International Journal of Computer Processing Of Languages
Vol. 22, No. 1 (2009) 21–47

World Scientific
www.worldscientific.com

Building a Chinese Collocation Bank

Ruifeng Xu[‡,*], Qin Lu[*,§], Kam-Fai Wong[†] and Wenjie Li[*,¶]

[‡]Department of Chinese, Translation and Linguistics,
City University of Hong Kong, Hong Kong
[*]Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[†]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, N.T., Hong Kong
[‡]*ruifeng.xu@cityu.edu.hk*
[§]*csluqin@comp.polyu.edu.hk*
[†]*kfwong@se.cuhk.edu.hk*
[¶]*cswjili@comp.polyu.edu.hk*

This paper presents the design and construction of an annotated Chinese collocation bank as the resource to support systematic research on Chinese collocations. The definition and properties are first studied. Based on a combination of different properties, a classification scheme is proposed to categorize Chinese collocations into four types. With the help of computational tools, bigram collocations and n-gram collocations of 3,643 headwords are manually identified in a 5-million-word corpus. Furthermore, for each identified bigram collocation, its dependency relation, chunking information and classification are annotated to produce a collocation bank. Currently, the Chinese collocation bank contains 23,581 bigram collocations and 2,752 n-gram collocations. The Chinese collocation bank is a valuable resource for Chinese collocation related research. Through statistical analysis on the collocation bank, some interesting characteristics of Chinese bigram collocations are presented in this paper.

Keywords: Chinese collocation; collocation bank; collocation annotation; collocation classification.

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 say 历史-包袱 (*historical baggage*)

[‡]Correspondence author.

rather than 历史-行李 (*historical luggage*) even though 包袱 (*baggage*) and 行李 (*luggage*) are synonymous. However, no one can explain why 历史 (*historical*) must collocate with 包袱 (*baggage*). Generally speaking, collocations are close-knit and frequently used word combinations. The collocated words always have syntactic or semantic relations but they cannot be accounted for directly by general syntactic or semantic rules. They mainly reflect the habitual use of natural language, as said in Firth's famous dictum that "*a word is characterized by the company it keeps*" [1]. 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 are widely employed in many natural language processing (NLP) tasks, such as in word sense disambiguation, machine translation, information retrieval and natural language generation [2–5].

Although the importance of collocation is well known, it is difficult to compile a comprehensive list of collocations. The definition and scope of collocation has always been subjected to debate [6–8]. This has brought much confusion in the research related to collocation. Traditional linguists manually identified and compiled typical collocations based on expert knowledge [9]. However, the coverage and consistency of these works can be problematic [5]. Meanwhile, a simple collocation list without much context information is insufficient to support many NLP tasks. Even though there are some attempts to acquire collocation knowledge through automatic extraction from electronic text [5, 10, 11], these techniques are mainly based on co-occurrence analysis using statistical techniques combined with some syntax analysis. Their performances are not satisfactory for direct use [12, 13]. Collocations cover a wide spectrum ranging from idioms to free word combinations. Some collocations are very rigid and some are flexible but many are in between. Finding features that can characterize collocations in the whole spectrum are difficult. This explains why previous research failed to have satisfactory performance. Collocations are word combinations, which co-occur within a short context. However, not all such co-occurrences are true collocations. Thus, further examinations are needed to identify true collocations from co-occurred word pairs. A large-scale corpus with true collocations identified and annotated is referred to as a ***collocation bank***. A collocation bank is an indispensable resource for natural language processing research especially for collocation related research. Context of true collocations can provide valuable information not only on the nature of collocations, but also clues for improving the performance of automatic collocation extraction approaches. A collocation bank can also serve as a common standard to compare the performance of different collocation extraction algorithms. Work on collocation bank construction is very limited. Kosho *et al*. presented their

works of manual collocation identification and annotation on Japanese text [14]. The Turkish Treebank included manual collocation annotation in its annotation among other things [15]. Furthermore, Tutin used shallow analysis based on finite state transducers and lexicon-grammar proposed by Gross [16] to identify and annotate collocations in a French corpus [17]. This collocation bank further provided the lexical functions of collocations.

