Annotating Chinese Collocations with Multi Information

Ruifeng Xu[1], Qin Lu[1], Kam-Fai Wong[2], Wenjie Li[1]

[1] Department of Computing,
The Hong Kong Polytechnic University,
Kowloon, Hong Kong

[2] Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong,
N.T., Hong Kong

{csrfxu,csluqin,cswjli}@comp.polyu.edu.hk kfwong@se.cuhk.edu.hk

Abstract

This paper presents the design and construction of an annotated Chinese collocation bank as the resource to support systematic research on Chinese collocations. With the help of computational tools, the *bi*-gram and *n*-gram collocations corresponding to 3,643 headwords are manually identified. Furthermore, annotations for *bi*-gram collocations include dependency relation, chunking relation and classification of collocation types. Currently, the collocation bank annotated 23,581 *bi*-gram collocations and 2,752 *n*-gram collocations extracted from a 5-million-word corpus. Through statistical analysis on the collocation bank, some characteristics of Chinese *bi*-gram collocations are examined which is essential to collocation research, especially for Chinese.

1 Introduction

Collocation is a lexical phenomenon in which two or more words are habitually combined and commonly used in a language to express certain semantic meaning. For example, in Chinese, people will say *历史-包袱* (*historical baggage*) rather than *历史 -行李* (*historical luggage*) even though *包袱* (*baggage*) and *行李*(*luggage*) are synonymous. However, no one can argue why *历史* must collocate with *包袱*. Briefly speaking, collocations are frequently used word combinations. The collocated words always have syntactic or semantic relations but they cannot be generated directly by syntactic or semantic rules. Collocation can bring out different meanings a word can carry and it plays an indispensable role in expressing the most appropriate meaning in a given context. Consequently, collocation knowledge is widely employed in natural language processing tasks such as word sense disambiguation, machine translation, information retrieval and natural language generation (Manning et al. 1999).

Although the importance of collocation is well known, it is difficult to compile a complete collocation dictionary. There are some existing corpus linguistic researches on automatic extraction of collocations from electronic text (Smadja 1993; Lin 1998; Xu and Lu 2006). These techniques are mainly based on statistical techniques and syntactic analysis. However, the performances of automatic collocation extraction systems are not satisfactory (Pecina 2005). A problem is that collocations are word combinations that co-occur within a short context, but not all such co-occurrences are true collocations. Further examinations is needed to filter out pseudo-collocations once co-occurred word pairs are identified. A collocation bank with true collocations annotated is naturally an indispensable resource for collocation research. (Kosho et al. 2000) presented their works of collocation annotation on Japanese text. Also, the Turkish treebank, (Bedin 2003) included collocation annotation as one step in its annotation. These two collocation banks provided collocation identification and co-occurrence verification information. (Tutin 2005) used shallow analysis based on finite state transducers and lexicon-grammar to identify and annotate collocations in a French corpus. This collocation bank further provided the lexical functions of the collocations. However to this day, there is no reported Chinese collocation bank available.

Proceedings of the Linguistic Annotation Workshop, pages 61–68,
Prague, June 2007.

In this paper, we present the design and construction of a Chinese collocation bank (acronymed *CCB*). This is the first attempt to build a large-scale Chinese collocation bank as a Chinese NLP resource with multiple linguistic information for each collocation including: (1) annotating the collocated words for each given headword; (2) distinguishing *n*-gram and *bi*-gram collocations for the headword; (3) for *bi*-gram collocations, *CCB* provides their syntactic dependencies, chunking relation and classification of collocation types which is proposed by (Xu and Lu 2006). In addition, we introduce the quality assurance mechanism used for *CCB*. *CCB* currently contains for 3,643 common headwords taken from "*The Dictionary of Modern Chinese Collocations*" (Mei 1999) with 23,581 unique *bi*-gram collocations and 2,752 unique *n*-gram collocations extracted from a five-million-word segmented and chunked Chinese corpus (Xu and Lu, 2005).

