguistics, like the research of Chinese compounds processing and representation, Mok, L (2009) manipulated the semantic transparency and word frequency, found that the higher the word frequency is, the more transparent the word semantics is. Besides, the meanings of the morphemes are easily identified in the compound with comparatively high word frequency and semantic transparency. In Pollatsek, A (2005), word frequency and semantic transparency are regarded as parameters again to analyze the Finland compound processing and representation. The result found that semantic transparency works with word transparency decides correct outputs in processing.

Normally in psycholinguistic research, researchers would manipulate more than one factor like typical compounds and pseudowords together to infer to the working models in compound processing, i.e, in the research of Mok and Pollatsek, both of them chose some typical compounds and design the same number of pseudowords as well. It is true that the result by the method of choosing typical compounds with intention can reflect typical mental representations in different semantic situations. Meanwhile, it is unavoidable to mislead that brain always work in typical instead of common situations.

In Semantics study, the common distributions of semantic transparency for a limited number of common frequent rather than typical compounds is often taken as the objective in the research. For example, in Dong's (2011), 500 Chinese compounds by frequencies with bisyllables from Modern Frequency Chinese Dictionary were selected and analyzed. Dong classified 5 ways of semantic transparency for these 500 compounds, and found that most of the compounds are fully transparent or partially transparent, semantic transparency can significantly influence the learning process, and transparent compounds could decrease the difficulties for non-native learners.

Truly, this research reflected partially the common semantic features of Chinese common compounds, but how semantic transparency reflects the word frequency, and how word frequency influences the distributions of semantic transparency, is there any relationship between these two variables? Little evidence could be found. And it is believed that one single language study with limited data objects is not strongly persuasive to prove any significance in linguistic studies.

Therefore, current study will focus on two language comparisons in the light of psycholinguistics and semantics, semantic transparency of 1000 English and 1000 Chinese common frequent compounds will be analyzed, and the relationship of the variables –semantic transparency and word frequency will be correlated with the help of SPSS 16.0. The following hypothesis will be tested in the research:

■What are the common distributions of semantic transparency in English and Chinese high frequent compounds?

■What are the relationships between semantic transparency and word frequency?

III. Data Collection

In defining compound, we referred to the approach used in Huang (1998)'s study, that is, compounds constitute two lexical items connected by syntactic rule, and this compound can be analyzed into two or more meaningful elements or morphemes. 2000 high frequent compounds nouns with two meaningful morphemes were chosen to build an English-Chinese compound nouns corpus in the research. Compound nouns take up large percent in both English and Chinese. So compound nouns analysis is meaningful to compound processing and representations.

As for the data collection, 1000 English compounds nouns were selected from 11521compound words in *Longman Dictionary of Contemporary English* (2004) by frequency index from British National Corpus (BNC). It is shown that the frequency index of these 1000 English compounds nouns distribute from 11433 to 140, which can reflect the commonality of the research data in an English Spoken country. Another 1000 Chinese compounds were selected from *A Frequency Dictionary of Mandarin Chinese (2009)*, the data in the dictionary are comparatively update, which cover materials of spoken, novels, news, etc. on the basis of 50 million Chinese words. In the corpus, we selected 1000 compound nouns according to the frequency index labeled in the dictionary, in order to guarantee the form and the speech in one-to-one correspondence, Modern Chinese Dictionary (5^{th} edition) has been used to match the frequent meaning with the noun form of the compounds.

Because of the different semantic relationship between the constituent morphemes, and the whole word meaning, semantic transparency has been graded as different degrees. (Libben 2003, Li & Li 2008) The typical (Libben, 2003) includes four degrees, they are TT (Transparent-Transparent), OT (Opaque-Transparent), TO (Transparent-Opaque) and OO (Opaque-Opaque). In the current research, in the principle of the relationship between morpheme meaning and word meaning as Libben, the semantic transparency of the compounds will be graded as 3, like TT (Transparent-Transparent), PT (Partially-Transparent), which include OT and TO in Libben's classification, and OO

(Opaque-Opaque).

IV. Results and Discussion

A. Semantic Transparency Distributions

According to the classification of semantic transparency, the 2000 compound nouns had been labeled with different degrees. It was found that the semantic transparency distribution of English is uneven as that of Chinese in Table 1.

TABLE 1.
SEMANTIC TRANSPARENCY DISTRIBUTIONS IN ENGLISH AND CHINESE

English Compounds (N=1000)				Chinese Compounds (N=1000)		
	TT	PT	OO	TT	PT	OO
M	493.13	457.81	614.52	343.28	420.71	570.09
SD	662.284	875.505	1041.694	428.45	562.059	839.066
Per%	57%	27.7%	15.3%	55.7%	33.3%	11%

TT accounted for the top, and OO for its obscurity, accounted for the fewest in the corpus in both languages. The result on Chinese compound nouns was consistent with Dong's (2011). Meanwhile, the mean of the word frequency in Chinese is not as large as that of in English, and the standard deviations of English in all of the types are remarkable, for their large densities in word frequency distribution in the corpus.