This paper presents the design and construction of a Chinese collocation bank (acronymis *CCB*). This is the first attempt to build a large scale Chinese collocation bank. The annotation is done using a headword driven approach. Each given keyword is regarded as a headword and the collocations related to the headword are annotated. *CCB* provides multiple linguistic information including (1) annotation of collocated words for each given headword, (2) distinction of n-gram from bigram collocations for the headwords, (3) annotation of syntactic dependencies, chunking relation and classification of bigram collocations. Well-designed quality assurance mechanisms are used for the construction of *CCB* using effective procedures and assistive tools. Currently, *CCB* contains 3,643 common headwords taken from *The Dictionary of Modern Chinese Collocations* [9] with 23,581 bigram collocations and 2,752 n-gram collocations extracted from a five-million word-segmented, part-of-speech (POS) tagged and chunked Chinese corpus [11]. Through the qualitative and quantitative analysis of annotated data, some characteristics of bigram collocations are revealed. Experiments show that *CCB* is a helpful resource to improve existing collocation extraction systems.

The rest of this paper is organized as follows. Section 2 presents the definition, the qualitative properties and the quantitative computational features of collocations. Based on these properties, collocation classification principles and rationales are presented. Section 3 presents the annotation guideline. Section 4 describes the practical issues in the annotation including corpus preparation, headword preparation, annotation flow, and quality assurance mechanisms. Section 5 gives the current status of *CCB* and the analysis of the annotated collocations. Section 6 estimates the effectiveness of *CCB* when it is used for improving an existed automatic collocation extraction system. Section 7 concludes this paper.

2. Definition, Properties and Classification of Collocations

2.1. Definition of collocation

Although collocations are commonly found in natural language use and they can be easily understood by people, a precise definition of collocation remains

elusive [8]. Various definitions and scopes have been proposed depending on different features of collocation. Church *et al*. defined collocations to "*a pair of words which co-occur more often than would be expected by chance*" [7]. This definition is only interested in collocations consisting of bigrams which can either be adjacent or interrupted. Richards and Schmidt defined collocations as "*a sequence of words or terms which co-occur more often than would be expected by chance*", which only considers collocations with adjacent words [18]. Both definitions emphasize the statistical significance of collocations. Yet, they both lack qualifiers to distinguish true collocations from free word combinations since the syntactic and semantic relations of the participating words are not apparent. A widely used definition is given by Benson as "*A collocation is an arbitrary and recurrent word combination*" [6], which emphasizes the semantic associations of collocated words. Allerton considered collocations for only the collocated words for which the whole meaning of the collocation cannot be predicted from individual components. This is considered quite rigid by many researchers.

To avoid confusion, collocation in this study is defined based on [8] with more concrete descriptions as stated below:

Definition: *A* ***collocation*** *is a recurrent and conventional expression containing two or more content word combinations that hold syntactic and/or semantic relations. More specifically, content words include noun, verb, adjective, adverb, determiner, etc.*

In this study, collocations are investigated for word combinations between content words only because these word combinations carry both syntactic and semantic information whereas function words may not carry independent semantic information. According to the number of collocated words that fit our definition, collocations can either be **bigram collocations** which contain only two words or **n-gram collocations** which contain more than two words. Collocated words can either be adjacent to each other, which are referred to as **uninterrupted collocation**, or separated by other words in between, which are referred to as **interrupted collocation**.

2.2. Properties of collocations

From a linguistic viewpoint, collocations can cover a large spectrum of word combinations which at one extreme are close to free word combinations with only syntactic bindings or at the other are very restrictive in usage such as idioms [20]. However, they do share certain common characteristics.

Firstly, collocations are recurrent co-occurrences as they are of habitual use [5]. Collocations occur frequently in similar contexts and they always appear in certain fixed patterns. This is the most important property of a collocation differing from random word combinations.

Secondly, collocations are of habitual use [5]. Many collocations cannot be described by general syntactic and semantic rules. For instance, we only say 浓茶 (*strong tea*), but not 烈茶 (*powerful tea*) even though both are syntactically sound and there is not much semantic differences. The choice of which is completely habitual. No syntax and semantic reason can explain the particular choice of words.