The rest of this paper is organized as follows. Section 2 presents some basic concepts. Section 3 describes the annotation guideline. Section 4 describes the practical issues in the annotation process including corpus preparation, headword preparation, annotation flow, and the quality assurance mechanism. Section 5 gives current status of *CCB* and characteristics analysis of the annotated collocations. Section 6 concludes this paper.

2 Basic Concepts

Although collocations are habitual expressions in natural language use and they can be easily understood by people, a precise definition of collocation is still far-reaching (Manning et al. 1999). In this study, we define a *collocation* as *a recurrent and conventional expression of two or more content words that holds syntactic and semantic relation*. Content words in Chinese include noun, verb, adjective, adverb, determiner, directional word, and gerund. Collocations with only two words are called *bi*-gram collocations and others are called *n*-gram collocations.

From a linguistic view point, collocations have a number of characteristics. Firstly, collocations are *recurrent* as they are of habitual use. Collocations occur frequently in similar contexts and they appear in certain fixed patterns. However, they cannot be described by the same set of syntactic or semantic rules. Secondly, free word combinations which can be generated by linguistic rules are normally considered compositional. In contrast, collocations should be *limited compositional* (Manning et al. 1999) and they usually carry additional meanings when used as a collocation. Thirdly, collocations are also *limited substitutable* and *limited modifiable*. Limited substitutable here means that a word cannot be freely substituted by other words with similar linguistic functions in the same context such as synonyms. Also, many collocations cannot be modified freely by adding modifiers or through grammatical transformations. Lastly, collocations are *domain-dependent* (Smadja 1993) and language-dependent.

3 Annotation Guideline Design

The guideline firstly determines the annotation strategy.

(1) The annotation of *CCB* follows the headword-driven strategy. The annotation uses selected headwords as the starting point. In each circle, the collocations corresponding to one headword are annotated. Headword-driven strategy makes a more efficient annotation as it is helpful to estimate and compare the relevant collocations.

(2) *CCB* is manually annotated with the help of automatic estimation of computational features, i.e. semi-automatic software tools are used to generate parsing and chunking candidates and to estimate the classification features. These data are present to the annotators for determination. The use of assistive tools is helpful to produce accurate annotations with efficiency.

The guideline also specifies the information to be annotated and the labels used in the annotation.

For a given headword, *CCB* annotates both *bi*-gram collocations and *n*-gram collocations. Considering the fact that *n*-gram collocations consisting of continuous significant *bi*-grams as a whole and, the *n*-gram annotation is based on the identification and verification of *bi*-gram word combinations and is prior to the annotation of *bi*-gram collocations.

For *bi*-gram annotation, which is the major interest in collocation research, three kinds of information are annotated. The first one is the syntactic dependency of the headword and its co-word in a *bi*-gram collocation . A syntactic dependency normally consists of one word as the governor (or *head*), a dependency type and another word serves as dependent (or *modifier*) (Lin 1998).Totally, 10 types of dependencies are annotated in *CCB*. They are listed in Table 1 below.

	Dependency Description	Example
ADA	Adjective and its adverbial modifier	极其/d 惨痛/a *greatly painful*
ADV	Predicate and its adverbial modifier in which the predicate serves as head	沉重/ad 打击/v *heavily strike*
AN	Noun and its adjective modifier	合法/a 收入/n *lawful incoming*
CMP	Predicate and its complement in which the predicate serves as head	医治/v 无效/v *ineffectively treat*
NJX	Juxtaposition structure	公正/a 合理/a *fair and reasonable*
NN	Noun and its nominal modifier	人身/n 安全/n *personal safety*
SBV	Predicate and its subject	财产/n 转移/v *property transfer*
VO	Predicate and its object in which the predicate serves as head	转换/v 机制/n *change mechanism*
VV	Serial verb constructions which indicates that there are serial actions	跟踪/v 报导/v *trace and report*
OT	Others	

Table 1. The dependency categories

The second one is the syntactic chunking information (a chunk is defined as a minimum non-nesting or non-overlapping phrase) (Xu and Lu, 2005). Chunking information identifies all the words for a collocation within the context of an enclosed chunk. Thus, it is a way to identify its proper context at the most immediate syntactic structure. 11 types of syntactic chunking categories given in (Xu and 2006) are used as listed in Table 2.

