The 2018 Conference on Computational Linguistics and Speech Processing
ROCLING 2018, pp. 199-213

On the Semantic Relations and Functional Properties of Noun-Noun Compounds in Mandarin

Shu-Ping Gong
Department of Foreign Languages
National Chiayi University
spgong@mail.ncyu.edu.tw

Chih-Hung Liu
Department of Foreign Languages
National Chiayi University
joker150258@hotmail.com

Abstract

The goal of this research was to determine the distribution of compounds in Mandarin Chinese from the aspect of semantics. In particular, the focus was on two types of compounds: compounds interpreted by semantic relations or by functional properties between constituents. We collected 880 compounds from a dictionary and categorized them into two types of noun-noun compounds in Mandarin, including relation-based compounds (e.g., 中國菜 *zhōngguócài* "Chinese food") and functional property-based compounds (e.g., 柳葉眉 *liǔyèméi* "arched eyebrows"). Finally, the frequency of occurrence of the two types of compounds was determined. The results showed that relation-based compounds occurred much more frequently than property-based compounds in our data (96.1% vs. 3.8%). In addition, it was found that within the relation-based compounds, noun-noun compounds using the FOR relation (e.g., 信紙 *xìnzhǐ* "letter paper") had the highest rates of occurrence (37.6%), while the CAUSE relation (e.g., 刀傷 *dāoshāng* "wounds by a knife") had the lowest rates of occurrence (0.7%). On the other hand, the functional property-based compounds, almost always referring to objects in the NATURAL KIND domain, take metaphorical meanings from individual constituents. Our study suggests that the relation-based meanings for interpreting compounds are common in daily conversation, which could be the dominant strategy people use to interpret novel compounds. This research has practical implications for natural language processing in dealing with segmentation of compounds and multiword expressions in Mandarin and even recognition of novel word combinations.

Keywords: Relation-based Compounds, Property-based Compounds, Natural Kind, Artifact, Corpus, Mandarin.

1. Introduction

Compounds are frequent and pervasive in daily language. Compounds, which are lexemes that consist of two or more words, are not interpreted with ease. Indeed, even though compounds are highly frequent, sometimes it is not easy to interpret (novel) compounds for several reasons. First, the semantic relations within compounds are complex [1]. For example, Gagne & Shoben [2] found that people used more than ten semantic relations to interpret compounds, such as 日月潭紅茶 *rìyuètán hóngchá* "Black Tea of Sun Moon Lake" as interpreted by the PLACE relation. Second, the meanings of some compounds are idiosyncratic, which cannot be inferred from the literal meanings of individual constituents [3]. For example, the compound 片花 *piànhuā* indicates "trailers of movies" and does not refer to "a piece of flower". Third, some compounds have more than one meaning and the appropriate meanings are determined by the context [1]. For example, the compound 出爐 *chūlú* refers to something "freshly baked". This can be food (e.g., bread) or written/innovative products (e.g., theses, books, or movies).

In this study, we focus on compounds. In particular, it is found that compounds can be interpreted by semantic relations or by functional property principles. First, when compounds are interpreted by semantic relations among constituents, these kinds of compounds are called relation-based compounds [2]. Such compounds are derived from the thematic relation between morphemes/words. For example, the compound 檸檬汁 *níngméngzhī* "lemon juice" in Mandarin Chinese belongs to this type. Second, when compounds are interpreted by using the functional properties involved in the meanings of words, these types of compounds are the property-based ones [4] that are constructed via mapping a property from one word to another word. For example, the compound 蝴蝶蘭 *húdiélán* "Phalaenopsis orchid" is this type.

In previous studies, two theories were proposed for accounting for how compounds are understood: The competition-among-relations-in-nominals (CARIN) and the dual-process theory. The CARIN theory proposed by Gagne and Shoben [2] accounted for relation-based compounds, suggesting that these compounds were interpreted based on the semantic relations between modifiers and head nouns. A high frequency of relations between words facilitated the processing of compounds, whereas a low frequency of relations between words inhibited the interpretation of compounds.