Thirdly, collocations are to a limited extent compositional. Brundage introduced the term "compositional" to describe the property where the meaning of an expression could be predicted from the meanings of its components [21]. Free word combinations can be generated by linguistic rules and their meanings are the combinations of their components. Thus, they are normally considered compositional. At the other extreme, idioms are always non-compositional, implying that their meanings are different from the combination of their components. Collocations are in between free combinations and idioms. They are expected to be compositional to a limited extent. In other words, collocations are expected to have additional meaning beyond their literal meaning. On the other hand, for those word combinations that have little additional meaning over the literal meaning, they can also be regarded as collocations if they show close semantic restrictions between their components.

Fourthly, collocations are limitedly substitutable and limitedly modifiable. Limitedly substitutable here refers the fact that a word in a collocation cannot be replaced freely by other words in its context, even for its synonyms. In other words, even if they are both syntactically or semantically identical (near-identical), they may still not be replaced. Also, collocations cannot be modified freely by adding modifiers or through grammatical transformations. For example, a collocation of 斗志-昂扬 (*high spirit*) does not allow further modification or insertion between two words.

Fifthly, most collocations are grammatically well-formed. Especially, most bigram collocations have direct syntactic relation or dependency, except for the idiomatic ones. Therefore, collocations always have both syntactic and semantic relations. Meanwhile, many collocations have fixed linear order. For example, in Verb-Object Chinese collocations, the verb always appears before object. That is "弹/v 钢琴/n" (*play piano*) and "踢/v 足球/n" (*play football*) are correct collocations while "钢琴/n 弹/v", "足球/n 踢/v"are wrong.

Sixthly, collocations are domain dependent. The collocations frequently used in one domain may be seldom used in another domain, especially the technical collocations. For example, "专家/n 系统/n" means expert system which is frequently used in computer science, but it seldom appears in general. Meanwhile, some collocations are more frequently used verbally while some others are always used in formal text. It means that collocations are domain dependent.

Lastly, collocations are language dependent, which means that some collocations in one language are no longer collocations when translated to another language, e.g., when a Chinese collocation 开/v 枪/n is translated to a single word *fire* in English. Meanwhile, many collocations cannot be translated simply word for word such as *play ball* as 打/v 球/n and *play piano* as 弹/v 钢琴/n. This can be attributed to the conventional patterns used in different languages.

2.3. Classification of Chinese collocations

Collocations cover a wide spectrum ranging from idioms to free word combinations. Some collocations are very rigid and some are flexible. There are also collocations that are in between. This is one of the reasons why different collocation definitions were proposed in the literature. Previous works in collocation extraction and collocation bank construction have not attempted to distinguish different kinds of collocations, leading to a rather weak linguistic consistency. Based on linguistic properties and co-occurrence statistics of typical collocations, this work classifies collocations according to their internal associations which are based on compositionality, substitutability, modifiability, order altering ability and statistical significance. Collocations with similar internal associations are classified into one of the following four types. Since collocations are language dependent, the English translation of some examples in Chinese collocations may not be not good examples of collocations in English.

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 in 缘木求鱼 (*climbing a tree to catch a fish, which is a metaphor for a fruitless endeavor*). Most Type 0 collocations are already listed as idioms in a dictionary. Some domain specific terms are also Type 0 collocations. For example, the term 蓝牙 (*Blue-tooth*), which is a made up word in computer science to refer to a wireless

communication protocol, has no relation to either 蓝 (*blue color*) or 牙 (*tooth*). Type 0 collocations must have fixed forms. Their components must be non-substitutable and non-modifiable allowing no syntactic transformation and no internal lexical substitutions. Their components cannot be shifted around or modified by adding or removing any component. As a fixed habitual use, it can certainly be verified by statistics. But the strong internal association measures by themselves makes verification on statistics unnecessary.

Type 1: Fixed collocation

Type 1 collocations are very limited in compositionality, and they have certain fixed forms which are non-substitutable and non-modifiable. For example, the meaning of the collocation 外交/n 豁免权/n (*diplomatic immunity*), is not completely different from that of its components. Yet this combination in itself carries a special meaning which makes it non-divisible. None of the words in a Type 1 collocation can be substituted by any other words to retain the same meaning. Thus, Type 1 collocations are limited compositional. Also they do not support order altering. Finally, Type 1 collocations normally have strong co-occurrence significance to support them.

Type 2: Strong collocation

Type 2 collocations are also limited in compositionality. They are also to a very limited extent substitutable. In other words, their components can only be substituted by few synonyms and the newly formed word combination retain similar meaning.