	Description	Examples
BNP	Base noun phrase	[市场/n 经济/n]NP *market economy*
BAP	Base adjective phrase	[公正/a 合理/a]BAP *fair and reasonable*
BVP	Base verb phrase	[顺利/a 启动/v]BVP *successfully start*
BDP	Base adverb phrase	[已/d 不再/d]BDP *no longer*
BQP	Base quantifier phrase	[数千/m 名/q]BQP 士兵/n *several thousand soldiers*
BTP	Base time phrase	[早上/t 8 时/t]BTP *8:00 in the morning*
BFP	Base position phrase	[蒙古/ns 东北部/f]BFP *Northeast of Mongolia*
BNT	Name of an organization	[烟台/ns 大学/n]BNT *Yantai University*
BNS	Name of a place	[江苏/ns 铜山/ns]BNS *Tongshan, Jiangsu Province*
BNZ	Other proper noun phrase	[诺贝尔/nr 奖/n]BNZ *The Nobel Prize*
BSV	S-V structure	[领土/n 完整/a]BSV *territorial integrity*

Table 2. The chunking categories

The third one is the classification of collocation types. Collocations cover a wide spectrum of habitual word combinations ranging from idioms to free word combinations. Some collocations are very rigid and some are more flexible. (Xu and Lu 2006) proposed a scheme to classify collocations into four types according to the internal association of collocations including compositionality, non-substitutability, non-modifiability, and statistical significance. They are,

Type 0: ***Idiomatic Collocation***

Type 0 collocations are fully non-compositional as its meaning cannot be predicted from the meanings of its components such as *缘木求鱼* (*climbing a tree to catch a fish, which is a metaphor for a fruitless endeavour*). Some terminologies are also Type 0 collocations such as *蓝 牙*(*Blue-tooth*) which refers to a wireless communication protocol. Type 0 collocations must have fixed forms. Their components are non-substitutable and non-modifiable allowing no syntactic transformation and no internal lexical variation. This type of collocations has very strong internal associations and co-occurrence statistics is not important.

Type 1: ***Fixed Collocation***

Type 1 collocations are very limited compositional with fixed forms which are non-substitutable and non-modifiable. However, this type can be compositional. None of the words in a Type 1 collocation can be substituted by any other words to retain the same meaning such as in *外交/n 豁免权/n* (*diplomatic immunity*). Finally, Type 1 collocations normally have strong co-occurrence statistics to support them.

Type 2: ***Strong Collocation***

Type 2 collocations are limitedly compositional. They allow very limited substitutability. In other words, their components can only be substituted by few synonyms and the newly generated word combinations have similar meaning, e.g., *缔结/v 同盟/n* (*alliance formation*) and *缔结/v 联盟/n* (*alliance formation*). Furthermore, Type 2 collocations allow limited modifier insertion and the order of components must be maintained. Type2 collocations normally have strong statistical support.

Type 3: ***Loose Collocation***

Type 3 collocations have loose restrictions. They are nearly compositional. Their components may be substituted by some of their synonyms and the newly generated word combinations usually have very similar meanings. Type 3 collocations are modifiable meaning that they allow modifier insertions. Type 3 collocations have weak internal associations and they must have statistically significant co-occurrence.

The classification represents the strength of internal associations of collocated words. The annotation of these three kinds of information is essential to all-rounded characteristic analysis of collocations.

4 Annotation of *CCB*

4.1 Data Preparation

CCB is based on the PolyU chunk bank (Xu and Lu, 2005) which contains chunking information on the People's Daily corpus with both segmentation and part-of-speech tags. The accuracies of word segmentation and POS tagging are claimed to be higher than 99.9% and 99.5%, respectively (Yu et al. 2001). The use of this popular and accurate raw resource helped to reduce the cost of annotation significantly, and ensured maximal sharing of our output.

The set of 3, 643 headwords are selected from "*The Dictionary of Modern Chinese Collocation*" (Mei 1999) among about 6,000 headwords in the dictionary. The selection was based both on the judgment by linguistic experts as well as the statistical information that they are commonly used.