As for the distributions of different degrees of semantic transparency, we analyzed the most prominent one, TT, and found that the constituents of word meaning related to morpheme meanings can be basically identified as two kinds, one is C=A+B. This type can be found both in English and in Chinese. For example,

In English

lunchtime: the time in the middle of the day when people usually eat their lunch.

lunch: a meal eaten in the middle of the day.

time: minutes or hours etc.

In Chinese

车主 (Car owner): owner of a vehicle.

In the corpus, the type of C=A+B is the most popular with English TT compounds, like *story book*, *newspaper*, *social service*, etc. the compound meaning is a combination of the morpheme meaning, the representation process is the process of morphological decomposition of the words, and both morphemes in a compound contribute systematically to the meaning of the compound word as a whole.

The other TT compounds is C=A=B, that means the compound meaning is represented by morphemes, and either of them contributes systematically to the meaning of compound word. But different from the previous type, the word meaning C is not a kind of meaning combination of the morphemes, like A+B, in some aspect, the meaning of C is equivalent to either morpheme meaning A or B. And either A or B is apparent in C. It is found that comparing with English compounds, it is more prominent in Chinese, about 33% in the corpus. For example, 朋友 (friend), the morphemes 朋 and 友 mean friend, they overlap each other in meaning and contribute individually to the whole word. To compute the meaning of 朋友 is like the computation of any morpheme, like 朋 or 友, both can facilitate decisively the processing of the whole word.

From mental lexicon processing and representation, whether the word is the type of C=A+B or C=A=B, when computing the meanings of the compounds, the mind will speculate on the two morphemes meanings automatically, especially for processing novel words. If the two morphemes are completely transparent in semantics, that will reduce the bearing load of the brain in processing and representation. In processing TT word meanings, the mind doesn't project in TIME or SPACE, instead, the morphemes meanings can represent the most meanings of words, and this way will definitely facilitate the communications. From this point, the processing and representation of TT compounds doesn't need to waste too much time or energy, so it is in line with the "economy principle", and this may explain the reason why the number of TT compounds is larger in both languages in the corpus.

Different from TT compounds, for PT compounds, the word meaning cannot be completely decomposed from the morpheme meanings, and it has a semantic relationship with only one constituent morhpeme. For example, in English.

shotgun: a long gun fired from the shoulder that shoots many small round balls at one time, used especially for killing birds or animals.

shot: when someone fires a gun, or the sound that this makes.

gun: a weapon from which bullets are fired.

The meaning of gun overlaps the meaning of the word shotgun, while, the meaning of shot is not apparent to the meaning of the word shotgun.

In Chinese, the meaning of 抱歉 (be sorry/ regret) is not the easily decomposed meaning of the constituent morphemes, 抱(hold or carry in the arms) plus 歉 (feel sorry/ apologize), and only 歉 overlaps the meaning of the word 抱歉(be sorry / regret), morpheme 抱 is opaque related to the compound meaning.

In both of these shotgun and 抱歉 (be sorry / regret), part of the morpheme meanings overlap the meaning of words,

like gun in shotgun and 歉 in 抱歉, and other morpheme meanings have to be transformed and projected either in time or in space, to process the word meaning, which will cost more time in computation. And for the relations between word and morpheme, only one morpheme is transparent.

In addition, we identified two types of PT compound in the corpus. One is OT, i.e Opaque A +Transparent B. Another is TO, i.e Transparent A + Opaque B. English compounds are typical rightmost-centered, namely the morpheme on the right side is decisive for the word semantics and morphology, etc. And according to the research done by Libben, Gibson, Yoon, and Sandra (2003), it was found that the English compounds with opaque heads took longer to recognize than the compounds with transparent heads. This is because of an effect of the opacity of the morphological head that occurs at the right morpheme of English words. So for English PT compounds, the degree of semantic transparency of OT is higher than that of TO. For example, OT compound, shotgun and TO compound, fire engine. In shotgun, the rightmost morpheme gun determines the semantic and morphological category of the word, and its meaning overlaps the meaning of the word shotgun, so the word semantic is easier and more transparent to infer. While, for TO compound, like fire engine, the rightmost morpheme engine, is opaque in meaning, that is, it showed little hint in the word meaning, so the degree of whole word semantic transparency has been decreased by this decisive morpheme.

