Computational Linguistics and Chinese Language Processing
Vol. 14, No. 3, September 2009, pp. 305-340

Corpus, Lexicon, and Construction: A Quantitative Corpus Approach to Mandarin Possessive Construction[1]

Cheng-Hsien Chen*

Abstract

Taking Mandarin Possessive Construction (MPC) as an example, the present study investigates the relation between lexicon and constructional schemas in a quantitative corpus linguistic approach. We argue that the wide use of raw frequency distribution in traditional corpus linguistic studies may undermine the validity of the results and reduce the possibility for interdisciplinary communication. Furthermore, several methodological issues in traditional corpus linguistics are discussed. To mitigate the impact of these issues, we utilize phylogenic hierarchical clustering to identify semantic classes of the possessor NPs, thereby reducing the subjectivity in categorization that most traditional corpus linguistic studies suffer from. It is hoped that our rigorous endeavor in methodology may have far-reaching implications for theory in usage-based approaches to language and cognition.

Keywords: Discourse-functional Grammar, Construction Grammar, Quantitative Corpus Linguistics, Possession, Clustering.

1. Introduction

It has been observed that grammatical structures or patterns often serve as routinized formats, fulfilling specific communicative purposes in our daily interaction (Biq, 2001; Chui, 2000; Huang, 2003; Ono & Thompson, 1996; Tao & Thompson, 1994; Thompson & Couper-Kuhlen, 2005; Thompson & Hopper, 2001; Wray, 2002). Speakers' knowledge of their native languages is argued to consist of "a structured inventory of conventional linguistic units, a unit

[1] An earlier version of this paper was presented at The 2008 International Conference on Language, Communication and Cognition, Brighton, UK, 4-7 August. The author is grateful to the audience for fruitful discussion and insightful comments. Sincere gratitude also goes to Stefan Gries, Shuanfan Huang, Chiarung Lu, Iwen Su, Shuchuan Tseng for invaluable advice. Any remaining errors remain the author's responsibility.

* TIGP-CLCLP, Academia Sinica and National Taiwan University
E-mail: alvinworks@gmail.com

being defined in processing terms as a cognitive routine" (Langacker, 1991, p.: 511: 511). In other words, language may provide indicative evidence for our cognitive understanding of the world (Croft, 2001; Fillmore & Atkins, 1992; Fillmore, Kay, & O'Connor, 1988; Grady, 1997; Lakoff, 1993; Lakoff & Johnson, 1980; Tyler & Evans, 2003).

While considerable research has been devoted to a corpus-based approach to constructional schemas, rather little attention has been paid to the methods that are used to further "interpret" the observations. The state of art is that, after surveying the behavioral patterns of a target construction, most cognitive or discourse-functional linguists may still resort to an introspective and intuitive method to identify its sub-patterns. While we do not wish to deny the important role that introspection ultimately plays in the advancement of theorizing, we expect a bottom-up procedure may lend more objectivity, thus credibility, to the empirical results. Therefore, a burgeoning research paradigm - corpus linguistics - now utilizes corpora to investigate the usage patterns and the semantic profiles of these conventional schemas in pursuit of a thorough understanding of our cognitive conceptualization.

A traditional corpus linguistic study on discourse-functional or cognitive grammar often adopts the following approach, as shown in Figure 1 (Biq, 2004a, 2004b, 2004c; Chang, 2002; Chui, 2000; Liu, 2002; Su, 1998, 2004; Tao, 2003b; Wang, Katz, & Chen, 2003).

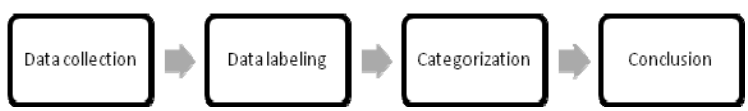

Figure 1. A typical procedure for traditional corpus linguistic studies

Initially, all the target constructions are collected from a corpus (Data collection). Second, all the relevant target constructions are manually labeled according to some researcher-defined features (Data labeling). Third, based on those manually-labeled features, the analyst tries to identify the "types" of these constructions and generate descriptive statistics to obtain a general distribution of the categories identified (Categorization). Finally, conclusions and implications are drawn on the basis of the constructional types of the highest raw frequency counts (Conclusion). It should be noted that such a working pipeline in traditional corpus linguistics has established itself in previous decades as more and more researchers submit to the view that corporal data reflect our grammatical knowledge. Therefore, if one commits him or herself to such a functional view of grammar, one would first collect data from the corpus, label them, categorize them into groups, and make generalizations based on the collected data. We take issue with the detailed procedure of how each step in the pipeline is achieved, namely, how the data is collected, how the data is labeled, and how the data is categorized.

Take Biq's (2004b) study on the patterns of Mandarin stative verb *hao* 'good' for example. In order to find the co-occurrence patterns of *hao*, she first collects all the relevant instances from a 15-hour spoken database and narrows her emphasis to two collocation patterns *hao* + *le* and *hai* + *hao*. Instances containing the target pattern are then further categorized on the basis of her operationally-defined syntactic and pragmatic criteria, and the distribution of these identified types is given in Table 1.

Table 1. The various senses/functions of hao+ le found in conversational data

SENSE/FUNCTION	NUMBER	PERCENTAGE
Topic transition	11	7.10%
Hao = resultative	13	8.30%
Hao = SV	21	13.50%
Conditional	46	29.50%
Recommendatory	65	41.60%
TOTAL	156	100.00%

Finally, conclusions are drawn based on the distribution of the type frequency. While such a traditional corpus linguistic approach has been widely adopted by most discourse-functional and cognitive grammarians, several methodological issues may merit more careful consideration.

In the first step of a traditional corpus linguistic study, the size of the corpus has long been a controversial issue in that small samples may undermine the validity of the results. For instance, even though "Recommendatory" serves as the most frequent type in Biq's observation, 156 tokens may still undermine the credibility of the distribution or even increase the possibility of the by-chance observation. Nevertheless, for most discourse-functional linguists who work on spontaneous speech, the problem of the sample size may appear inevitable due to the enormous manual labor of spoken corpus construction. Given this limitation of the spoken corpus, it is suggested that analysts might as well pursue further statistical analysis so as to increase the confidence level of their numbers. That is, given the maximal recall of the target construction in a corpus of considerable size, it would be theoretically more convincing if the distribution could be statistically tested so as to compensate for the deficiency in small-scale sampling.

With respect to the second step, the features for identifying the types of the target construction are often criticized for being researcher-dependent and lacking basis for cross-analyst comparison. In the case of Biq's study, linguists may differ in how they categorize the pragmatic functions of *hao* + *le*, thus leading to difficulty in comparing different analyst's categorizations of the same construction. Of crucial importance is the third step in a traditional corpus linguistic approach, where only the distribution of the raw

frequency is consulted when conclusions are being drawn. While we acknowledge the fact that frequency plays a crucial role in the formation of our grammatical knowledge (Bybee, 2005; Bybee & Hopper, 2001), we believe that such frequency effects should exclude the by-chance possibilities due to sampling in corpus linguistics. In Table 1, chances are that in daily interaction, recommendatory speech acts may be frequent in general, thus contributing to its higher frequency in the *hao* + *le* co-occurrence patterns. If that is the case, it could be argued that, for all the constructions capable of performing recommendatory acts, this use will eventually emerge as the most frequent type among its pragmatic categorization. In other words, inferential statistics are needed to test whether the recommendatory act is indeed far more frequent than expected. In view of these potential challenges, a quantitative corpus linguistic approach has emerged (c.f. Baayen (2008) for an overview).

In a quantitative corpus linguistic study, the analyst's subjectivity is hoped to be reduced to a minimum. In terms of sampling, (semi-)automatic retrieval of the target pattern is usually adopted to ensure a better recall rate in a large balanced corpus. Second, the features for categorizing the constructional tokens are rigorously *quantified* in an operationally-defined way so that inter-analyst comparison of the results can be easily made. Most crucially, the categorization of the target patterns is made in a bottom-up procedure to replace the analyst's manual efforts as well as subjective factors. Gries and Stefanowitsch (to appear) adopts hierarchical agglomerative cluster analysis to objectively determine semantic classes of constructional sub-patterns. Furthermore, hierarchical cluster analysis has proven itself useful in a wide range of linguistic analyses such as semantic profiles of polysemy (Divjak & Gries, 2006), typology (Croft, 2008), language phylogeny (Atkinson & Gray, 2005; Dunn, Terrill, Reesink, Foley, & Levinson, 2005; McMahon & McMahon, 2003), grammaticization (Hilpert, 2007), and language development (Wiechmann, 2008). Such sophistication in the analytic process facilitates the communication between discourse-functional grammarians of different research paradigms.

The present study, therefore, aims to investigate the interaction between lexicon and construction in a quantitative corpus-based approach. With special focus on a case study of Mandarin Possessive Construction (NP1-*DE*-NP2), this paper addresses one fundamental question for every potential constructional schema: Does this constructional schema have any basic semantic patterns or any other sub-patterns? Specifically, the predictions are: 1) If NP1-*DE*-NP2 Construction has a basic meaning, the NP1-NP2 pairs will yield us such semantic sub-patterns as the major category. 2) If NP1-*DE*-NP2 Construction has *no* basic meaning, the NP1-NP2 pairs will yield us some other heterogeneous sub-patterns, or none. Meanwhile, we will compare the rank-ordering of the raw frequency counts in a traditional corpus linguistic approach with our sophisticated measures to illustrate the potential danger in relying on the former for theorizing.

The rest of the paper is structured as follows. In Section 2, a brief layout of our methodological framework - collostructional analysis - is introduced with a special emphasis on the covarying collexeme analysis. Section 3 will briefly describe the data source and our research methods and demonstrate the inferential statistics used in the evaluation of our data. Results and discussion will be provided in Section 4, illustrating the weaknesses of traditional corpus linguistic studies and the strengths of quantitative corpus linguistic studies. Section 5 concludes this paper with directions for future research and theoretical implications.

2. Lexicon and Construction

While the importance of constructional schemas has come to be the central focus of discourse-functional grammarians (c.f., Croft & Cruse, 2004), it still remains unclear how these constructional analyses can be compared and evaluated given that different linguists resort to different evidence and methods. For instance, some linguists may base their description of the constructional profiles on their own native intuition without quantitative corpus data (Fillmore, *et al.*, 1988; Kay & Fillmore, 1999; Langacker, 2003; Michaelis, 2003; Michaelis & Lambrecht, 1996; Tyler & Evans, 2003). Traditional corpus linguists may take a step further to capitalize on the raw frequency distribution of the words occurring in the target construction (Biq, 2004a; Dancygier & Sweetser, 2000; Goldberg, 1998; Liu, 2002; Su, 2002, 2004; Wang, *et al.*, 2003). Methodologically speaking, little headway has been made in examining the statistical validity of the traditional quantification and little attempt has been made to define an operational method for an analyst to generate the semantic classes of a constructional schema. Occupying the niche, collostructional analysis, proposed by Stefanowitsch and Gries (2003), provides a more rigorous approach to identifying the meaning of a grammatical construction.