4.2 Corpus Preprocessing

The *CCB* annotations are represented in XML. Since collocations are practical word combinations and word is the basic unit in collocation research, a preprocessing module is devised to transfer the chunked sentences in the PolyU chunk bank to word sequences with the appropriate labels to indicate the corresponding chunking information. This preprocessing module indexes the words and chunks in the sentences and encodes the chunking information of each word in two steps. Consider the following sample sentence extracted from the PolyU chunk bank:

确保/v[人民/n 群众/n]BNP 的/u[生命/n 财产/n 安全/an]BNP

(*ensure life and property safety of the people*)

The first step in preprocessing is to index each word and the chunk in the sentence by giving incremental word ids and chunk ids from left to right. That is,,

[W1]确保/v [W2]人民/n [W3]群众/n [W4]的/u
[W5]生命/n [W6]财产/n [W7]安全/an [C1]BNP [C2]BNP

where, [*W1*] to [*W7*] are the words and [*C1*] to [*C2*] are chunks although chunking positions are not included in this step. One Chinese word may occur in a sentence for more than one times, the unique word ids are helpful to avoid ambiguities in the collocation annotation on these words.

The second step is to represent the chunking information of each word. Chunking boundary information is labeled by following initial/final representation scheme. Four labels, *O/B/I/E*, are used to mark the isolated words outsides any chunks, chunk-initial words, words in the middle of chunks, and chunk-final words, respectively. Finally, a label *H* is used to mark the identified head of chunks and *N* to mark the non-head words.

The above sample sentence is then transferred to a sequence of words with labels as shown below,

<labeled> [W1][O_O_N][O]确保/v [W2][B_BNP_N][C1] 人民/n [W3][E_BNP_H][C1]群众/n [W4][O_O_N][O]的/u [W5][B_BNP_N][C2]生命/n [W6][I_BNP_N][C2]财产/n [W7][E_BNP_N][C2]安全/an </labeled>

For each word, the first label is the word ID. The second one is a hybrid tag for describing its chunking status. The hybrid tags are ordinal with respect to the chunking status of boundary, syntactic category and head, For example, *B_BNP_N* indicates that current word is the beginning or a *BNP* and this word is not the head of this chunk. The third one is the chunk ID if applicable. For the word out of any chunks, a fixed chunk ID *O* is given.

4.3 Collocation Annotation

Collocation annotation is conducted on one headword at a time. For a given headword, an annotators examines its context to determine if its cooccurred word(s) forms a collocation with it and if so, also annotate the collocation's dependency, chunking and classification information. The annotation procedure, requires three passes. We use a headword *安全/an* (*safe*), as an illustrative example.

Pass 1. Concordance and dependency identification

In the first pass, the concordance of the given headword is performed. Sentences containing the headwords are obtained, e.g.

S1: *遵循/v [确保/v 安全/an]BVP 的/u 原则/n*
(follow the principles for ensuring the safety)
S2: *确保/v [人民/n 群众/n]BNP 的/u[生命/n 财产/n 安全/an]BNP*
(ensure life and property safety of people)
S3: *确保/v 长江/ns [安全/an 度汛/v]BVP*
(ensure the flood pass through Yangzi River safely)

With the help of an automatic dependency parser, the annotator determines all syntactically and semantically dependent words in the chunking context of the observing headword. The annotation output of ***S1*** is given below in which XML tags are used for the dependency annotation.

S1:*<sentence>遵循/v [确保/v 安全/an]BVP 的/u 原则/n*

<labeled> [W1][O_O_N][O]遵循/v [W2][B_BVP_H][C1] 确保/v [W3][E_BNP_N][C1]安全/an [W4][O_O_N][O]的 /u [W5][O_O_N][O]原则/n </labeled>
<dependency no="1" observing="安全/an" head="确保/v" head_wordid="W2" head_chunk ="B_BVP_H" head_chunkid="C1" modifier="安全/an" modifier_wordid="W3" modifier _chunk="E_BVP_N" modifer_chunkid="C1" relation="VO" > </dependency>
</sentence>

Dependency of word combination is annotated with the tag <dependency> which includes the following attributes:

-<dependency> indicates an identified dependency
-no is the id of identified dependency within current sentence according to ordinal sequence
-observing indicates the current observing headword
-head indicates the head of the identified word dependency
-head_wordid is the *word id* of the head
-head_chunk is the hybrid tags for labeling the chunking information of the head
-head_chunkid is the *chunk id* of the head
-modifier indicates the modifier of the identified dependency
-modifier_wordid is the *word id* of the modifier
-modifier_chunk is the hybrid tags for labeling chunking information of the modifier
-modifier_chunkid is the *chunk id* of the modifier
-relation gives the syntactic dependency relations labeled according to the dependency labels listed in Table 1.

In ***S1*** and ***S2***, the word combination *确保/v 安全/an* has direct dependency, and in ***S3***, such a dependency does not exist as *确保/v* only determines *度汛/v* and *安全/an* depends on *度汛/v*. The quality of *CCB* highly depends on the accuracy of dependency annotation. This is very important for effective characteristics analysis of collocations and for the collocation extraction algorithms.

Pass 2. *N*-gram collocations annotation

It is relatively easy to identify *n*-gram collocations since an *n*-gram collocation is of habitual and recurrent use of a series of *bi*-grams. This means that *n*-gram collocations can be identified by finding consecutive occurrence of significant *bi*-grams in certain position. In the second pass, the annotators focus on the sentences where the headword has more than one dependency. The percentage of all appearances of each dependent word at each position around the headword is estimated with the help of a program (Xu and Lu, 2006). Finally, word dependencies frequently co-occurring in consecutive positions in a fixed order are extracted as *n*-gram collocations.

For the headword, an *n*-gram collocation *生命/n 财产/n 安全/an* is identified since the co-occurrence percentage of dependency *生命/-NN-安全/an* and dependency *财产/n-NN-安全/an* is 0.74 is greater than a empirical threshold suggest in (Xu and Lu, 2006). This *n*-gram is annotated in **S2** as follows:

<ncolloc observing="安全/an" w1="生命/n" w2="财产/n" w3="安全/an" start_wordid="5"> </ncolloc>

where,
-<ncolloc> indicates an *n*-gram collocation
-w1, w2,..wn give the components of the *n*-gram collocation according to the ordinal sequence.
-start_wordid indicates the word id of the first component of the *n*-gram collocation.

Since *n*-gram collocation is regarded as a whole, its internal dependencies are ignored in the output file of pass 2. That is, if the dependencies of several components are associated with an *n*-gram collocation in one sentence, the *n*-gram collocation is annotated and these dependencies are filtered out so as not to disturb the bi-gram dependencies.

Pass 3. *Bi*-gram collocations annotation

In this pass, all the word dependencies are examined to identify *bi*-gram collocations. Furthermore, if a dependent word combination is regarded as a collocation by the annotators, it will be further labeled based on the type determined. The identification is based on expert knowledge combined with the use of several computational features as discussed in (Xu and Lu, 2006).

An assistive tool is developed to estimate the computational features. We use the program to obtain feature data based on two sets of data. The first data set is the annotated dependencies in the 5-million-word corpus which is obtained through **Pass 1** and **Pass 2** annotations. Because the dependent word combinations are manually identified and annotated in the first pass, the statistical significance is helpful to identify whether the word combination is a collocation and to determine its type. However, data sparseness problem must be considered since 5-million-word is not large enough. Thus, another set of statistical data are collected from a 100-million segmented and tagged corpus (Xu and Lu, 2006). With this large corpus, data sparseness is no longer a serious problem. But, the collected statistics are quite noisy since they are directly retrieved from text without any verification. By analyzing the statistical features from both sets, the annotator can use his/her professional judgment to determine whether a *bi*-gram is a collocation and its collocation type.