Furthermore, Type 2 collocations allow limited modifier insertion and the order of components must be maintained. Type 2 collocations normally have strong statistical support. For example, each word in the collocation "裁减/v 员额/n" (*reduce the posts*) cannot be substituted by their synonyms and the word order cannot be changed. Meanwhile, a modifier can be inserted in this collocation, giving a new n-gram collocation like "裁减/v 军队/n 员额/n" (*reduce the army post*) or "裁减/v 政府/n 员额/n" (*reduce the government post*). Thus, this collocation is a Type 2 collocation.

Type 3: Loose collocation

Type 3 collocations have loose restrictions. They are nearly compositional. Their components can be easily substituted by their synonyms without change in meaning. This means that more substitutions of its components are allowed

Table 1. Comparison of different types of collocations.

	Type 0	Type 1	Type 2	Type 3
Compositional	no	very limited	limited	nearly yes
Substitutable	no	no	very limited	limited
Modifiable	no	no	very limited	nearly yes
Statistical significance	not required	not required	required	strongly required
International association	strongest	very strong	strong	weak

but the substitution is not necessarily free. Type 3 collocations allow modifier insertion and component order alteration. They are also modifiable. Type 3 collocations have weak internal associations and they must have significant co-occurrence statistical support. Here are some examples: 合法/v 收入/n (*lawful income*), 正当/v 收入/n (*legitimate income*), and 合法/v 收益 (*lawful profit*).

Table 1 summarizes the differences among the four types of collocations in terms of compositionality, substitutability, modifiability, and statistical significance. It can be seen that the strength of internal associations from Type 0 to Type 3 are reduced. Furthermore, by classifying collocations into four types, the individual characteristics of each type of collocations in terms of compositional, substitutable, modifiable and internal association can be measured in terms of difference in degrees.

3. Annotation Guidelines

The first step in developing a collocation bank is the establishment of a set of clear, unambiguous, easy to understand and easy to follow annotation guideline. The annotation takes two annotation strategies.

(1) *The headword-driven strategy*. The annotation uses selected headwords as the starting points. In each annotation cycle, the words co-occurred within the observing headword are collected as candidates. The word combinations that are collocations according to our definition are identified and annotated. Headword-driven strategy makes a more targeted annotation, which uses the headword as the focal point so that degree of associations of different characteristics of collocation candidates can be easily obtained for quantitative measures.

(2) *Manual annotation with the help of assistive tools for acquisition of computational features*. Manual annotation can achieve good accuracy especially for semantic associations and dependency determination. It is particularly effective for data sparseness problem because expert knowledge is a must without statistical support. But, inconsistency in human judgment is inevitable. Furthermore, experts often need supportive data to accurately analyze the characteristics of observing instances. Quantitative analysis and comparison of several related collocations are much easy to acquire through computer tools than relying on human judgment. Thus, the annotation of *CCB* takes a semi-automatic strategy. A software tool is used to acquire the classification features and candidates based on this tool are presented to the expert annotators for selection.

The annotation guidelines specify the information to be identified and the labels used in the annotation. For a given headword, *CCB* annotates both bigram collocations and n-gram collocations. For example, given a headword 合作/n (*co-operation*), both the bigram collocations e.g. 全面/a 合作/n (*all-rounded co-operation*) and n-gram collocations e.g. 南/f 南/f 合作/vn (*co-operation between south earth countries*) are annotated. For a collocation, every occurrence of it in the corpus is visited and only true collocations is annotated. For example, in the context 加强/v 两国/n 之间/f 全面/a 合作/n (*enhance the all-rounded co-operation between two nations*), 全面/a and 合作/n is a true collocation. However, in the context 全面/ad 发展/v 两国/n 友好/a 合作/n 关系/n (*all-rounded develop the friendly co-operation relationship between two nations*), 全面/a and 合作/n are not collocated words even though it appeared together. In fact, *全面/ad* is collocated with 发展/v and 合作/n is collocated with 友好/a in this occurrence. The verification of true collocation instances depicts the characteristics of collocations more accurately.

For bigram collocations, three kinds of information are annotated. The first one is the syntactic dependency of the 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). A dependency describes the direct syntactic relationship between two words. Totally, 10 types of important dependencies are annotated in *CCB*. They are listed in Table 2 below with examples.