On the other hand, the dual-process theory, proposed by Wisniewski [4], involved two independent operations: relation-linking and property-mapping in the process of interpreting compounds. In particular, compounds were first interpreted using the relation-linking operation. If this operation did not yield a meaning, property-mapping was activated and used to interpret the meaning of the compound.

Even though past studies have discussed how the two types of compounds were understood, most have only investigated these compounds in English. In fact, little research has determined the frequency of occurrence of the two types of compounds in Mandarin. Compounds in Mandarin are very pervasive in daily language. Compounds, one kind of multiword expression, are constructed according to some semantic principles, i.e., semantic features and functional property mapping. It is necessary to know the distribution of compound types in order to know how people process compounds. Therefore, the current research intends to examine the distribution of relation-based and property-based compounds in Mandarin Chinese.

2. Background

As mentioned above, two theories have been proposed to describe how compounds are interpreted and understood, namely the competition-among-relations-in-nominals (CARIN) theory [2] and the dual-process theory [4]. The CARIN theory proposes that the meaning of compounds can be unified by a particular thematic relation between modifiers and head nouns, while the dual-process theory proposes that compounds can be interpreted via two independent operations, either relation-linking or property-mapping. In the following sections, each theory will be presented in detail.

2.1 Relation-based Compounds

Gagne and Shoben [2] proposed the CARIN theory to deal with how compounds are interpreted and comprehended. They proposed that speakers will select a specific thematic relation, connecting the modifier and the head noun, to unify the ultimate interpretation. Three experiments testing compounds with three thematic relations supported their theory: highly frequent for both constituents (HH, e.g., "mountain bird"), highly frequent only for the modifier (HL, e.g., "mountain magazine"), and highly frequent only for the noun (LH, e.g., "gas cloud cloud"). Their experimental results demonstrated that the typical usages of modifiers (i.e., HH and HL) can ease the comprehension of the combinations, as the CARIN theory predicted. Moreover, knowledge of highly frequent thematic relations can also influence the ease of comprehending modifier-noun compounds, though the relational information of the modifier is not the sole factor in the ease of interpretation.

Gagne [5] investigated whether property-based compounds were as common as relation-based compounds, and whether the similarity between the modifier and the head noun would affect the interpretation of property-based compounds. Two kinds of novel compounds were tested (Table 1).

Gagne [5] showed that: (1) relation-based compounds were processed more readily and that some properties were added to the newly conceptual combination after the selection of the general relation in the unitary meaning; (2) as predicted by the CARIN theory [2], the general relation of the modifier was the basis for unifying the relation-based compound; and (3) relation-based compounds were easily adopted by speakers, which suggests that speakers tended to interpret novel compounds through a relation-based approach. For example, people tend to interpret the compound "mountain bird" as "a bird living on a mountain", instead of "a large bird". Nevertheless, the current research found that the selection of compounds for the corpus-based studies might have been influenced by the unbalanced proportion of relation-based and property-based compounds.

Table 1. Examples from Gagne's (2000) Experiment

Property-based		Relation-based	
Similar Nouns	Dissimilar Nouns	High Relation Frequency	Low Relation Frequency
Coat, shirt	Coffee, sword	Plastic toy	Water bird

Gagne and Spalding [6] conducted experiments to discover the effect of the lemma frequency and the family frequency of each constituent's position. They found that both the lemma frequency and the positional family frequency affected the processing of compounds. For example, people's interpretation of the compound "doghouse" was affected by the family members of the structures "dog + __" and "__ + house", but not by the family members of the structures "__ + dog" and "house + __".

To conclude, the CARIN theory proposes that compounds connected via thematic relations between modifiers and head nouns are easier to interpret and comprehend. Past studies [2, 5-6] have supported this proposition. In addition, other factors, including the frequency of the modifiers and the types of head nouns, play an important role in compound processing.