Collostructional analysis represents one rigorous corpus-based methodology in discourse-functional linguistics. It makes theoretical commitments to a holistic and symbolic view of linguistic units and, at the same time, bases its quantitative methods on sophisticated statistical analyses. This empirical approach not only flavors the research of usage-based grammar with a more serious emphasis on statistical evaluation but also refreshes the direction of corpus linguistics with a more construction-specific focus on lexico-structural relations. It serves as an umbrella term, referring to research that investigates the correlation/association between words and constructional schemas.

We would like to briefly introduce the terminology and principles in collostructional analysis for the ease of the following exposition. First, lexemes that are attracted to a particular construction are referred to as *collexemes* of the construction. Crucially, collostructional analysis considers the overall distribution of the words in the corpora in calculating the association strength of those words to a specific constructional schema. The

association strength between a collexeme and a construction is measured by submitting all the raw frequency counts of each word in the specific slot of the construction to the Fisher-Yates Exact Test (Pedersen, 1996). Each word occurring in the slot of the construction will be ordered by *collostrength* - defined as the log-transformed *p*-value (to the base of 10) with a positive/negative sign that indicates attraction/repulsion to the construction. This association measure allows a cognitive linguist to probe into human conceptualization through a quantitative study of the relation between words and constructional schemas.

In addition, as constructional schemas often encode a relational meaning, observations on pairs of collexemes in a construction may play an even more crucial role in the identification of the construction semantic profile. Under a usage-based cognitive-linguistic framework, grammatical patterns have been studied in terms of colligations, *i.e.*, linear co-occurrence preferences and restrictions held between specific lexical items and its surrounding syntagmatic contexts (Bybee & Scheibman, 1999; Hunston & Francis, 1999; Scheibman, 2002; Thompson, 2002; Thompson & Hopper, 2001). All of these findings point to the hypothesis that the meaning of one construction relies on the words co-occurring most often with the construction. The assumption behind this reasoning is: a word may occur in a construction if it is semantically compatible with the meaning of the construction (Goldberg, Casenhiser, & Sethuraman, 2004; Stefanowitsch & Gries, 2005). Following this hypothesis, we would expect that, given a construction with two variable slots, observations on the co-occurring patterns in these slots may yield useful empirical evidence for the (semantic) relation encoded by the construction. In this respect, Gries and Stefanowitsch (2004) extend collostructional analysis to covarying collexeme analysis, and seek pairs of collexemes that are statistically attracted to each other within a construction (*i.e.*, *covarying collexemes***)**. Furthermore, Gries and Stefanowitsch (to appear) further adopt a clustering-based approach to identify the potential sub-patterns of covarying collexemes in reflection of the semantic profiles of the target construction.

The present study is by and large compatible with Gries and Stefanowitsch (to appear), to which it is indebted for part of its general outlook, but poses some rather different questions, which we will identify in Section 3. Therefore, in order to investigate the semantic coherence of MPC, a closer look at the correlation between NP1 and NP2 in MPC may present itself as a rewarding endeavor. In Section 3, we will provide a more detailed illustration of our hypotheses and methods.

3. Method

The present study adopts a quantitative corpus-based approach. Initially, the data was collected from the Academia Sinica Balanced Corpus of Mandarin Chinese. This is the major Chinese corpus with detailed parts-of-speech tagging, and it includes a fairly wide range of

genres and styles (mostly formal registers). Instances of Mandarin Possessive Construction (MPC) "NP1 + *DE* + NP2" were automatically retrieved via regular expressions[2]. Retrieval of the constructional instances was done in Java scripts written by the author.

Subsequently, we looked for quantified operationally-defined features to further categorize our MPC tokens. Unlike a traditional corpus linguistic approach, we aimed to reduce the involvement of the analyst's judgment to a minimum. Nevertheless, after collecting instances of the target pattern, traditional corpus linguists usually adopt two types of methods to categorize the target construction. One method is to first formulate the possible semantic categories that the target construction tokens may belong to, then label each token with an appropriate category label. In other words, this approach packages all the categorization process into the analyst's mind and the reader could only see the overall distribution of these predetermined semantic categories. How each target token is categorized into certain semantic category (*i.e.*, the operationally defined criteria) is often obscure to the readers. The other method that a traditional corpus linguistic study may adopt is to formulate a set of researcher-dependent features, usually nominal variables, then manually mark each token with the values of each feature. Then, the analyst categorizes all of the tokens according to the feature values in an introspective fashion. The disadvantage of this approach is obvious. On the one hand, the features tagged for each token usually vary from linguist to linguist and are often categorical and not quantified. On the other hand, even though the features are operationally workable, an introspective way of categorization invites a considerable degree of subjectivity in determining the clusters from the dataset. For instance, the semantic relations can be summarized into 10 labels as in Stefanowitsch (2003) or can be further elaborated into 35 as in Moldovan *et al*. (2004). Different linguists may have different labels and it would be hard to determine if two similar labels are truly semantic equivalents in both analysts' minds, thereby reducing the possibility of comparing the conclusions from different studies. Therefore, both ways of traditional corpus linguistic studies may lead to difficulty in comparing research findings with each other. More challenges to traditional corpus linguistic approach will be elaborated in Section 4.3.

Following Gries and Stefanowitsch (to appear), we adopt a specific type of hierarchical clustering algorithm known as neighbor-joining clustering. A typical process of hierarchical cluster analysis includes: 1) comparing pairwise (dis)similarities between the items in a (dis)similarity matrix via a vector-based representation of the items; 2) successively

[2] Based on the POS tagging principles elaborated in CKIP Technical Report 95-02/98-04, only nouns tagged as Na, Nc, Nd, and Nh were included as our relevant MPC constructional instances. We excluded proper names (Nb), determiners (Ne), classifiers (Nf), and postpositions (Ng). Furthermore, for all the nouns preceding *DE*, we retrieved the rightmost noun as our possessor NP1; for all the nouns following *DE*, we retrieved the first noun tagged with Na, Nc, Nd, or Nh as our possessed NP2.

amalgamating all items into clusters based on the (dis)similarity matrix, which reaches maximal intra-cluster similarity and inter-cluster dissimilarity; 3) visualizing the hierarchical structure of the datasets in the form of a tree-like dendrogram. Specifically, neighbor-joining clustering is often used in phylogeny estimation in biology (Saitou & Nei, 1987), aiming to reconstruct phylogenetic trees from evolutionary distance data under the principle of minimum evolution. Dunn *et al.* (2005) also successfully extends neighbor-joining clustering to the reconstruction of phylogeny in Oceanic languages. Our reason for choosing this algorithm lies in the assumption that constructional semantic profiles evolve similarly to phylogenic evolution in the sense that different semantic patterns of a construction, like different senses of a lexical word, may form a structured polysemy (Goldberg, 1995; Tyler & Evans, 2003). Furthermore, it is suggested that structured polysemy usually emerges from the conventional usage of high frequency via conceptual mechanisms of metaphor and metonymy (Hopper & Traugott, 1993; Traugott & Dasher, 2002; Tyler & Evans, 2003). In other words, semantics of a constructional schema is argued to evolve with language use (Bybee, 1998; Hopper, 1987; Huang, 1998; Tao, 2003a). It is this emergent or evolutionary nature of grammar and semantics that leads us to the decision of adopting phylogenetic clustering in our study. Specifically, in neighbor-joining clustering, not every node on the bottom should be collapsed into one ancestor node. This flexibility allows the possibility that not every sub-pattern comes from one prototypical pattern of the constructional instantiations.

From a perspective of the discourse-functional approach to language, the meaning of a word or a construction is defined by how speakers use it in their daily interaction (Scheibman, 2002; Tao, 2003b; Thompson & Couper-Kuhlen, 2005). In order to look for the semantic coherence encoded by MPC, two possibilities may be pursued: 1) to cluster NP1 based on NP2; 2) to cluster NP2 based on NP1. In the present study, we choose the former approach on a discourse functional basis. It has been observed that the possessor NP in MPC often serves as a topic to which new information encoded by the possessed NP is attached. Therefore, the clustering patterns of the NP1 may shed light on the overall semantic domains of the MPC instance. Furthermore, in the MPC context, the cooccurrence pattern of each NP1 with their NP2 may serve as traces on how speakers frequently make reference to the possessor NPs, thus reflecting the semantic coherence of NP1. That is, a look at how each NP1 is correlated with different types of NP2 in MPC may shed light on their similarity in their references of their possessed entities. If two types of NP1 are correlated with similar types of NP2, they are more inclined to form a semantically coherent class, where their possessed entities are of great similarity. For instance, if in NP1 position, *shi4chang3* 'market' and *chan3pin3* 'product' often co-occur with similar groups of NP2 such as *gong1zuo4* 'job', *xu1qiu2* 'demand', *qing2kuang4* 'condition', *fan3ying4* 'reaction' in MLC, they may easily form a cluster, thus suggesting their similarity in their reference of their possessed entities (*i.e.*, both being

conceptualized as consisting of similar groups of entities). Also, if an abstract entity and a concrete entity are clustered together at an early stage, they may be argued to bear great resemblance in metaphorical conceptualization. Based on these correlation patterns, we can then infer if semantic coherences do exist among different types of NP1. Our working assumption is:

-If MPC has coherent meanings, the NP1 clustering will yield us such semantic sub-patterns as prominent categories at the early stage;
-If MPC has grammaticized as a pure syntactic formative, the NP1 clustering will yield us more heterogeneous sub-patterns, or none.

As clustering approaches are sensitive to the problem of data sparseness and often yield their best results when applied to moderately frequent cases (Kaufman & Rousseeuw, 2005 [1990]), we make a compromise that strikes a balance between the representativeness of the sample and the efficiency of the algorithm. Figure 2 shows the relationship between covarying NP frequency threshold and data preservation percentage. We choose to include 83% of the NP2 by setting a threshold of 5 for the frequency of NP2 and cluster only the top 100 frequent NP1, amounting to 32% of all the NP1. That is, only those covarying NP2 occurring at least 5 times in our original dataset are considered a feature for NP1 in the subsequent vector representation and clustering.

After data filtering, we transform each type of NP1 into vectors based on their association with each covarying collexeme NP2, as tabulated in Table 3. Now that we have a definition for the features or dimensions of each NP1 vector (NP2 of frequency larger than 5), we need measures of association between each NP1 and a given feature (*i.e.*, each type of NP2). It has been observed that cooccurrence raw frequency, as shown in Table 3, is a poor measure of association between a word and a context feature (Jurafsky & Martin, 2008 [2000], p.: 661: 661; Manning & Schütze, 1999, p.: 156: 156). We may require a weighting or measure of association that asks how much more often than chance the feature co-occurs with each type of NP1. Following Gries and Stefanowitsch (to appear), we adopt collostrength from covarying collexeme analysis as our measure of association between each type of NP1 and its covarying NP2 feature.[3]

[3] For further justification for the use of *p*-values as a measure of association strength, please refer to Footnote 6 in Stefanowitsch and Gries (2003). In this analysis, we also tried *t* score as our measure of association, as suggested in Manning and Schütze (1999), and the results were similar to what we had obtained from collostrength measure.