Different from English, Huang (1998), after analyzing 24,000 bisyllabic modern Chinese compounds, proposed that Chinese is neither left-centered nor right-centered, it is 'headless language in its compounding morphology'. (Huang, 1998) That is, any one component morpheme cannot fully determine the whole compounds either in semantics or in morphology. Like TO compound 抱歉 (be sorry/regret) and OT compound 当局 (authorities), none of the morphemes can decide the semantic categories of the words, the computation of compound meanings need the support of two morphemes integration, therefore, differ from the English OT compounds, the mental representation of Chinese OT compounds has to rely on the semantic integration of the constituent morphemes. More bearing load may cost more time and energy to compute, which will increase difficulties in communications. This might be the reason for the smaller number of PT compounds in the corpus, comparing with TT compounds.

In OO compounds, the meanings of both morphemes are opaque to the meanings of the words, and morphological decomposition would yield wrong representation of the words. Like the English compound, black sheep, it means someone who is regarded by other members of their family or group as a failure or embarrassment. But one morpheme black means having the darkest color and sheep means a farm animal that is kept for its wool and its meat. Neither black nor sheep has direct relations with the word black sheep. This situation is the same in Chinese, like 旗下(subordinate), the morpheme 旗 means flag, and 下 means under or below. In both of these two compounds, the morphological decomposition does little contribution to the whole-word recognition, the meanings of morphemes are not apparent in the meanings of words, and the computation process rely on the whole lexical form instead of decomposed morphemes. Therefore, OO compounds are considered as the most 'word-like' (Libben et al., 2003) or the most 'unitised' (Mok, 2009) comparing with TT and PT compounds.

It is likely concluded that the decomposing effect of meaning computation from morphemes decreases with the degree of transparency. The more semantically opaque morphemes meanings are as to compound meaning, the less efficient will be the process of morphological decomposition in facilitating whole-word meaning computation. Libben et al. (2003) showed that opaque compounds had a much stronger repetition effect than TT and PT compounds, which means that this type of compounds were more difficult to process, and more difficult to be comprehended in communication, because of more time and energy consuming. This may the reason why there is smaller number of OO compounds in the corpus, and may even in the daily use.

B. Relations between Semantic Transparency and Word Frequency

As discussed in 3.1, the more transparent the compounds are, the larger number the words is, like TT compounds accounted for the most in both languages in the corpus. And the reason TT is popular may because it is easily to be understood and comprehended in communication. Generally speaking, the easier to be communicated, the more frequent the word will be in daily use, this is like Matthew Effect. In the research, we hypothesize that this effect would be represented by the relationship between semantic transparency and word frequency.

To test that there are some kind of relations between semantic transparency and word frequency in both languages, two variables were labeled with different numbers and degrees. For the word frequency, it was marked from 1 to 5 according to the numbers in mean and standard deviation, and each number represent 200 compounds, the larger the number is, the higher the word frequency is. As for the semantic transparency, the 3 types of compounds, TT, PT and OO have been graded with numbers from 1 to 3, the more opaque the word is, the larger the number is. After all of these have been done, the two variables were tested on Pearson correlation coefficient and the bilateral inspection. The result is in Table 2.

TABLE 2.
PEARSON CORRELATION BETWEEN SEMANTIC TRANSPARENCY AND WORD FREQUENCY

		Semantic transparency	Word frequency
English compound semantic transparency	Pearson Correlation	.007	1
	Sig. (2-tailed)	.833	
	N	1000	1000
Chinese compound semantic transparency	Pearson Correlation	.086**	1
	Sig. (2-tailed)	.007	
	N	1000	1000
**. Correlation is significant at the 0.01 level (2-tailed).			

In Table 2, for the high frequent English 1000 compounds, r=0.007<1, p=0.833>0.01, which means that the relation between word frequency and semantic transparency is not significant, the change of semantic transparency cannot remarkably reflect the change of word frequency, three degrees of semantic transparency have spread loosely across the 5 levels of word frequency, from the least frequent to the most frequent in the corpus. Therefore, there is no direct relationship between the two variables in English.

But for Chinese in the corpus, r=0.086, p=0.007<0.01, inspection level is 0.01, word frequency can significantly reflect semantic transparency. When the number of word frequency grows bigger, the degree of semantic transparency becomes higher, vice versa. For example, TT compound, the word frequency is more concentrated in the level of 5 and 4. In contrast, the lower word frequency of compound tends to be opaque in semantic transparency. Like, OO compound, the word frequency is more likely to be concentrated in the level of 2 and 1. This shows that Chinese compound word frequency can obviously reflect the change of semantic transparency, the two variables have direct correlations.