In the example sentences, two collocations are identified. Firstly, *安全/an 度汛/v* is classified as a Type 1 collocation as they have only one peak co-occurrence, very low substitution ratio and their co-occurrence order nearly never altered. Secondly, *确保/v 安全/an* is identified as a collocation. They have frequent co-occurrences and they are always co-occurred in fixed order among the verified dependencies. However, their co-occurrences are distributed evenly and they have two peak co-occurrences. Therefore, *确保/v 安全/an* is classified as a Type 3 collocation. These *bi*-gram collocations are annotated as illustrated below,

<bcolloc observing="安全/an" col="度汛/v" head="度汛/v" type= "1" relation="ADV">
<dependency no="1" observing="安全/an" head="度汛/v" head_wordid="W4" head_chunk ="E_BVP_H" head_chunkid="C1" modifier=" 安 全 /an" modifier_wordid="W3" modifier _chunk="B_BVP_N" modifer_chunkid="C1" relation="ADV" ></dependency></bcolloc>

where,

-**<bcolloc>** indicates a *bi*-gram collocation.

-**col** is for the collocated word.

-**head** indicates the head of an identified collocation

-**type** is the classified collocation type.

-**relation** gives the syntactic dependency relations of this *bi*-gram collocation.

Note that the dependency annotations within the *bi*-gram collocations are reserved.

4.4 Quality Assurance

The annotators of *CCB* are three post-graduate students majoring in linguistics. In the first annotation stage, 20% headwords of the whole set was annotated in duplicates by all three of them. Their outputs were checked by a program. Annotated collocation including classified dependencies and types accepted by at least two annotators are reserved in the final data as the *Golden Standard* while the others are considered incorrect. The inconsistencies between different annotators were discussed to clarify any misunderstanding in order to come up with the most appropriate annotations. In the second annotation stage, 80% of the whole annotations were then divided into three parts and separately distributed to the annotators with 5% duplicate headwords were distributed blindly. The duplicate annotation data were used to estimate the annotation consistency between annotators.

5 Collocation Characteristic Analysis

5.1 Progress and Quality of *CCB*

Up to now, the first version of *CCB* is completed. We have obtained 23,581 unique *bi*-gram collocations and 2,752 unique *n*-gram collocations corresponding to the 3,643 observing headwords. Meanwhile, their occurrences in the corpus are annotated and verified. With the help of a computer program, the annotators manually classified *bi*-gram collocations into three types. The numbers of Type 0/1, Type 2 and Type 3 collocations are 152, 3,982 and 19,447, respectively.

For the 3,643 headwords in The Dictionary of Modern Chinese Collocations (Mei 1999) with 35,742 bi-gram collocations, 20,035 collocations appear in the corpus. We call this collection as Mei's Collocation Collection (MCC). There are 19,967 common entries in MCC and CCB, which means 99.7% collocations in MCC appear in CCB indicating a good linguistic consistency. Furthermore, 3,614 additional collocations are found in CCB which enriches the static collocation dictionary.

5.2 Dependencies Numbers Statistics of Collocations

Firstly, we study the statistics of how many types of dependencies a *bi*-gram collocation may have. The numbers of dependency types with respect to different collocation types are listed in Table 3.

Collocations	1 type	2 types	>2 types	Total
Type 0/1	152	0	0	152
Type 2	3970	12	0	3982
Type 3	17282	2130	35	19447
Total	21404	2142	35	23581

Table 3. Collocation classification versus number of dependency types

It is observed that about 90% *bi*-gram collocations have only one dependency type. This indicates that a collocation normally has only one fixed syntactic dependency. It is also observed that about 10% *bi*-gram collocations have more than one dependency type, especially Type 3 collocations. For example, two types of dependencies are identified in the *bi*-gram collocation *安全/an-国家/n*. They are *安全/an-AN-国家/n* (*a safe nation*) which indicates the dependency of a noun and its nominal modifier where *国家/n* serves as the head, and *国家/n-NN-安全/an* (*national security*) which indicates the dependency of a noun and its nominal modifier where 安全/an serves as the head. It is attributed to the fact that the use of Chinese words is flexible. A Chinese word may support different part-of-speech. A collocation with different dependencies results in different distribution trends and most of these collocations are classified as Type 3. On the other hand, Type 0/1 and Type 2 collocations seldom have more than one dependency type.