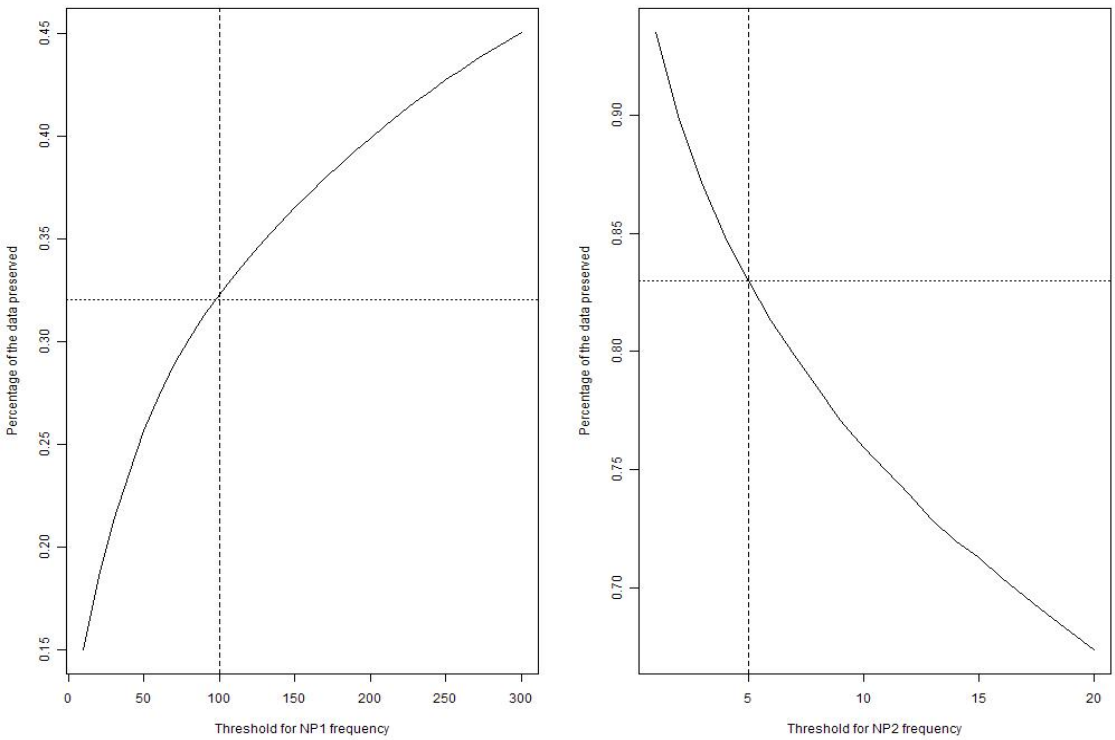

Figure 2. Threshold for NP1 and NP2 and the percentage of data perserved

For example, let us consider the distribution of *zheng4fu3* 'government' and *zheng4ce4* 'policy' in MPC (*i.e.*, *zheng4fu3 DE zheng4xe4* 'the policy of the government') as shown in

Table 2 (parentheses indicate expected frequencies and italics indicate observed frequencies). Applying the Fisher-Yates Exact test to this table yields a *p*-value of 1.11e-59, corresponding to a $p_{\log 10}$-value, *i.e.*, collostrength, of 58.95. This extreme *p*-value indicates that the association between *zheng4fu3* and *zheng4ce4* in MPC is a relatively strong one.

Table 2. The distribution of* zheng4fu3 *and* zheng4ce4 *in Mandarin Possessive Construction

	zheng4ce4	Other NP2	Row Totals
zheng4fu3	40(*1*)	410(*449*)	450
Other NP1	230(*269*)	207829(*207790*)	209059
Column Totals	270	208239	208509

Table 4 shows part of the co-occurrence table with the collostrength of each covarying collexeme pair in the cell. Higher collostrength may suggest a stronger association between

NP1 and NP2.

Table 3. Co-occurrence table of the NP1 (row) with the covarying NP2 (column) in MPC (raw frequency count as assoication measure)

NP1 \ NP2	*ren2* 'man'	*sheng1huo2* 'life'	*xin1* 'heart'	*wen4ti2* 'problem'	*hai2zi5* 'child'	*she4hui4* 'society'	…
ta1 'he'	63	49	49	17	23	6	
wo3 'I'	46	35	152	25	90	2	
zi4ji3 'self'	24	102	26	26	55	8	
ren2 'man'	21	44	40	20	5	6	
ta1 'she'	27	21	39	12	18	1	
wo3men5 'we'	11	55	25	8	48	110	
ni3 'you'	21	25	28	11	38	3	
ta1men5 'they'	11	45	19	8	18	11	
Tai2wan1 'Taiwan'	13	9	3	8	2	18	
…							

Table 4. Co-occurrence table of the NP1 (row) with the covarying NP2 (column) in MPC (collostrength as assoication measures)

NP1 \ NP2	*ren2* 'man'	*sheng1huo2* 'life'	*xin1* 'heart'	*wen4ti2* 'problem'	*hai2zi5* 'child'	*she4hui4* 'society'	…
ta1 'he'	0.248008	0.063422	1.211847	1.18E-05	0.006771	0	
wo3 'I'	0.040219	0.006473	Inf	0.02988	Inf	0	
zi4ji3 'self'	6.10E-07	Inf	0.045449	0.064862	5.841638	6.70E-07	
ren2 'man'	0.019297	3.341083	5.632644	0.548443	0.000205	0.000827	
ta1 'she'	0.171652	0.071817	5.323306	0.033664	0.550925	8.00E-08	
wo3men5 'we'	3.46E-05	7.69897	1.44626	0.003065	Inf	Inf	
ni3 'you'	0.187166	0.858615	3.448672	0.110468	Inf	0.00024	
ta1men5 'they'	0.003388	7.522879	1.313479	0.037517	1.585381	0.298076	
Tai2wan1 'Taiwan'	0.03479	0.008238	0.000383	0.07051	0.000245	2.061541	
…							

Next, we compute pairwise distance matrix among these 100 types of NP1. As summarized in Jurafsky and Martin (2008 [2000]: 663-667), correlation similarity measures are more prone to detect and to use curvature of vectors in multidimensional space; these measures may work better for word similarity in information/document retrieval as compared to distance dissimilarity measures. Moreover, according to Manning and Schütze (1999: 299), among all the distance-based measures, the cosine is the most frequently-used measure in the comparison of semantic similarity (*c.f.*, Curran (2004)).[4] Therefore, we compute a pairwise cosine distance matrix and submit this matrix to neighbor-joining clustering. The statistical evaluation is computed in R scripts written by the author, using the *ape* package developed by Paradis (2004).

Furthermore, we compare the lists ordered by raw frequency and collostrength, respectively, by submitting these two rank-orderings to Friedman's rank test for correlated samples. This test is the nonparametric analogue of the one-way repeated-measures ANOVA, often being applied to test if two rank-orderings differ significantly. By so doing, we demonstrate the degree to which raw frequency overlaps with the collostrength, thus highlighting the potential danger in relying on the raw frequency in theorizing.

Before the discussion of the results, let us briefly turn to the question of why we chose Mandarin Possessive Construction as our pilot study. Even though we name this construction as "possessive" here, its constructional meaning is not as uncontroversial as the naming suggests. The reason for choosing this as our target construction is mainly due to the cross-linguistic complexity of possessive or genitive constructions (Baron, Herslund, & Sorensen, 2001; Dong, 2003; Heine, 2001; Lyons, 1986; Nikiforidou, 1991; Stefanowitsch, 2003; Taylor, 1996). As implied in its alias as "associative phrases" (Li & Thompson, 1981), MPC has been notorious for its encoding of diversified semantic relations between two NPs to the extent that Li and Thompson (1981) even argue that "the precise meaning…is determined entirely by the meanings of the two noun phrases involved". While a typical possessive construction may encode a semantic relation of "possession", including ownership, kinship, and component-part relations (Nikiforidou, 1991; Stefanowitsch, 2003; Taylor, 1996), it is still unclear whether MPC indeed exhibits semantic coherences in its distributional patterns, or should better be analyzed as a semantic-general syntactic formative. Hopefully, the empirical evidence from the covarying collexemes may help solve this controversial issue.

---

[4] Curran (2004) evaluated a wide range of similarity measures by comparing the results with gold-standard thesauri and concluded that Dice and Jaccard methods perform best as measures of vector similarity. As a result, we also computed the similarity matrix based on these methods and submitted them to hierarchical clustering. The results were by and large similar to what we had obtained from the cosine similarity measure. Therefore, we shall base our discussion on the results from the cosine similarity measure.

4. Results and Discussion

208509 tokens of MPC were extracted from the Academia Sinica Corpus of Mandarin Chinese. These MPC instances consist of 26005 types of NP1 and 25987 types of NP2, amounting to 159645 types of NP1-NP2 pairs, *i.e.*, covarying collexemes. Each distinct type of NP1-NP2 was submitted to statistical evaluations, and the results are as follows.

4.1 Semantic Coherence in MPC

We cluster NP1 according to its covarying NP2 in MPC. After filtering our infrequent NP1 and NP2 types, we cluster the most frequent 100 NP1 according to their covarying NP2 of frequency larger than 5. This boils down to a 100 by 4372 contingency table with 58470 tokens of MPC in total, as shown earlier in Table 3. All of the possessor NPs (NP1) are then transformed into vector representations based on their collostrength with each type of covarying collexeme NP2, as shown previously in Table 4. The cooccurrence measures between NP1 and NP2 serve as criteria for classification of the NP1. We then compute the cosine distance between each pair of NP1 and submit the distance matrix to neighbor-joining clustering to obtain a tree-like representation of the NP1 categorization.

In a tree size with 100 tips (*i.e.*, 100 types of NP1), the information that is supposed to be summarized is likely to be no longer visible. Therefore, instead of plotting out the whole tree, obscuring the clustering information that is sought, we choose to plot only a portion of the full dendrogram at a time, while indicating its context - how it relates to the rest of the tree. In the following illustration, the original dendrogram is divided into three parts, where the whole tree is plotted on the left and the subtree on the right. The location of the subtree is indicated with the color on the whole tree.

The results from Figure 3 to Figure 5 are moderately revealing in that several coherent semantic frames in NP1 emerge from the dendrograms. Specifically, 7 semantically coherent categories emerge from the amalgamative process: Human, Time, Country, Enterprise, Culture, Knowledge, and Institution. Based on these correlation patterns, we suggest that semantic coherences do exist among different types of NP1, supporting the claim that MPC has not fully grammaticized as a pure syntactic formative.