2.2 Property-based Compounds

As proposed by Wisniewski [4] and Wisniewski and Love [7], property-based compounds are interpreted by mapping a property from one word to another word. They found that people mainly interpreted noun-noun compounds via aligning the property with the head noun to acquire the combined meaning, or by extracting a property from one constituent and transferring it to another constituent to obtain the compound. For example, the compound "shark lawyer" should be interpreted as "a lawyer is as truculent as a shark", not as "a lawyer for a shark".

Wisniewski [4] conducted a study to investigate the hypothesis of relation-linking via novel noun-noun compound interpretations and the distribution of relation-based compounds via two interpretation tasks. In addition, Wisniewski (1996) tested whether the higher similarity between constituents within the compound would promote the adoption of a property-based interpretation and found that although the relation-based approach might have facilitated the interpretation of compounds, property-based interpretations were not rare. Wisniewski [4] also suggested that speakers might possess a two-process mechanism for processing compounds. Moreover, the higher similarity between the constituents within the compound might have promoted the adoption of property-based interpretation.

Furthermore, Wisniewski and Love [7] discussed whether speakers first considered the relational information of the constituents and then mapped the property of the modifier to obtain the compound meaning if the semantic relation failed to interpret the meaning. For example, the compound "robin hawk" can be interpreted as "a hawk that preys on robins" or "a small hawk". They found that the higher the similarity between the constituents, the easier it was to interpret property-based compounds.

Wisniewski and Love [7] suggested that the relation-based approach and the property-based approach were both adopted by speakers. Finally, although there was a significant difference between the relation-based approach (70.9%) and the property-based approach (29.1%), this result also revealed that property-based interpretations of compounds occurred frequently, which was different from the findings of the CARIN theory [2].

To conclude, although the relation-based approach to interpreting compounds was dominant in comprehending noun-noun compounds, the property-based approach was often used to interpret compounds when the relation-based approach failed to produce the meanings of compounds.

3. Goals of This Study

How people construct different concepts to form a novel compound word is still debatable. Some have supported the relation-based approach, while others have suggested the property-based approach. One area that has yet to be investigated is whether processing may have something to do with the frequency of compound usages in daily language. While previous studies have focused on the processing of compounds, few studies have investigated the distribution of compounds from the perspective of corpus linguistics. In particular, no studies have determined the distribution of the types of compounds in Mandarin Chinese. Therefore, this research collected compounds from a dictionary and counted the frequencies of the compounds using both relation-based and property-based theories.

Accordingly, our research question is as follows: What is the distribution of relation-based and property-based compounds in Mandarin Chinese? In particular, we would like to know whether relation-based compounds occur more frequently than property-based ones in Mandarin.

4. Methods

The goal of this corpus-based study was to investigate the distribution of relation-based and property-based noun-noun compounds in Mandarin Chinese. Noun-noun combinations were collected from a dictionary and then categorized into relation-based and property-based categories. Finally, the frequency of occurrence of the relation-based and property-based compounds was examined.

4.1 Data Collection and Analysis

This corpus-based study replicated Gagne’s [5] analysis of compounds. Compounds were collected from a Chinese classifier dictionary [8]. The reason for using a classifier dictionary was that the compounds listed in this type of dictionary were more concrete than normal compounds, since compounds in Mandarin Chinese can have abstract meanings. That is, compounds preceded by classifiers were considered evidence that the meaning of these compounds could be either objects or referring to something concrete in the real world.

In addition, noun-noun compounds were selected according to the following four principles. First, one-character nouns preceded by classifiers were excluded, such as 人 *rén* "people", 棋 *qí* "chess", and 玉 *yù* "jade". Second, if the noun-noun compounds could not be segmented into two parts, or the head noun was followed by a modifier, they were removed; for example, in the Chinese compound 雪花 *xuěhuā* "snowflakes", the head noun precedes the modifier. Third, binding words such as 葡萄 *pútáo* "grape" and 蝴蝶 *húdié* "butterfly" and reduplicative words such as 星星 *xīngxīng* "stars" were removed from the data. Fourth, if one constituent of the compound did not belong to the syntactic category of the noun, the compound was deleted from the data; for example, the Chinese phrase 釣魚 *diàoyú* "fish" acts as a verbal phrase in the Chinese compound 釣魚竿 *diàoyúgān* "fishing rod", so it was deleted from the data.