The second annotation for bigram collocation is syntactic chunking information. A chunk is defined as a minimum non-nesting phrase [11]. Normally, a chunk has a lexical word as its head. Essentially, a chunk consists of continuous words and contains neither no nesting components nor overlapping with other phrases. Chunking information identifies all the context words for a collocation

within an enclosed chunk or related chunks. It helps to identify the proper context of a bigram collocation within the appropriate syntactic structures. 11 types of syntactic chunking categories proposed in [11] are adopted in this study. They are listed in Table 3 with examples.

Table 2. The dependency categories.

	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 3. The syntactic chunking categories.

	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*

The third one is the classification of each bigram collocation as given in Section 2. The classification represents the type of collocations based on the strength of internal associations of the collocations in terms of compositionality, substitutability, modifiability, order altering and statistical significance. The annotation of the three kinds of information is essential to all-rounded characteristic analysis of bigram collocations.

Considering that n-gram collocations consisting of continuous significant bi-grams as a whole and they are mostly in some fixed patterns, no additional information for n-gram collocations will be annotated. In order not to mix n-grams with bi-gram information, the component bigrams within an n-gram collocation are not further annotated and will not affect bigram statistics.

4. Annotation of the Collocation Bank

4.1. Corpus data preparation and headword set preparation

CCB is annotated using People's Daily corpus, a widely used high quality manually segmented corpus with part-of-speech tags by Peking University [22]. The use of this popular corpus significantly reduced the cost of annotation of this work It also makes our *CCB* widely usable. The chunking information provided in [11], which is also annotated on the People's Daily corpus, are used to build *CCB*.

A linguistic resource, *The Dictionary of Modern Chinese Collocation* [9] listed 35,742 bigram collocations corresponding to nearly 6,000 headwords. The occurrence statistics of these collocations in the People's Daily corpus are collected. Only collocations with occurrence of more than an empirical threshold and fit the collocation definition in this work are considered for annotation, which results in a total of 23,581 bigram collocations corresponding to 3,643 headwords. These 3,643 headwords form the headword set for *CCB* annotation.

4.2. Corpus preprocessing

The *CCB* is annotated using XML. Before collocation related annotation is done, preprocessing works are done to properly indexed words and chunks and then label them with appropriate labels for collection of collocation related information. The first work in preprocessing is to index each word and each chunk in each sentence from left to right. For example, a chunked sentence "遵循/v [确保/v# 安全/an]BVP 的/u 原则/n (*follow the principles for ensuring the safety*)" (# flags the head of a chunk) is labeled as

[W1]遵循/v [C1][BVP][W2]确保/v [W3]安全/an [/C1] [W4]的/u [W5]原则/n

where [W1] to [W5] are the word ids and [C1] is the chunk id. The second work is to give a hybrid tag for each word according to its syntactical information and positions in the chunk. A hybrid tag contains three pieces of information in the form of "*BoundaryInd_SynCat_HeaderInd*" where *BoundaryInd* is the boundary indicator which can take four labels *O/B/I/E* to mean a word either *O*utside or in the *B*egin, *I*nside, or *E*nd of a chunk. *SynCat* indicates the syntactic categories as given in Table 3. Normally a chunk contains two to four words with one content word as its header. The header information is indicated by *HeaderInd* with two values *H/N* to mean either a *H*eadword or a *N*on-headword. Based on this label definitions, the above sample sentence is then transferred to a sequence of words with labels as shown below:

[W1][O_O_N][O] 遵循/v [W2][B_BVP_H][C1] 确保/v [W3][E_BVP_N][C1] 安全/an [W4]][O_O_N][O] 的/u [W5][O_O_N][O] 原则/n

For example, [W2][B_BVP_H][C1] 确保/v indicates that this word is the second in the sentence. It is the beginning and the header of a base verb phrase while this phrase is the first in the sentence.

4.3. Collocation annotation

Collocation annotation is conducted for one headword at a time. The annotation involves three passes. The first pass identify all the dependency relations for the set of headwords. Based on the dependency relations, n-gram collocation are identified in the second pass, and are labeled as such. In the third pass, bi-gram collocations are identified. Furthermore, collocation types are also annotated for bigrams.