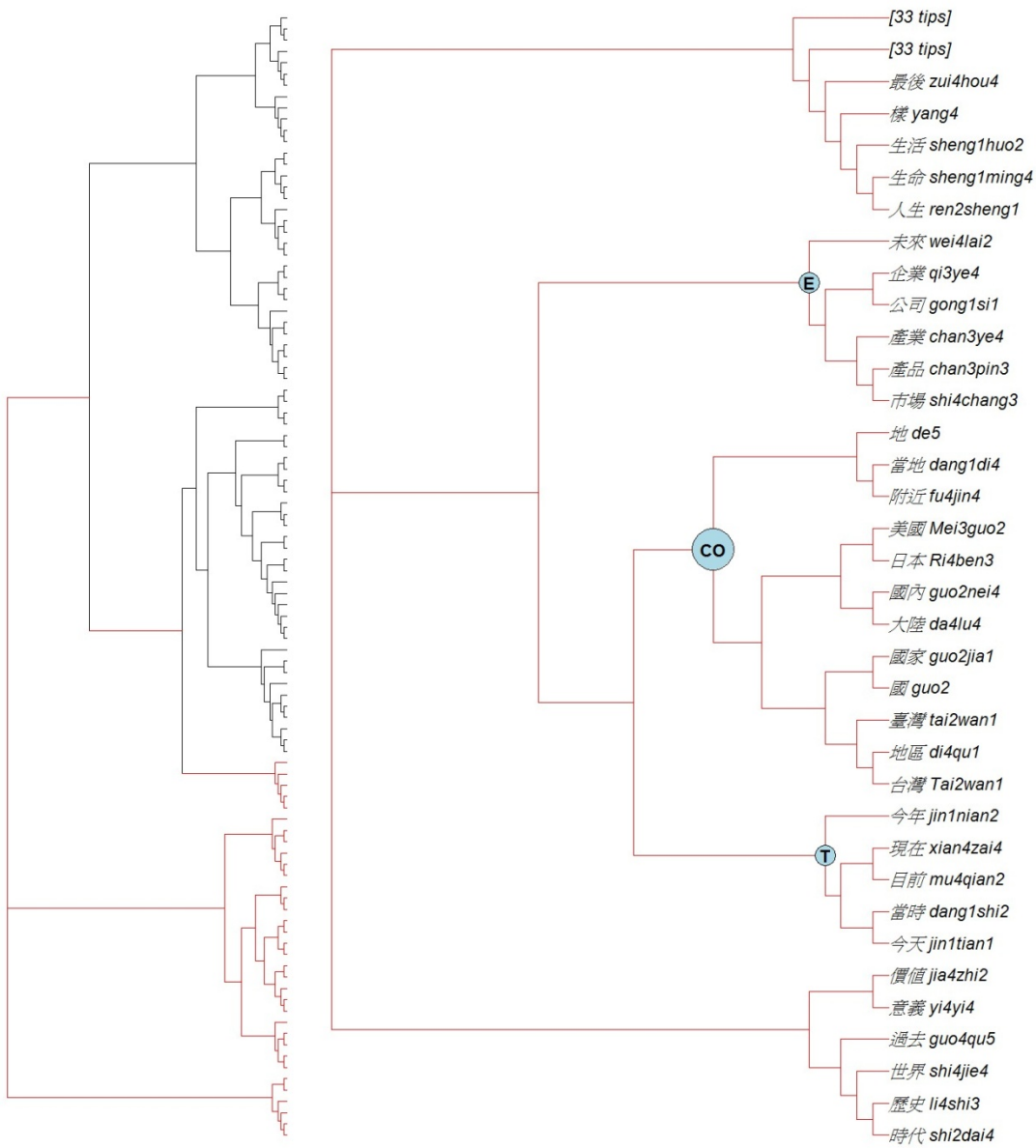

Figure 3. Subtree one of the dendrogram

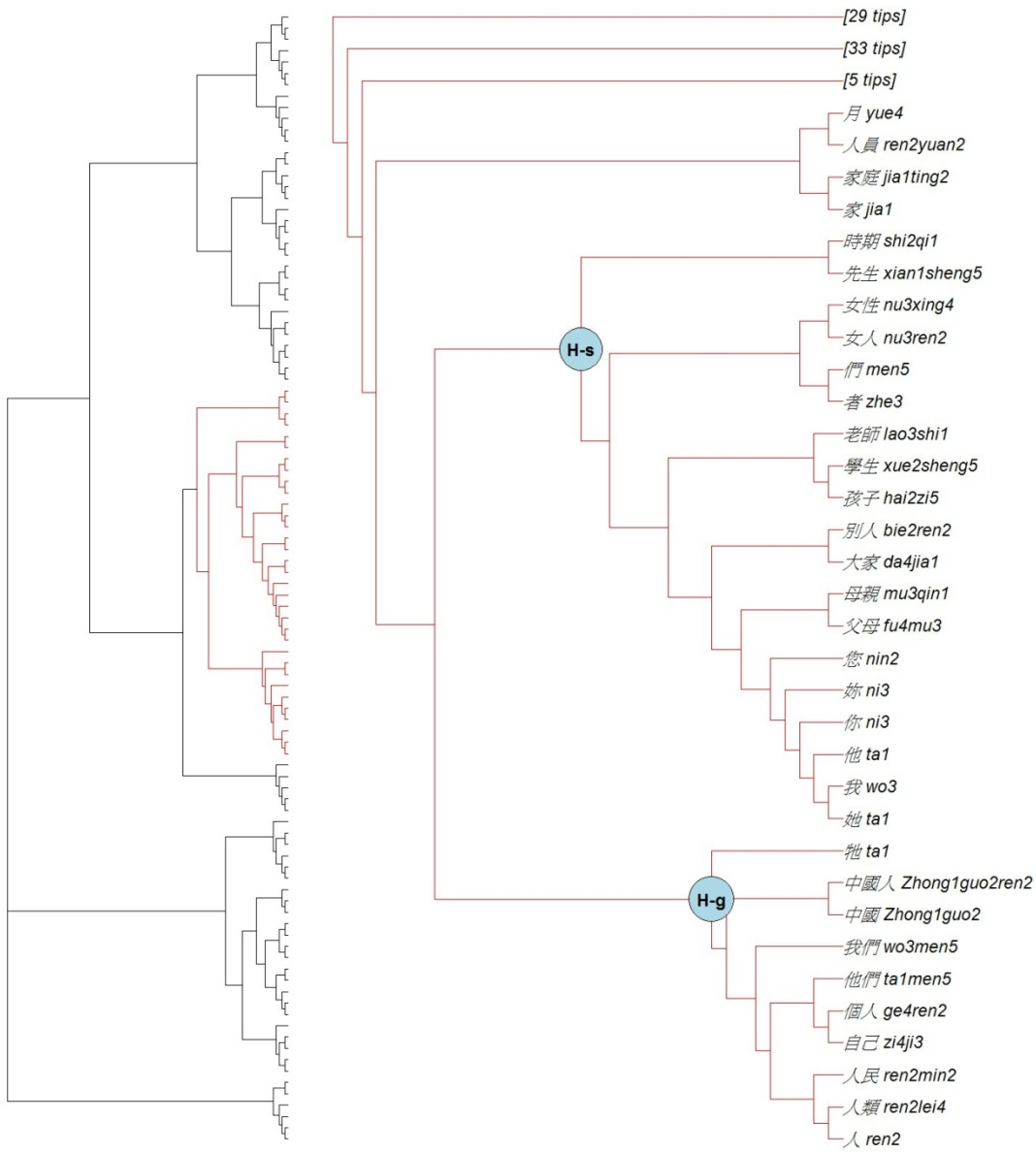

Figure 4. Subtree two of the dendrogram

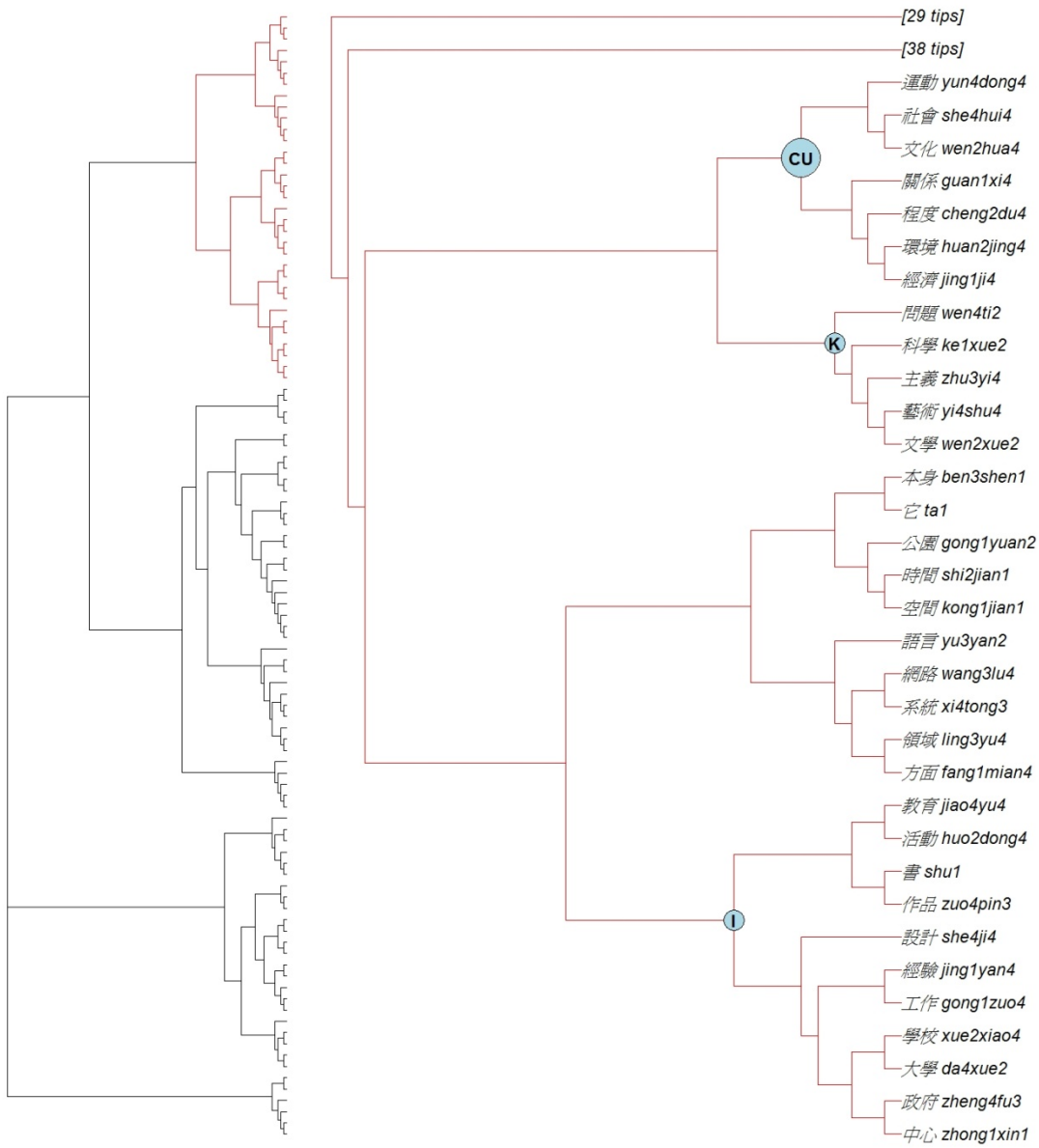

Figure 5. Subtree three of the dendrogram

Figure 6 shows the distribution of the significant covarying pairs in each semantic frame. Among all the significant covarying collexemes, about 35 percent of the NP1 falls into the HUMAN semantic frame. A further Chi-square test suggests that the distribution of these covarying pairs in different semantic frames is significant ($\chi^2_{(8)} = 862.25$, $p < 0.01$). Furthermore, among the semantic frames we identify, HUMAN presents itself as the most

concrete category. This may suggest that the HUMAN frame serves as a basis for the metaphorical extension of the possessive relations encoded by MPC and that other semantic frames may be argued to derive from this basis through cognitive mechanisms.