According to these four principles, 880 compounds were collected. Next, a concreteness rating test was conducted to exclude all the compounds that carried abstract meanings. Thirty-two undergraduate students participated in the rating test, which required them to rate compounds as having either abstract meanings or concrete meanings. Compounds were classified as concrete when 75% agreement was reached among the participants. After the rating task, 417 compounds were collected.

The 417 compounds were classified into two word-formation categories: relation-based compounds [2, 5] and property-based compounds [4]. If a Chinese compound could be interpreted by relation, it was placed into the relation-based category. For example, 書桌 *shūzhuō* "desk" is interpreted by the relation FOR, as this is a table for working or for studying. On the other hand, if one or more properties of a constituent were mapped to the other constituent, it was classified as property-based. For example, the Chinese compound 貝殼機 *bèikéjī* "clamshell phone" is a type of mobile phone, and the opening shape (i.e., a functional property) of the noun 貝殼 *bèiké* "shell" is transferred to the other noun to interpret this compound.

Table 2. Eleven Thematic Relations for Noun-Noun Combinations used in this study

Thematic Relations between words	Examples
MAKE	木椅 *mùyǐ* "wooden chairs"
IS	蘭花 *lánhuā* "orchid"
DERIVE	米酒 *mǐjiǔ* "rice wine"
LOCATE	田鼠 *tiánshǔ* "voles"
HAVE	繪本 *huìběn* "picture books"
CAUSE	高山症 *gāoshānzhèng* "altitude sickness"
FOR	窗簾 *chuānglián* "curtain"
USE	兒童椅 *értóngyǐ* "children’s chairs"
ABOUT	山雜誌 *shānzázhì* "mountain magazine"
DURING	冬雨 *dōngyǔ* "winter rain"
BY	學生故事 *xuéshēnggùshì* "student story"

Sixteen property-based compounds were obtained, and of the 401 compounds classified as relation-based, they were further divided into 11 thematic categories (Table 2) according to Gagne's classification [5].

Table 2 shows Chinese examples in the 11 thematic categories, such as the compound

田鼠 *tiánshǔ* "voles", which is classified into the category LOCATE because it is interpreted as "a kind of mouse located in the mountains". Another example is the compound 木椅 *mùyǐ* "wooden chairs", which is a lexical item in thematic relation to the noun (i.e., "chair") modified by what it is made of (i.e., "wood"). Finally, these 401 compounds were further grouped into eight categories. In the following section, the distribution of thematic relations between the constituents in Chinese compounds will be reported.

4.2 Results and Discussion

Of the 417 compounds collected, 401 (96.2%) compounds were classified as relation-based and 16 (3.8%) compounds were classified as property-based. The frequency of occurrence and the corresponding percentages of the 401 relation-based compounds are shown in Table 3. The FOR relation occurred the most frequently in Chinese compounds (37.6%), while the MAKE relation occurred the second most frequently (12.2%), and the BY and CAUSE relations occurred the least frequently.

For the 16 functional property-based interpretations, six categories were discovered, including FARMING AND PLANT, ANIMAL, PHYSICAL APPEARANCE, UNIVERSE, ARTIFACT, and PAPER DOCUMENT. Then, these six categories were placed into one of two domains, the NATURAL KIND domain (i.e., FARMING AND PLANT, ANIMAL, PHYSICAL APPEARANCE, and UNIVERSE), and the ARTIFACT domain (i.e., ARTIFACT and PAPER DOCUMENT; see Table 4). Moreover, findings similar to Wisniewski and Love (1998) were obtained.