5.3 Syntactic Dependency Statistics of Collocations

The statistics of the 10 types of syntactic dependencies with respect to different types of *bi*-gram collocations are shown in Table 4. *No.* is the number of collocations with a given dependency type *D* and a given collocation type *T*. The percentage of *No.* among all collocations with the same collocation type *T* is labeled as *P_T*, and the percentage of *No.* among all of the collocations with the same dependency *D* is labeled as *P_D*.

	Type 0/1			Type 2			Type 3			Total	
	No.	*P_T*	*P_D*	*No.*	*P_T*	*P_D*	*No.*	*P_T*	*P_D*	*No.*	*P_T*
ADA	1	0.7	0.1	212	5.3	11.5	1637	7.6	88.5	1850	7.2
ADV	9	5.9	0.3	322	8.1	11.2	2555	11.8	88.5	2886	11.2
AN	20	13.2	0.4	871	21.8	15.4	4771	22.0	84.3	5662	22.0
CMP	12	7.9	2.2	144	3.6	26.9	379	1.8	70.8	535	2.1
NJX	8	5.3	3.2	42	1.1	16.9	198	0.9	79.8	248	1.0
NN	44	28.9	0.9	1036	25.9	21.6	3722	17.2	77.5	4802	18.6
SBV	4	2.6	0.2	285	7.1	11.1	2279	10.5	88.7	2568	10.0
VO	26	17.1	0.5	652	16.3	12.5	4545	21.0	87.0	5223	20.2
VV	3	2.0	0.2	227	5.7	13.4	1464	6.8	86.4	1694	6.6
OT	25	16.4	7.7	203	5.1	62.5	97	0.4	29.8	325	1.3
Total	152	100.0	0.6	3994	100.0	15.5	21647	100.0	83.9	25793	100.0

Table 4. The statistics of collocations with different collocation type and dependency

Corresponding to 23,581 *bi*-gram collocations, 25,793 types of dependencies are identified (some collocations have more than one types of dependency). In which, about 82% belongs to five major dependency types. They are *AN*, *VO*, *NN*, *ADV* and *SBV*. It is note-worthy that the percentage of *NN* collocation is much higher than that in English. This is because nouns are more often used in parallel to serve as one syntactic component in Chinese sentences than in English.

The percentages of Type 0/1, Type 2 and Type 3 collocations in *CCB* are 0.6%, 16.9% and 82.5%, respectively. However, the collocations with different types of dependencies have shown their own characteristics with respect to different collocation types. The collocations with *CMP*, *NJX* and *NN* dependencies on average have higher percentage to be classified into Type 0/1 and Type 2 collocations. This indicates that *CMP*, NJX and *NN* collocations in Chinese are always used in fixed patterns and these kinds of collocations are not freely modifiable and substitutable. In the contrary, many *ADV* and *AN* collocations are classified as Type 3. This is partially due to the special usage of auxiliary words in Chinese. Many *AN* Chinese collocations can be inserted by a meaningless auxiliary word 的/u and many *ADV* Chinese collocations can be inserted by an auxiliary word 地/u. This means that many *AN* and *ADV* collocations can be modified and thus, they always have two peak co-occurrences. Therefore, they are classified as Type 3 collocations. 7.7% and 62.5% of the collocations with dependency *OT* are classified as Type 0/1 and Type2 collocations, respectively. Such percentages are much higher than the average. This is attributed by the fact that some Type 0/1 and Type 2 collocations have strong semantic relations rather than syntactic relations and thus their dependencies are difficult to label.

5.4 Chunking Statistics of Collocations

The chunking characteristic for the collocations with different types and different dependencies are examined. In most cases, Type 0/1/2 collocations co-occur within one chunk or between neighboring chunks. Therefore, their chunking characteristics are not discussed in detail. The percentage of the occurrences of Type 3 collocations with different chunking distances are given in Table 5. If a collocation co-occurs within one chunk, the chunking distance is 0. If a collocation co-occurs between neighboring chunks, or between neighboring words, or between a word and a neighboring chunk, the chunking distance is 1, and so on.