In the semantic classes generated, however, there are still quite a range of covarying pairs that are difficult to label with appropriate semantic categories (*i.e.*, OTHER in Figure 6). Nearly one-fifth of the NP1s do not yield coherent clustering patterns at the early stage of the dendrogram. While these clusters generated in the dendrogram may not be suggestive in reaching a coherent semantic category, they are revealing in the respect that they show how one entity is conceptualized similarly to another under the context of a possessive relation. On the top of the subtree in Figure 3, it is observed that *sheng1huo2*, *sheng1ming4*, and *ren2sheng1* are often portrayed as the "end point" (*zui4hou4*) in discussing their possessed properties. Similarly, the bottom of the subtree in Figure 3 shows that the properties of the abstract entities such as world, history, times, value, and meaning are often cast in the past background as those abstract NPs (*shi4jie4*, *li4shi3*, *shi2dai4*, *jia4zhi2*, *yi4yi4*) are clustered together with *guo4qu4*. Furthermore, in the middle of the subtree in Figure 4, it is suggested that time and space is conceptualized as one coherent domain as *shi2jian1* and *kong1jian1* are clustered together at the early stage. Of similar nature is the grouping of *yu3yan2* with *wang3lu4* and *xi4tong3*, suggesting that native speakers often conceptualize the Internet and language in a similar fashion. Instead of being a blow to the credibility of our clustering method, these cases may serve as *prima facie* evidence for the degree of grammaticization in MPC toward becoming a pure "associative" syntactic formative. This paradox should not come as a counter-expectation at all to discourse-functional grammarians as the more frequent a construction gets used the more its semantics gets bleached (Hopper & Traugott, 1993; Traugott & Dasher, 2002). Yet, compared with the other 76% of the semantically coherent clusters, this small portion of the heterogamous patterns may not necessarily stop us from claiming that MPC indeed has semantic coherence in its usage.

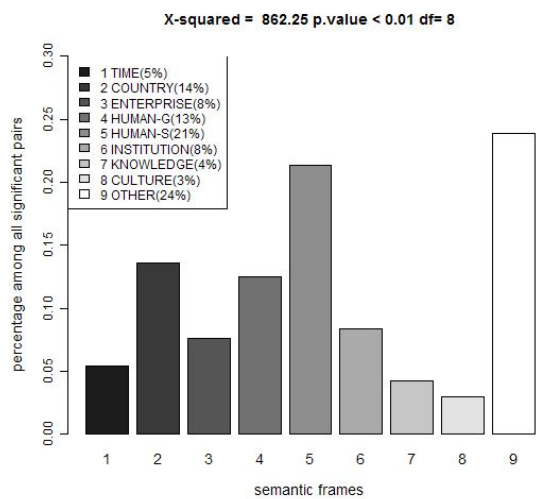

Figure 6. distribution of the significant covaryign collexemes in different semantic frames

Although the clusters are automatically yielded by the algorithm, what each cluster represents still relies on the analyst's manual labeling, thus drawing criticism that such endeavors are still introspective and subjective. Nevertheless, it should be noted that the distribution in Figure 6 differs greatly from the raw frequency distribution used in traditional corpus linguistic studies. First of all, collostrength, rather than raw frequency, is used to reduce the possibility of making a by-chance observation. Second, even though the label for each semantic category may be analyst-dependent, the members of each cluster are objectively generated by the quantified features and a sophisticated algorithm. When adopting the same algorithm on the same dataset, different quantitative corpus linguists will obtain the same clustering results, although their labels for those semantic frames may differ. This advantage provides the possibility for research on the same construction to compare their conclusions and theoretical implications.

On the one hand, we still need a more objective way to decide what kind of semantic relations are maintained in each semantic frame. In the current stage, synsets in WordNet provide a promising possibility for an automatic identification of semantic relations (c.f., Moldovan & Badulescu, 2005). The present study only provides a coarse-grained categorization for the semantic domains of the possessor NPs. With a semantically disambiguated and syntactically parsed corpus such as WordNet, we could conduct the covarying collexeme analysis on a "synset," rather than "word," basis. Furthermore, clustering possessor NP1 according to possessed NP2 (or the other way around) will not provide us a clear picture of the semantic relations encoded by MPC. To automatically identify such semantic relations between NP1 and NP2, we need to cluster the whole MPC according to its covarying lexemes/constructions, such as the coocurring predicates.

On the other hand, the labeling of the semantic frame for the clusters generated may be expected to proceed automatically in the near future by making reference to the hypernyms in Chinese WordNet as well. For instance, for all the NP1s that are clustered together, we can generate a list of their hyponyms for each sense of the NP1 in WordNet and look for potential higher-order semantic domains among all these NP1s. A sophisticated extension to the synonyms of these NP1s may also facilitate the search for a common superordinate domain. In other words, the analyst's subjectivity may be reduced to the minimum once Chinese Wordnet is available. Also, it should be noted that cluster analysis here is not intended to completely substitute for manual classification (or in any sense bearing absolute superiority over the latter). Instead, the goal here is to show that, in order to introduce findings and observations from discourse-functional linguistics into the modeling of natural language processing, an automatic constructional sense induction may be needed for efficient implementation.

4.2 A Closer Look at each Semantic Frames

Before we start to look at some examples from each cluster generated, we would like to emphasize that the labels of semantic relations in the following discussion are mainly for exposition[5]. Furthermore, we leave for future consideration whether it is feasible to reach a consensus among discourse-functional grammarians regarding a unanimous set of semantic relations (See 4.3 below for more discussion). Rather, these brief sketches of the covarying collexemes in each cluster are to support our claim that the clusters generated by the neighbor-joining algorithm are indeed semantically coherent.

First of all, two types of HUMAN frames - specific and generic - can be clearly identified in Figure 4. One consists of personal pronouns while the other includes noun phrases mostly referring to the generic idea of "people" or "human beings". The former semantic frame, dubbed as HUMAN-specific, demonstrates prototypical "ownership" (*e.g.*, *ta1 DE xiao3shuo1*), "component-whole" (*e.g.*, *ta1 DE shou3*) as well as "interpersonal relation" (*e.g.*, *ta1 DE zhang4fu5*) relation between NP1 and NP2 and the covarying collexemes of higher collostrength are included in (1)[6]. In the latter, the HUMAN-generic frame, NP2 often refers to the key components of a human life or human beings in general, thus maintaining a component-whole relation with the NP1. Typical examples are included in (2).

(1) HUMAN - specific (H-s)

她 ta1 'she'	丈夫 zhang4fu5 'husband'
她 ta1 'she'	女兒 nu3er2 'daughter'
他 ta1 'he'	小說 xiao3shuo1 'novel'
他 ta1 'he'	太太 tai4tai5 'wife'
我 wo3 'I'	心 xin1 'heart'
我 wo3 'I'	心情 xin1qing2 'mood'
他 ta1 'he'	手 shou3 'hand'
我 wo3 'I'	日記 ri4ji4 'diary'

[5] Our semantic relations are based on a more complete list of semantic relations proposed by Moldovan *et al.* (2004).

[6] The covarying collexemes listed as examples here are all of significant collostrength ($p < 0.01$).

(2) Humans - generic (H-g)

人 ren2 'man' 一生 yi1sheng1 'all one's life'
人 ren2 'man' 天性 tian1xing4 'nature'
自己 zi4ji3 'self' 生命 sheng1ming4 'life'
人民 ren2min2 'people' 生活 sheng1huo2 'life'
個人 ge4ren2 'individual' 自由 zi4you2 'freedom'
我們 wo3men5 'we' 社會 she4hui4 'society'
我們 wo3men5 'we' 孩子 hai2zi5 'child'
我們 wo3men5 'we' 祖先 zu3xian1 'ancestor'
自己 zi4ji3 'self' 家 jia1 'home'
人類 ren2lei4 'humanity' 理性 li3xing4 'sense'
人 ren2 'man' 尊嚴 zun1yan2 'dignity'

In Figure 3, three semantic frames are identified: TIME, COUNTRY, and ENTERPRISE. Typical significant covarying collexemes in the TIME frame are included in (3). The purpose of this Time frame appears to contextually "position" the NP2 within a specific temporal space denoted by NP1. Therefore, it can be observed that the prominent semantic relation is attribute-holder between NP1 and NP2.

(3) TIME (T)

當時 dang1shi2 'then' 心情 xin1qing2 'mood'
今天 jin1tian1 'today' 主題 zhu3ti2 'theme'
當時 dang1shi2 'then' 台灣 Tai2wan1 'Taiwan'
現在 xian4zai4 'modern' 年輕人 nian2qing1ren2 'young people'
目前 mu4qian2 'at the present time' 狀況 zhuang4kuang4 'condition'

For the COUNTRY frame, significant covarying collexemes are listed in (4). The components of a country are clearly shown in the covarying collexemes of this category as component-whole relation appears to be a dominant semantic relation in this semantic frame. Quite a range of fundamental components of a country manifest clearly, from concrete entities like *min2zhong4* or *ren2min2* to more abstract assets such as *zheng4zhi4*, *jing1ji4*, *wen2hua4*, and *fa3lu4*. As far as the purpose of the present study is concerned, this COUNTRY frame may be argued to exhibit a metaphorical conceptualization, where a basic possessive relation -

component-whole - is extended to a higher abstract level of political entities. Also, the prominence of this semantic category may reflect the nature of the material collected in Academia Sinica Corpus as local news accounts for the majority of the data sources.

(4) COUNTRY (CO)

地區 di4qu1 'area' 人民 ren2min2 '(the) people'
國家 guo2jia1 'country' 人民 ren2min2 '(the) people'
台灣 Tai2wan1 'Taiwan' 主權 zhu3quan2 'sovereignty'
台灣 Tai2wan1 'Taiwan' 民主 min2zhu3 'democracy'
當地 dang1di4 'local' 民俗 ming2shu2 'customs'
地區 di4qu1 'area' 民眾 min2zhong4 'people'
當地 dang1di4 'local' 居民 ju1min2 'resident'
國家 guo2jia1 'country' 法律 fa3lu:4 'law'
台灣 Tai2wan1 'Taiwan' 政治 zheng4zhi4 'politics'
美國 Mei3guo2 'America' 軍事 jun1shi4 'military affairs'
台灣 Tai2wan1 'Taiwan' 原住民 yuan2zhu4min2 'indigenous peoples'
大陸 da4lu4 'mainland' 煤 mei2 'coal'
大陸 da4lu4 'mainland' 經濟 jing1ji4 'economy'
日本 Ri4ben3 'Japan' 經濟 jing1ji4 'economy'

Let us now consider the ENTERPRISE frame, as illustrated in Figure 3. There is quite a bit noise in this group, where NP1 and NP2 may hold an ownership relation (*gong1si1 DE lao3ban3*), or producer-product (*gong1si1 DE chan3pin3*) and some other typical behaviors or expectations of a social institution (*shi4chang3 DE gong1xu1* and *shi4chang3 DE jing4zheng1*). Nonetheless, this may suffice as to argue that an ENTERPRISE frame is emergent from our daily uses of MPC as all these possessor NPs (NP1) bear great resemblance in reference with their possessed entities (NP2). Furthermore, the amalgamation of *wei4lai2* with this ENTERPRISE cluster may suggest that in the discourse context these enterprises are often cast as futuristic entities in that possibilities and potentials are more emphasized.