The ARTIFACT category occurred the most frequently in Chinese compounds via the property-based strategy, such as 鞭炮 *biānpào* "firecracker". The FARMING AND PLANT category occurred the second most frequently, including 蝴蝶蘭 *húdiélán* "Phalaenopsis orchid", 黃金扁柏 *huángjīn biǎnbǎi* "Oriental arborvitae", 黃金葛 *huáng jīn gě* "centipede tongavine", and 梯田 *tītián* "terraced fields". These findings are in line with those proposed by Wisniewski and Love [7].

Table 3: Distribution of Thematic Relations between Constituents

Thematic Relations	Examples	Frequency (%)
FOR	保險金 *bǎoxiǎnjīn* "insurance claims"	157 (37.6%)
MAKE	鐵門 *tiěmén* "iron gate"	51 (12.2%)
LOCATE	山豬 *shānzhū* "wild boar"	40 (9.5%)
IS	蘭花 *lánhuā* "orchid"	63 (15.1%)
USE	汽車 *qìchē* "automobile"	15 (3.5%)
ABOUT	卡通片 *kǎtōngpiàn* "cartoon film"	17 (4.0%)
HAVE	帆船 *fānchuán* "sailboat"	26 (6.2%)
DURING	年輕人 *niánqīngrén* "youngster"	10 (2.3%)
DERIVED	蜂蜜 *fēngmì* "honey"	15 (3.5%)
BY	中國菜 *zhōngguócài* "Chinese food"	4 (0.9%)
CAUSE	刀傷 *dāoshāng* "wounds by a knife"	3 (0.7%)
Total		401 (100%)

The third most frequently occurring category was PHYSICAL APPEARANCE, including 柳葉眉 *liǔyèméi* "arched eyebrows", 硃砂痣 *zhūshāzhì* "cinnabar mole", and 大花臉 *dàhuāliǎn* "painted face". Moreover, within the top three categories, the properties of SHAPE and COLOR were transferred from one constituent to another. For example, in the compound 蝴蝶蘭 *húdiélán* "Phalaenopsis orchid", the property of SHAPE is transferred from butterflies to describe the shape of the flowers. In another instance, in the compound 黃金葛 *huángjīngě* "centipede tongavine", the property of COLOR is transferred from gold to describe the color of the plant. These results show that the properties of SHAPE and COLOR were adopted as modifiers.

Table 4: Domains and Categories of Property-based Compounds

Domains	Categories	Examples
NATURAL KIND	FARMING AND PLANT	蝴蝶蘭 *húdiélán* "Phalaenopsis orchid"
	ANIMAL	金魚 *jīnyú* "gold fish"
	PHYSICAL APPEARANCE	柳葉眉 *liǔyèméi* "arched eyebrows"
	UNIVERSE	彗星 *huìxīng* "comet"
ARTIFACT	ARTIFACT	原子筆 *yuánzǐbǐ* "ball-point pen"
	PAPER DOCUMENT	黑函 *hēihán* "poison-pen letter"

In addition, the property-based compounds were classified according to the properties of the modifiers. As Table 5 shows, modifiers with the SHAPE property were the most frequently occurring (50%), while the second most frequently occurring was the COLOR property (31.2%), and the METAPHORICAL property occurred the least (18.7%).

Table 5: Distribution of Property-based Modifiers

Properties	Examples	Frequency (%)
COLOR	黃金扁柏 *huángjīn biǎnbǎi* "Oriental arborvitae"	5 (31.2%)
SHAPE	蝴蝶蘭 *húdiélán* "Phalaenopsis orchid"	8 (50.0%)
METAPHORICAL	黑函 *hēihán* "poison-pen letter"	3 (18.7%)
Total		16 (100%)

The corpus results showed that there were very few property-based compounds. This could have resulted from the stimuli collected. Most of the stimuli (i.e., relation-based compounds) analyzed in this study belonged to the ARTIFACT domain, while property-based compounds occurred more often in the NATURAL KIND domain.