	ADA	ADV	AN	CMP	NJX	NN	SBV	VO	VV	OT
0 chunk	56.8	53.1	65.7	48.5	70.2	62.4	46.5	41.1	47.2	86.4
1 chunk	38.2	43.7	28.5	37.2	15.4	27.9	41.2	35.7	41.1	13.5
2 chunks	5.0	3.2	3.7	14.2	14.4	9.7	11.0	17.6	9.6	0.1
>2chunks	0.0	0.0	2.1	0.1	0.0	0.0	1.3	5.6	2.1	0.0

Table 5. Chunking distances of Type 3 collocations

It is shown that the co-occurrence of collocations decreases with increased chunking distance. Yet, the behavior for decrease is different for collocations with different dependencies. Generally speaking, the *ADA*, *ADV*, *CMP*, *NJX*, *NN* and *OT* collocations seldom co-occur cross two words or two chunks. Furthermore, the occurrences of *AN*, *NJX* and *OT* collocations quickly drops when the chunking distance is greater than 0, i.e. these collocations tends to co-occur within the same chunk. In the contrary, the co-occurrences of *ADA*, *ADV*, *CMP*, *SBV* and *VV* collocations corresponding to chunking distance equals 0 and 1 decrease steadily. It means that these four kinds of collocations are more evenly distributed within the same chunk or between neighboring words or chunks. The occurrences of *VO* collocations corresponding to chunking distance from 0 to 3 with a much flatter reduction. This indicates that a verb may govern its object in a long range.

6 Conclusions

This paper describes the design and construction of a manually annotated Chinese collocation bank. Following a set of well-designed annotation guideline, the collocations corresponding to 3,643 headwords are identified from a chunked five-million word corpus. 2,752 unique *n*-gram collocations and 23,581 unique *bi*-gram collocations are annotated. Furthermore, each *bi*-gram collocation is annotated with its syntactic dependency information, classification information and chunking information. Based on *CCB*, characteristics of collocations with different types and different dependencies are examined. The obtained result is essential for improving research related to Chinese collocation. Also, *CCB* may be used as a standard answer set for evaluating the performance of different collocation extraction algorithms. In the future, collocations of all unvisited headwords will be annotated to produce a complete 5-million-word Chinese collocation bank.

Acknowledgement

This research is supported by The Hong Kong Polytechnic University (A-P203), CERG Grant (5087/01E) and Chinese University of Hong Kong under the Direct Grant Scheme project (2050330) and Strategic Grant Scheme project (4410001).

References

Bedin N. et al. 2003. The Annotation Process in the Turkish Treebank. In *Proc. 11th Conference of the EACL-4th Linguistically Interpreted Corpora Workshop- LINC*.

Kosho S. et al. 2000. Collocations as Word Co-occurrence Restriction Data - An Application to Japanese Word Processor. In *Proc. Second International Conference on Language Resources and Evaluation*

Lin D.K. 1998. Extracting collocations from text corpora. In *Proc. First Workshop on Computational Terminology*, Montreal

Manning, C.D., Schütze, H. 1999: *Foundations of Statistical Natural Language Processing*, MIT Press

Mei J.J. 1999. *Dictionary of Modern Chinese Collocations*, Hanyu Dictionary Press

Pecina P. 2005. An Extensive Empirical Study of Collocation Extraction Methods. In *Proc. 2005 ACL Student Research Workshop*. 13-18

Smadja. F. 1993. Retrieving collocations from text: Xtract, *Computational Linguistics*. 19. 1. 143-177

Tutin A. 2005. Annotating Lexical Functions in Corpora: Showing Collocations in Context. In *Proc. 2nd International Conference on the Meaning – Text Theory*

Xu R. F. and Lu Q. 2005. Improving Collocation Extraction by Using Syntactic Patterns, In *Proc. IEEE International Conference on Natural Language Processing and Knowledge Engineering*. 52-57

Xu, R.F. and Lu, Q. 2006. A Multi-stage Chinese Collocation Extraction System. *Lecture Notes in Computer Science, Vol. 3930*, Springer-Verlag. 740-749

Yu S.W. et al. 2001. *Guideline of People's Daily Corpus Annotation*, Technical Report, Peking University