(5) ENTERPRISE (E)

未來 wei4lai2 'future' 方向 fang1xiang4 'direction'
未來 wei4lai2 'future' 走向 zou3xiang4 'trend'
未來 wei4lai2 'future' 主人翁 zhu3ren2weng1 'master (of one's own destiny, etc.)'
市場 shi4chang3 'market' 主流 zhu3liu2 '(n) main stream of a fluid'
公司 gong1si1 '(business) company' 老闆 lao3ban3 'boss'
產品 chan3pin3 'goods' 行銷 xing2xiau1 'marketing'
市場 shi4chang3 'market' 佔有率 zhan4yau3lu4 'percentage of coverage'
市場 shi4chang3 'market' 供需 gong1xu1 'supply and demand'
公司 gong1si1 '(business) company' 股東 gu3dong1 'stockholder'
產品 chan3pin3 'goods' 品質 pin3zhi4 'quality'
公司 gong1si1 '(business) company' 董事長 dong3shi4zhang3 'chairman of the board'
市場 shi4chang3 'market' 競爭 jing4zheng1 'to compete'

In the subtree of Figure 5, three more semantic frames are identified: CULTURE, KNOWLEDGE, and INSTITUTION. Typical examples of the first frame are illustrated in (6). The NP1 in the CULTURE frame often refers to the products out of our socialization, such as *she4hui4*, *wen2hua4*, and *yun4dong4*. Typical cases of a component-whole relation in this frame may include *she4hui4 DE cheng2yuan2*, *yun4dong4 DE chuan4shi3ren2*, or *wen2hua4 DE ren2qun2*. Of particular interest here is that most possessive relations maintained between these covarying collexemes are also deemed metaphorical in the sense that the possessor and the possessed refer to abstract social entities, rather than concrete animate subjects.

(6) CULTURE (CU)

社會 she4hui4 'society' 成員 cheng2yuan2 'member'
運動 yun4dong4 'movement' 創始人 chuang4shi3ren2 ‘founder'
文化 wen2hua4 'culture' 人群 ren2qun2 'a crowd'
社會 she4hui4 'society' 良心 liang2xin1 'conscience'
社會 she4hui4 'society' 現象 xian4xiang4 'appearance'
文化 wen2hua4 'culture' 精髓 jing1sui3 'marrow'
文化 wen2hua4 'culture' 影響 ying3xiang3 'influence'
文化 wen2hua4 'culture' 差異 cha1yi4 'difference'
文化 wen2hua4 'culture' 產物 chan3wu4 'product'

Another semantic frame identified in Figure 5 - KNOWLEDGE - illustrates possessive relations in a variety of knowledge-based domains, such as *ke1xue2*, *zhu3yi4*, *yi4shu4,* and *wen2xue2*. Of particular interest here is the inclusion of *wen4ti2* into this cluster. In other words, the former disciplines such as *ke1xue2*, *zhu3yi4*, and *yi4shu4* may be argued to behave similarly to *wen4ti2* under the context of making references to their possessed entities (*i.e.*, NP2). This amalgamated pattern may suggest that the other disciplines in this KNOWLEDGE frame are often viewed as a question to which we quest for a possible solution or answer.

(7) KNOWLEDGE (K)

科學 ke1xue2 'science' 方法 fang1fa3 'method'
問題 wen4ti2 'problem' 方法 fang1fa3 'method'
主義 zhu3yi4 'creed' 色彩 se4cai3 'tint'
藝術 yi4shu4 'art' 形式 xing2shi4 'form'
問題 wen4ti2 'problem' 時候 shi2hou5 'time'
藝術 yi4shu4 'art' 創作 chuang4zuo4 'to create'
問題 wen4ti2 'problem' 答案 da2an4 'answer'
問題 wen4ti2 'problem' 辦法 ban4fa3 'means'
問題 wen4ti2 'problem' 關鍵 guan1jian4 'crucial'
問題 wen4ti2 'problem' 癥結 zheng1jie2 'bottleneck'
文學 wen2xue2 'literature' 性格 xing4ge2 'nature'
科學 ke1xue2 'science' 知識 zhi1shi5 'intellectual'

The final semantic frame - INSTITUTION - refers to goal-oriented social formations, ranging from concrete entities like *xue2xiao4*, *da4xue2*, and *shu1*, to more abstract ones like *zhong1xin1*, *huo2dong4*, and *jiao4yu4*. In terms of basic possessive relations, NP2 in this frame often consists of the components of NP1 such as *zhong1xin1 DE ren2yuan2*, *xue2xiao4 DE lao3shi1*, *shu1 DE zuo2zhe3*, *zheng4fu3 DE fa3ling4*, *da4xue2 DE xiao4zhang3*, and *xue2xiao4 DE she4bei4*. Nonetheless, a look at the NP2 shared by the NP1 in this frame suggests the goal-oriented nature of this category, as in *zheng4fu3 DE zhu3zhang1*, *jiao4yu4 DE mu4di4, hau2dong4 DE mu4di4,* and *shu1 DE zhu3zhi3*. More examples are listed in (8).

(8) INSTITUTION (I)

中心 zhong1xin1 'center' 人員 ren2yuan2 'staff'
學校 xue2xiao4 'school' 老師 lao3shi1 'teacher'
書 shu1 'book' 作者 zuo2zhe3 'author'
政府 zheng4fu3 'government' 法令 fa3ling4 'decree'
大學 da4xue2 'university' 校長 xiao4zhang3 'president'
學校 xue2xiao4 'school' 設備 she4bei4 'equipment'
活動 huo2dong4 'activity' 內容 nei4rong2 'content'
書 shu1 'book' 內容 nei4rong2 'content'
教育 jiao4yu4 'to educate' 內容 nei4rong2 'content'
政府 zheng4fu3 'government' 主張 zhu3zhang1 'to advocate'
活動 huo2dong4 'activity' 目的 mu4di4 'purpose'
教育 jiao4yu4 'to educate' 目的 mu4di4 'purpose'
政府 zheng4fu3 'government' 決策 jue2ce4 'decision'
書 shu1 'book' 主旨 zhu3zhi3 '(n) gist'

4.3 Raw Frequency and Collostrength

As the rank-ordering of the raw frequency has been greatly utilized in the literature of traditional corpus linguistic studies, we would now like to express some issues with the validity of this approach. In order to examine the relationship between the raw frequency (*i.e.*, the counts of the covarying collexemes in our collected sample) and the collostrength (*i.e.*, the association strength of the covarying collexemes with each other in the construction), we compare the ordering of these two measures for the most frequent N covarying collexemes. The procedure is as follows. First, the most frequent N covarying collexemes are selected and their corresponding raw frequency and collostrength are submitted to Friedman's rank test to see if the rank-ordering of the raw frequency and the collostrength differs significantly among these top frequent N cases. The results are shown in Table 5.

Table 5. The p-values from Friedman's rank test and Kendall's τ coefficient for the ordering of raw frequency and collostrength among the top frequent N covarying collexemes

For top frequent N covarying collexemes	Friedman test *p*-value	Kendall's τ
3	0.083265	1
4	0.0455	0.666667
5	0.025347	0.4

6	0.014306	0.466667
7	0.008151	0.52381
8	0.004678	0.357143
9	0.0027	0.055556
10	0.001565	0.022222
11	0.000911	0.163636
12	0.003892	0.090909
13	0.002282	0.102564
14	0.001341	0.230769
15	0.000789	0.314286
16	0.000465	0.383333
17	0.000275	0.441176
18	0.000162	0.503268
19	9.60E-05	0.54386
20	5.70E-05	0.463158

Table 5 illustrates the correlation between the raw frequency and the collostrength for the most frequent N covarying collexemes. The second column lists the *p*-value from Friedman test and the third column gives the Kendall's τ coefficient as the degree of correspondence between the two rankings of raw frequency and collostrength. As can be seen, while raw frequency may have explanatory power in the topmost frequent cases, the rank ordering itself may be legitimately applied only to the most frequent cases (N < 7). Starting from the most frequent 7 covarying collexemes, the rank-ordering of the raw frequency differs significantly from that of the collostrength ($\chi^2_{(1)} = 7$, p-value < 0.01). Furthermore, Kendall's τ coefficient shows the association strength of the rankings between raw frequency and the collostrength weakens with the inclusion of more covarying collexeme types. In other words, a study based on the most frequent 6 covarying collexemes may yield the same conclusions as one based on the covarying collexemes of the top 6 collostrength. Nonetheless, a study based on more than 6 covarying collexemes is likely to yield somewhat different patterns from one based on a more statistically sophisticated measure, *i.e.*, collostrength. Whether the index for rank-ordering is statistically sophisticated may be trivial for the most frequent few cases. Yet, as far as the majority of the covarying collexemes are concerned, the statistical sophistication of the rank-ordering index is non-trivial and crucial in drawing conclusions. Nevertheless, what most traditional corpus-based studies do is to base their theorizing on the rank ordering of the raw frequency in all cases, which in our view may seriously undermine the validity of such corpus-based endeavor. Therefore, we suggest that a certain level of sophistication is

needed in the use of the raw frequency in traditional corpus-based studies.[7]

Let us now take a closer look at the differences between the ordering of raw frequency and that of collostrength. Table 6 shows the top 20 covarying collexemes sorted by their raw frequency in a descending order. If an analyst bases a study on the ordering of the raw frequency, they may easily reach the conclusion that the possessors in MPC overwhelmingly fall into the human category. Nevertheless, the high frequency of the covarying pairs in table 6 may derive from the fact that those NP1 are indeed words of high frequency in the overall corpora. If the frequency of the NP1 is high, the pairs containing NP1 are expected to be higher. In other words, the significance of the high constructional frequency may be diminished by the frequency of its parts. Most importantly, it remains unclear whether the observed frequency is significantly higher than the expected.

Table 6. A list of covarying collexemes ranked by their respective raw frequency

NP1	NP2	N	Collostrength
我 wo3 'I'	心 xin1 'heart'	152	101.9261
我們 wo3men5 'we'	社會 she4hui4 'society'	110	81.6629
自己 zi4ji3 'self'	生活 sheng1huo2 'life'	102	27.86079
他 ta1 'he'	作品 zuo4pin3 'works (of art)'	100	40.82442
我 wo3 'I'	孩子 hai2zi5 'child'	90	41.54748
我 wo3 'I'	手 shou3 'hand'	78	40.03977
他 ta1 'he'	話 hua4 'dialect'	77	28.52847
自己 zi4ji3 'self'	身體 shen1ti3 '(human) body'	77	46.91627
人 ren2 'man'	生命 sheng1ming4 'life'	72	49.52095
她 ta1 'she'	手 shou3 'hand'	67	46.39019
自己 zi4ji3 'self'	生命 sheng1ming4 'life'	66	28.51615
月 yue4 'moon'	時間 shi2jian1 'time'	66	94.69845
我 wo3 'I'	朋友 peng2you5 'friend'	65	28.84368
他 ta1 'he'	人 ren2 'man'	63	1.48E-05
他 ta1 'he'	朋友 peng2you5 'friend'	61	21.59611
他 ta1 'he'	手 shou3 'hand'	58	19.41221

[7] For a thorough review of statistical measures of association, please refer to Chapter 20 in Jurafsky and Martin (2008 [2000]).