The differences in the relationship between the two variables may be interpreted from the aspect of the different lexical structure distributions in the two languages. In English, although compounding is the most productive in word formation, compound words are not the main type, 45% English words are single morphemes, and compound only accounts to 25% (Dupuy 1974). This can be also found from the English compound word frequency, the distributions are not so concentrated as that in Chinese. And in Chinese, compound words are the main type, especially bimorphemic compounds, which can be accounted for73.6% of the total number, so the word frequency can significantly reflect the distributions of semantic transparency.

V. CONCLUSIONS

Based on the corpus of English-Chinese high frequent compounds, semantic transparency was analyzed qualitatively and quantitatively, and the two languages word frequencies were correlated with their semantic transparency as well. The results showed that both in English and Chinese, TT compounds are the most prominent and OO compounds are the least. As for the relationship of the two variables, English compound is not as significant as that of Chinese. For the similarity of the two languages compounds, this reflects language 'economic' principle, and for the differences in the relationship between semantic transparency and word frequency, this may be interpreted as the differences between the two language lexical structure distributions. The research provides an insight to analyze the characteristics of multimorphemic words in both psycholinguistic and semantic aspects.

There is also some future work that could be done to improve the credibility of the current research. First, the lexical sample in the corpus could be enlarged to include more words not only within the limit of frequency and compound nouns. Second, the word semantic transparency division can be more objective by repeating division work from different participants.

ACKNOWLEDGEMENT

This work was supported in part by a grant from China Scholarship Council (No. 2011307315) and Chinese Degree and Postgraduate Education Society (B1-2013Y03-005)

REFERENCES

[1] A Frequency Dictionary of Mandarin Chinese. (2009). New York: Routledge.
[2] Algeo, J. (1991). Among the new words. *American Speech* 66. 2, 71-80.
[3] Cruise, D. A. (1991). Lexical semantics. Cambridge: Cambridge University Press.
[4] Dong, Y. W. (2011). The study of semantic transparency in Chinese bisyllable word. *Chinese Journal of International* 2. 1, 178-187.
[5] Dupuy H. J. (1974). The rationale, development and standardization of a basic word vocabulary test. Washington: U. S. Government Printing Office.
[6] Fehringer, C. (2012). The lexical representation of compound words in English: evidence from aphasia. *Language Sciences* 34, 65-75.

[7] Huang, S. (1998). Chinese as a headless language in compounding morphology. In J. L. Packard (ed.). *New approaches to Chinese word formation: Morphology, phonology and the lexicon in modern and ancient Chinese*. New York: Mouton de Gruyter, 261-284.
[8] Laudana, A., & Burani, C. (1995). Distributed properties of derivational affixes: implications for processing. In L. B. Feldman (ed.), *Morphological aspects of language processing*. Hillsdale, NJ: Erlbaum, 345-364.
[9] Li, C. N., & Thomson, S. A. (1981). Mandarin Chinese: A functional reference grammar. Berkeley, CA: University of California Press.
[10] Li, J. X., & Li, Y. M. (2008). Semantic transparency in word meaning. *Language Study* 7, 60-65.
[11] Libben, G., Gibson, M., Bom Yoon, Y. & Sandra, D. (2003). Compound fracture: the role of semantic transparency and morphological headness. *Brain and Language* 84, 50-64.
[12] Libben, G. (1998). Semantic transparency in the processing of compounds: consequences for representation, processing, and impairment. *Brain and Language* 61, 30-44.
[13] Longman Dictionary of Contemporary English (4th edn）(2009）Essex: Pearson Education Ltd.
[14] Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review*, 101, 3-33.
[15] Modern Chinese Dictionary (5thedn) (2005). Beijing: Commercial Press.
[16] Mok, L. W. (2009). Word-superiority effect as a function of semantic transparency of Chinese bimorphemic compound words. *Language and Cognitive Process* 24, 1039-1081.
[17] Pollatsek, A., Hyona, J., & Bertram, R. (2005). The role of semantic transparency in the processing of Finnish Compound Words. *Language and Cognitive Processes* 20, 261-290.
[18] Schreuder, R., & Baayen, H. (1995). Modeling morphological processing. In L. B. Feldman (ed.), *Morphological aspects of language processing*. Hillsdale, NJ: Erlbaum, 345-364.
[19] Wang, C. M, & Peng, D. L., (1999). The roles of surface frequencies, cumulative morpheme frequencies, and semantic transparencies in the processing of compound words. *ACTA Psychologica Sinca*. 3, 266-273.
[20] Wang, C. M, & Peng, D. L., (2000). The access representation of polymorphemic words: decomposed or whole? *Psychological Science*. 23. 4, 395-398.

Wenyan Ma was born in Shenyang, R.R China. She got a Ph. D. in linguistics from Communication University of China in 2010, and currently is a lecturer at School of Foreign Languages of Beijing Institute of Technology. Her research interest includes psycholinguistics and Teaching English for Specific purposes.