5. Conclusion and future work

The goal of this study was to determine the distribution of relation-based and property-based compounds in Mandarin Chinese. Our study intended to determine the frequency of occurrence of relation-based and property-based compounds in Mandarin Chinese. The results showed that relation-based compounds occurred much more frequently than property-based compounds (96.1% vs. 3.8%). Furthermore, it was found that within the relation-based compounds, noun-noun compounds using the FOR relation (e.g., 信紙 *xìnzhǐ* "letter paper") had the highest rates of occurrence (37.6%), while the CAUSE relation (e.g., 刀傷 *dāoshāng* "wounds by a knife") had the lowest rates of occurrence (0.7%). Thus, our study suggests that relation-based word formations were the most commonly found compounds in Mandarin Chinese. Finally, it was found that the property-mapping principles were likely applied to interpret compounds in the ARTIFACT domain.

To answer our research question of "What is the distribution of relation-based and property-based compounds in Mandarin Chinese?", it was found that the frequency of relation-based compounds was much higher than property-based compounds. This distribution is consistent with the prediction of the CARIN theory [2].

Our findings show that there is a tendency that the compounds in the NATURAL KIND domain would likely be interpreted by property-mapping between constituents, while the ones in the ARTIFACT domains would likely be interpreted by relation-linking between constituents. However, our data have insufficient compounds to further analyze this hypothesis. In the future, a study will be conducted in order to collect more compounds from more categories, including ANIMAL, PLANT, and FOOD, to determine whether more property-based compounds occur in the NATURAL KIND than in the ARTIFACT domains.

To conclude, this research has shown that relation-based compounds occurred more frequently than property-based compounds. In addition, the property-based compounds were more likely to occur in the NATURAL KIND domain. These corpus findings suggest that the CARIN theory [2] can better predict the distribution of compound types compared with the dual-process theory [4].

It is hoped that this research will offer more cross-linguistic evidence with which to evaluate the CARIN theory and the dual-process theory regarding compounds. This study has implications for computer processing in dealing with how machines learn to recognize compounds and even novel word combinations. This study has shown that many thematic relations or property-mapping strategies exist in the search for compounds or word combinations, which can help machines to acquire possible thematic relations to interpret novel compounds or concept combinations.

Acknowledgments

This research was supported by grants from the Ministry of Science and Technology (MOST 105-2410-H-415-025) to the first author. We would like to thank the anonymous reviewers for their comments on this work. Remaining errors are our sole responsibility.

References

[1] R. Girju, D. Moldovan, M. Tatu, and D. Antohe, "On the semantics of noun compounds", Computer Speech and Language, vol. 19, pp. 479–496, 2005.

[2] C. L., Gagne and E. J. Shoben, "Influence of Thematic Relations on the Comprehension of Modifier-Noun Combinations". *Experimental Psychology: Learning, Memory, and Cognition,* vol. 23, no. 1, pp. 71-87, 1997.

[3] M. Constant , G. Eryiğit , J.. Monti , L. V. D. Plas , C. Ramisch , M. Rosner and A. Todirascu, "Multiword Expression Processing: A Survey", Computational Linguistics, vol. 43, no. 4, pp. 837-892, 2017.

[4] E. J., Wisniewski, "Construal and Similarity in Conceptual Combination", *Memory and Language*, vol. 35, no. 24, pp. 434-453, 1996.

[5] C. L., Gagne, "Relation-Based Combinations versus Property-Based Combinations: A Test of the CARIN Theory and the Dual-Process Theory of Conceptual Combination", *Memory and Language,* vol. 42, no. pp. 365-389, 2000.

[6] C. L., Gagne and T. L. Spaldin, "Constituent Integration during the Processing of Compound Words: Does It Involve the Use of Relational Structures?" *Memory and Language,* vol. 6*0*, pp. 20-35, 2009.

[7] E. J., Wisniewski and B. C. Love, "Relations versus Properties in Conceptual Combination" *Memory and Language,* vol. 38, pp. 177-202, 1998.

[8] C.-R., Huang, K.-J., Chen, and Q.-X. Lai, *Chinese Classifier Dictionary*. Taipei: Mandarin Daily News, 1996.《黃居仁、陳克健、賴慶雄 (1996)。*常用量詞詞典*。 臺北市：國語日報。》