自己 zi4ji3 'self'	孩子 hai2zi5 'child'	55	17.00112
自己 zi4ji3 'self'	能力 neng2li4 'ability'	55	16.34834
我們 wo3men5 'we'	生活 sheng1huo2 'living'	55	15.38543
我 wo3 'I'	意思 yi4si5 'idea'	51	28.90842
我 wo3 'I'	話 hua4 'dialect'	51	14.85332
類型 lei4xing2 'type'	人 ren2 'man'	51	51.11585
他 ta1 'he'	心 xin1 'heart'	49	9.045286
他 ta1 'he'	生活 sheng1huo2 'life'	49	1.698984
方面 fang1mian4 'respect'	問題 wen4ti2 'problem'	48	29.34759
我們 wo3men5 'we'	孩子 hai2zi5 'child'	48	23.34222
我 wo3 'I'	眼睛 yan3jing1 'eye'	47	24.80487
我 wo3 'I'	人 ren2 'man'	46	1.54E-06
他 ta1 'he'	父親 fu4qin1 'father'	46	25.52748
你 ni3 'you'	忠告 zhong1gao4 'advice'	45	79.65025

Even though we have adopted collostrength of the covarying collexemes as a reference or approximation to their association to the construction, we still do not know what kind of semantic relations MPC encodes most often. A traditional corpus linguist may proceed to label the semantic relations between NP1 and NP2 manually. In order to demonstrate a traditional corpus linguistic approach, we take the top 20 covarying collexeme pairs as an illustration. Table 7 shows the top 20 covarying collexeme pairs that are significantly attracted to each other in MPC. The list is ranked according to their collostrength in a descending order. In the rightmost column, we manually label these significant covarying collexemes with possible semantic profiles, *i.e.*, a semantic relation between NP1 and NP2. Our labels for the semantic relations in Table 7 are purely descriptive, as stated in Section4.2; no theoretical significance is attached to the precise labels used to characterize the semantic relations.

Table 7. A list of covarying collexemes ranked by their respective collostrength

NP1	NP2	N	Collostrength	Semantic relation
不久 bu4jiu3 'not long (after)'	將來 jiang1lai2 'future'	35	107.7124	Idiom
我 wo3 'I'	心 xin1 'heart'	152	101.9261	Component-Whole
月 yue4 'moon'	時間 shi2jian1 'time'	66	94.69845	Attribute-Holder
我們 wo3men5 'we'	社會 she4hui4 'society'	110	81.6629	ownership
你 ni3 'you'	忠告 zhong1gao4 'advice'	45	79.65025	Participant-Event
政府 zheng4fu3 'government'	政策 zheng4ce4 'policy'	40	59.58043	Participant-Event
魔王 mo2wang2 'fiend'	左手 zuo3shou3 'left-hand'	14	51.66222	Component-Whole
類型 lei4xing2 'type'	人 ren2 'man'	51	51.11585	Attribute-Holder
人 ren2 'man'	生命 sheng1ming4 'life'	72	49.52095	ownership
最後 zui4hou4 'final'	獵人 lie4ren2 'hunter'	19	49.42944	idiom
媒體 mei2ti3 'media'	報導 bao4dao3 'coverage'	24	49.00789	Participant-Event
自己 zi4ji3 'self'	身體 shen1ti3 '(human) body'	77	46.91627	Component-Whole
她 ta1 'she'	手 shou3 'hand'	67	46.39019	Component-Whole
問題 wen4ti2 'problem'	癥結 zheng1jie2 'bottleneck'	20	42.58891	Component-Whole
龍 long2 'dragon'	傳人 chuan2zen2 'heir'	12	41.89946	idiom
我 wo3 'I'	孩子 hai2zi5 'child'	90	41.54748	Interpersonal relations
因素 yin1su4 'element'	影響 ying3xiang3 'influence'	22	41.51503	Participant-Event
他 ta1 'he'	作品 zuo4pin3 'works'	100	40.82442	ownership
我 wo3 'I'	手 shou3 'hand'	78	40.03977	Component-whole
生命 sheng1ming4 'life	意義 yi4yi4 'meaning'	37	40.03224	Attribute-

(force)'				Holder
學者 xue2zhe3 'scholar'	社區 she4qu1 'community'	16	38.85216	ownership
異樣 yi4yang4 'discrimination'	眼光 yan3guang1 'judgment'	14	38.56742	Attribute-Holder
國王 guo2wang2 'king'	新衣 xin1yi1 'new clothes'	11	38.10723	idiom
動詞 dong4ci2 'verb'	論元 lun4yuan2 'argument'	11	38.03616	Component-Whole
人 ren2 'man'	一生 yi1sheng1 'all one's life'	35	37.90407	Participant-Event
挫折 cuo4zhe2 'setback'	時候 shi2hou5 'time'	22	36.72331	Time-Event
瘤子 liu2zi5 'lump'	老公公 lao3gong1gong1 'old man'	9	36.0871	Attribute-Holder
他 ta1 'he'	妻子 qi1zi5 'wife'	44	35.55265	Interpersonal Relations
方面 fang1mian4 'respect'	知識 zhi1shi5 'knowledge'	30	34.97364	Attribute-Holder
用戶 yong4hu4 'user'	需求 xu1qiu2 'requirement'	19	34.68546	Participant-Event

In Table 7, several covarying collexemes of low frequency do jump out as prominent instances of MPC, such as *mo2wang2 DE zuo3shou3*, *long2 DE chuan2zen2*, *guo2wang2 DE xin1yi1*, and *dong4ci2 DE lun4yuan2*. These significant pairs are not only indicative of the constructional semantic profiles but also suggestive of the topics covered in the corpora. Crucially, these phrases would not have emerged on the analyst's list if one had adopted only raw frequency as their measure of association.

Interpretable as it may seem, even the results based on the ordering of the collostrength still raise several methodological issues. Although we have flavored a traditional corpus linguistic approach with a quantitative nature using collostrength, such a traditional approach still needs to face the fact that a predetermined list of semantic relations is needed in order to label all the covarying pairs. It comes as no surprise that our labeling for the semantic relations in Table 7 may draw adverse criticism from researchers of a different paradigm. Linguists differ greatly in the number of possible semantic relations encoded by possessive constructions and different linguists may adopt different terms. For instance, the semantic relations can be summarized into 10 labels as in Stefanowitsch (2003) or can be further elaborated into 35 as in Moldovan *et al.* (2004). Furthermore, while a small sample of the significant covarying collexemes may be indicative of the basic semantic profiles of MPC, there is still a potential drawback. We choose the top 20 covarying collexemes only for

demonstration of a traditional corpus linguistic approach. A traditional corpus linguistic study could have chosen the top 200, 2000, or even 20000. In other words, to the size of the sample from the ordering list may have great impact on the validity of the results. As long as a traditional corpus linguist likes to investigate all the semantic relations between NP1 and NP2 in MPC, they are bound to face these potential challenges. Most importantly, it would be difficult for them to bypass the issue of how to classify all the MPC tokens in an objective way. While a wedge of cheese like the top 20 (or more) covarying collexemes may be suggestive for the semantic coherence of MPC, a step further can be made to include more data so as to generate the semantic coherence of MPC in a more objective fashion. This is exactly the niche we are trying to occupy.

5. Concluding Remarks

Based on our empirical investigation, the overall results suggest that Mandarin Possessive Construction does exhibit a considerable degree of semantic coherence that holds between covarying collexemes, and the relative consistency among different sets of covarying collexemes. In addition, we further ensure the objectivity in identifying semantic classes of the possessor NPs by submitting a sample of covarying collexemes into phylogenic hierarchical clustering. The generated dendrogram appears to support the claim that semantic coherence does hold between covarying collexemes of the construction in question and NP1 exhibits several clear semantic classes where possessive relations are often contextualized, namely, HUMAN, COUNTRY, ENTERPRISE, INSTITUTION, KNOWLEDGE, and CULTURE. Nevertheless, some of the clusters have failed to manifest a coherent category of their own. While the prominent semantic frames identified may explain why most linguists still recognize this construction as a possessive construction in Mandarin, these heterogeneous clusters may account for the fact that some would describe it as a pure contextually-driven formative for any possible association. Therefore, noise in our results may serve as preliminary evidence for its degree of grammaticization toward becoming a pure "associative" syntactic formative.

The purpose of the present study should be clear. While construction grammar has emerged as one of the dominant theoretical frameworks in the usage-based research paradigm, its insights may be further supported by more quantitative empirical data. It is argued that covarying collexeme analyses may serve as a compelling approach in identifying constructional sub-patterns, thus lending more credibility to the empirical results. Also, various statistical tools may not only facilitate the difficult task of categorization for the analysts but reduce the subjectivity of the judgment to the minimum as well.

Furthermore, with more and more quantitative methods being incorporated into linguistic studies, these findings are more likely to be taken seriously by other interdisciplinary scholars.

Differences in methodology only widen the gap for the possible interdisciplinary interaction and comparison. Crucially, while other disciplines like biology, psychology, and cognitive science have long been viewing classification as a quantitative problem and have been using computer programs to identify a best parsimonious tree from an unorganized dataset, it would be less advantageous for traditional linguists to opt for an intuition-based approach where classifications are acceptable as long as scholars of the same research paradigm agree that they are acceptable. Even though traditional corpus linguistics has made a step further in contributing a great deal to the linguistic theorizing in general, such an approach does not typically produce data which are interpretable and usable by neighboring disciplines, especially in natural language processing. While other disciplines provide results based on rigorous quantitative design, they would hardly buy the story of linguists who generate conclusions via purely descriptive statistics. Therefore, a more rigorous quantitative method may serve as an objective platform where more interdisciplinary dialogue on human cognition can be made. While discourse-functional and cognitive linguists are sifting the wheat from the chaff in the massive harvest of corpus data, it is hoped that such rigorous emphasis on methodology may lend more objectivity and credibility to their revealing insights.

References

Atkinson, Q. D., & Gray, R. D. (2005). Curious parallels and curious connections: Phylogenetic thinking in biology and historical linguistics. *Systematic Biology,* 54(4), 513-526.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baron, I., Herslund, M., & Sorensen, F. (2001). *Dimensions of possession*. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Biq, Y.-O. (2001). The grammaticalization of *Jiushi* and *Jiushishou* in Mandarin Chinese. *Concentric: Studies in English Literature and Linguistics,* 27(2), 53-74.

Biq, Y.-O. (2004a). Construction, reanalysis, and stance: 'V *yi ge* N' and variations in Mandarin Chinese. *Journal of Pragmatics,* 36, 1655-1672.

Biq, Y.-O. (2004b). From collocation to idiomatic expression: The grammaticalization of *hao* phrases/constructions in Mandarin Chinese. *Journal of Chinese Language and Computing,* 14(2), 73-95.

Biq, Y.-O. (2004c). People, things, and stuff: general nouns in spoken Mandarin. *Concentric: Studies in Linguistics,* 30(1), 41-64.

Bybee, J. (1998). The emergent lexicon. *Chicago Linguistic Society,* 34, 421-435.

Bybee, J. (2005). Mechanisms of change in grammaticization: The role of frequency. In B. D. Joseph & R. D. Janda (Eds.), *The Handbook of Historical Linguistics* (pp. 602-623). Malden, MA: Blackwell Publication.

Bybee, J., & Hopper, P. J. (2001). *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins.

Bybee, J., & Scheibman, J. (1999). The effect of usage on degree of constituency: the reduction of *don't* in American English. *Linguistics, 37*, 575-596.

Chang, M.-H. (2002). Discourse functions of *Anne* in Taiwanese Southern Min. *Concentric: Studies in English Literature and Linguistics,* 28(2), 85-115.

Chui, K. (2000). Morphologization of the degree adverb *HEN*. *Language and Linguistics,* 1(1), 45-59.

Croft, W. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Croft, W. (2008). Evolutionary Linguistics. *Annual Review of Anthropology,* 37(1), 219-234.

Croft, W., & Cruse, D. A. (2004). *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Curran, J. R. (2004). *From distributional to semantic similarity.* Unpublished dissertation, University of Edinburgh, Edinburgh, UK.

Dancygier, B., & Sweetser, E. E. (2000). Constructions with *if*, *since* and *because*: Causality, epistemic stance, and clause order. In E. Couper-Kuhlen & B. Kortmann (Eds.), *Cause, condition, concession, contrast: Cognitive and discourse perspectives* (pp. 111-142). Berlin: Mouton de Gruyter.

Divjak, D. S., & Gries, S. T. (2006). Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory,* 2(1), 23-60.

Dong, C.-R. (2003). A cognitive account of possessive construction. *Foreign Languages and Their Teaching,* 169, 60-63.

Dunn, M., Terrill, A., Reesink, G., Foley, R. A., & Levinson, S. C. (2005). Structural phylogenetics and the reconstruction of ancient language history. *Science,* 309(5743), 2072-2075.

Fillmore, C. J., & Atkins, B. T. (1992). Toward a frame-based lexicon: The semantics of *risk* and its neighbors. In A. Lehrer & E. F. Kittay (Eds.), *Frames, Fields, and Contrasts* (pp. 75-102). Hillsdale, NJ: Lawrence.

Fillmore, C. J., Kay, P., & O'Connor, M. K. (1988). Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language,* 64, 501-538.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar approach to argument structure*. Chicago: Chicago University Press.

Goldberg, A. E. (1998). The emergence of the semantics of argument structure constructions. In B. MacWhinney (Ed.), *The emergence of language* (pp. 197-212). Hillsdale, NJ.: Lawrence Erlbaum Associates.

Goldberg, A. E., Casenhiser, D., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics,* 15, 286-316.

Grady, J. (1997). *Foundations of meaning: Primary metaphors and primary scenes.* Unpublished dissertation, University of California Berkeley, Berkeley, CA.

Gries, S. T., & Stefanowitsch, A. (2004). Co-varying collexemes in the *into*-causative. In M. Achard & S. Kemmer (Eds.), *Language, Culture and Mind* (pp. 225-236). Stanford, CA: CSLI Publications.

Gries, S. T., & Stefanowitsch, A. (to appear). Cluster analysis and the identification of collexeme classes. In J. Newman & S. Rice (Eds.), *Experimental and empirical methods in the study of conceptual structure, discourse, and language* (pp. 73-90). Stanford, CA: CSLI Publications (Available at: http://www.linguistics.ucsb.edu/faculty/stgries/research/ClusteringCollexemes.pdf).

Heine, B. (2001). Ways of explaining possession. In I. Baron, M. Herslund & F. Sorensen (Eds.), *Dimensions of possession* (pp. 311-328). Amsterdam and Philadelphia, PA.: John Benjamins.

Hilpert, M. (2007). *Germanic Future Constructions: A Usage-based Approach Grammaticalization.* Unpublished dissertation, Rice University, Houston, TX.

Hopper, P. J. (1987). Emergent grammar. *Berkeley Linguistics Society,* 13, 139-157.

Hopper, P. J., & Traugott, E. C. (1993). *Grammaticalization*. Cambridge: Cambridge University Press.

Huang, S. (1998). Emergent lexical semantics. In S. Huang (Ed.), *Selected papers from the second international symposium on languages in Taiwan* (pp. 129-150). Taipei: Crane.

Huang, S. (2003). Doubts about complementation: A functionalist analysis. *Language and Linguistics,* 4(2), 429-455.

Hunston, S., & Francis, G. (1999). *Pattern Grammar. A corpus-driven approach to the lexical grammar of English*. Amsterdam and Philadelphia, PA.: John Benjamins.

Jurafsky, D., & Martin, J. H. (2008 [2000]). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd edn ed.). Upper Saddle River, NJ: Prentice Hall.

Kaufman, L., & Rousseeuw, P. J. (2005 [1990]). *Finding groups in data: An introduction to cluster analysis* (2nd edn ed.). Hoboken, NJ: Wiley.

Kay, P., & Fillmore, C. J. (1999). Grammatical constructions and linguistic generalizations: The *What's X doing Y?* construction. *Language,* 75(1), 1-34.

Lakoff, G. (1993). The contemporary theory of metaphor. In A. Ortony (Ed.), *Metaphor and Thought* (2nd edn ed., pp. 202-251). Cambridge: Cambridge University Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Langacker, R. W. (1991). *Foundations of cognitive grammar: Descriptive application* (Vol. 2). Stanford, CA: Stanford University Press.

Langacker, R. W. (2003). Constructions in cognitive grammar. *English Linguistics,* 20, 41-83.

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: A Functional Reference Grammar*. Berkeley and Los Angeles: University of California Press.

Liu, M.-C. (2002). *Mandarin Verbal Semantics: A Corpus-based Approach*. Taipei: Crane Publishing Co.

Lyons, C. (1986). The syntax of English genitive constructions. *Journal of Linguistics,* 22(1), 123-143.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

McMahon, A., & McMahon, R. (2003). Finding Families: Quantitative Methods in Language Classification. *Transactions of the Philological Society,* 101(1), 7-55.

Michaelis, L. A. (2003). Word meaning, sentence meaning, and syntactic meaning. In H. Cuykens, R. Dirven & J. R. Taylor (Eds.), *Cognitive approaches to lexical semantics* (pp. 163-209). Berlin and New York: Mouton de Gruyter.

Michaelis, L. A., & Lambrecht, K. (1996). Toward a construction-based model of language function: The case of nominal extraposition. *Language,* 72, 215-247.

Moldovan, D., & Badulescu, A. (2005, October). *A semantic scattering model for the automatic interpretation of genitives.* Paper presented at the Human language technology conference and conference on empirical methods in natural language processing, Vancouver.

Moldovan, D., Badulescu, A., Tatu, M., Antohe, D., & Girju, R. (2004, 6 May). *Models for the semantic classification of noun phrases.* Paper presented at the HLT-NAACL Worksop on Computational Lexical Semantics, Boston, MA.

Nikiforidou, K. (1991). The Meanings of the Genitive: A Case Study in Semantic Structure and Semantic Change. *Cognitive Linguistics,* 2(2), 149-205.

Ono, T., & Thompson, S. A. (1996). Interaction and Syntax in the Structure of Conversational Discourse: Collaboration, Overlap, and Syntactic Dissociation. In E. H. Hovy & D. R. Scott (Eds.), *Computational and Conversational Discourse: Burning Issues, an Interdisciplinary Account* (pp. 67-96). Heidelberg: Springer-Verlag.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics,* 20(2), 289-290.

Pedersen, T. (1996). Fishing for exactness. *Proceedings of the SCSUG 96 in Austin, TX*, 188-200.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution,* 4(4), 406-425.

Scheibman, J. (2002). *Point of View and Grammar: Structural patterns of subjectivity in American English conversation*. Amsterdam: John Benjamins Publishing Company.

Stefanowitsch, A. (2003). Constructional semantics as a limit to grammatical alternation: The two genitives of English. In G. Rohdenburg & B. Mohndorf (Eds.), *Determinants of Grammatical Variation in English*. Berlin and New York: Mouton de Gruyter.

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics,* 8(2), 209-243.

Stefanowitsch, A., & Gries, S. T. (2005). Covarying collexemes. *Corpus Linguistics and Linguistic Theory,* 1(1), 1-43.

Su, L. I.-w. (1998). Conversation coherence: the use of *ranhou* in Chinese spoken discourse *Collected Papers of the Second Interactional Symposium on Languages in Taiwan* (pp. 167-182). Taipei: Crane.

Su, L. I.-w. (2002). Why a construction - That is the question! *Concentric: Studies in English Literature and Linguistics,* 28(2), 27-42.

Su, L. I.-w. (2004). Subjectification and the use of the complementizer *SHUO*. *Concentric: Studies in Linguistics,* 30(1), 19-40.

Tao, H. (2003a). Toward an emergent view of lexical semantics. *Language and Linguistics,* 4(4), 837-856.

Tao, H. (2003b). A usage-based approach to argument structure: 'remember' and 'forget' in spoken English. *International Journal of Corpus Linguistics,* 8(1), 75-95.

Tao, H., & Thompson, S. A. (1994). The discourse and grammar interface: Preferred clause structure in Mandarin conversation. *Journal of the Chinese Language Teachers Association,* 29(3), 1-34.

Taylor, J. R. (1996). *Possessives in English: An Exploration in Cognitive Grammar*. Oxford: Oxford University Press.

Thompson, S. A. (2002). "Object complements" and conversation: towards a realistic account. *Studies in Language,* 26(1), 125-164.

Thompson, S. A., & Couper-Kuhlen, E. (2005). The clause as a locus of grammar and interaction. *Discourse Studies,* 7(4-5), 481-506.

Thompson, S. A., & Hopper, P. J. (2001). Transitivity, clause structure, and argument structure: Evidence from conversation. In J. Bybee & P. J. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 27-60). Amsterdam and Philadelphia: John Benjamins.

Traugott, E. C., & Dasher, R. B. (2002). *Regularity in semantic change*. Cambridge: Cambridge University Press.

Tyler, A., & Evans, V. (2003). *The Semantics of English Prepositions: Spatial Scenes, Embodied Meaning, and Cognition*. Cambridge: Cambridge University Press.

Wang, Y.-F., Katz, A., & Chen, C.-H. (2003). Thinking as saying: *shuo* ('say') in Taiwan Mandarin conversation and BBS talk. *Language Sciences,* 25(5), 457-488.

Wiechmann, D. (2008). On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory,* 4(2), 253-290.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.