Pass 1. Concordance and dependency identification

In this pass, all the sentences containing the headwords are first obtained through a concordance tool. Take the headword 安全/an (*safe*), as an illustrative example. 安全/an (*safe*) occurs in the following three sentences:

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 a Chinese dependency parser provided by Harbin Institute of Technology, China, the annotators examine the dependencies in the context of the headword. A dependency relation is a triple consisting of a *head*, a dependency relationship type and a *modifier*. It is annotated by the tag <dependency> which includes the following attributes:

<dependency>	: dependency tag,
no	: dependency id in the current sentence,
headword	: headword id,
head	: id of the head in this dependency,
modifier	: id of the modifier in this dependency,
relation	: dependency type according to Table 2.

Then, the dependencies of the headword 安全/an (*safe*) can be annotated for **S1** to **S3** 安全 as follows:

S1: [W1][O_O_N][O] 遵循/v [W2][B_BVP_H][C1] 确保/v
[W3][E_BVP_N][C1] 安全/an [W4]][O_O_N][O] 的/u
[W5][O_O_N][O] 原则/n
<dependency no="1" headword="W2" head="W2" modifier="W3"
relation="VO"> </dependency>

S2: [W1][O_O_N][O] 确保/v [W2][B_BNP_N][C1] 人民/n
[W3][E_BNP_H][C1] 群众/n]BNP [W4] 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
<dependency no="1" headword="W7" head="W1" modifier="W7"
relation="VO" > </dependency>
<dependency no="2" headword="W7" head="W7" modifier="W5"
relation="NN" > </dependency>
<dependency no="2" headword="W7" head="W7" modifier="W6"
relation="NN" > </dependency>

S3: [W1]][O_O_N][O] 确保/v [W2]][O_O_N][O] 长江/ns
[W3][B_BVP_N][C1] 安全/an [W4][E_BNP_H][C1] 度汛/v
<dependency no="1" headword="W3" head="W4" modifier="W3"
relation="ADV" > </dependency>

It can be seen that in **S1** and **S2**, the word bigram 确保/v 安全/an has direct dependency. However, such a dependency does not exist in **S3** as 确保/v only determines 度汛/v and 安全/an depends on 度汛/v. This shows that the dependency annotation can provide information on whether certain co-occurrences are syntactically related or not. The annotated dependency triples will be analyzed in the following passes to identify true collocations.

Pass 2. N-gram collocations annotation

In this pass, the annotators focus on the sentences where the headword has more than one dependency to identify n-grams. It is relatively easy to identify n-gram collocations since an n-gram collocation is a longer word sequence with stable and recurrent use. The components of an n-gram collocation must also be recurrent and appear in a fixed position. This means that n-gram collocations can be identified by finding consecutive occurrence of significant dependency triples. The dependencies frequently co-occurring in consecutive positions and in fixed order are extracted as n-gram collocations. An n-gram is annotated with the following attributes.

<ncolloc> : n-gram collocation tag,
no : n-gram id in the current sentence,
headword : headword id,
start : id of the first word of the n-gram collocation,
end : id of the last word of the n-gram collocation.

For the example headword 安全/an, an n-gram collocation 生命/n 财产/n 安全/an is identified in **S2** which is annotated as follows:

<ncolloc no="1" headword="安全/an" start="W5" end="W7"> </ncolloc>

An n-gram collocation is regarded as a whole. No annotation of the internal syntactic and semantic relations of n-gram collocations is performed. To avoid duplicate annotation, if an n-gram collocation is identified, the dependency relations for its component bigrams are removed so that Pass 3 will not revisit the contained bi-grams.

Pass 3. Bigram collocations annotation

In this pass, dependencies are examined to identify bigram collocations. Once a dependent word pair is regarded as a collocation, its collocation type is also

annotated. The type annotation is based on human knowledge and the use of the assistive tool which will be introduced in Section 4.4. A bigram collocation is annotated with the following attributes:

<bcolloc> : bigram collocation Tag,
headword : headword id,
col_id : id of the collocated word,
type : collocation type.

In the example sentence **S1** and **S2**, the word pair 确保/v 安全/an is observed, it has strong co-occurrence frequency with fixed order and the co-occurrences are distributed evenly in two peak positions (observation is based on the statistics in the whole corpus). Therefore, 确保/v 安全/an is classified as Type 3. In **S3**, 安全/an 度汛/v is classified as a Type 1 collocation because it has strong occurrence frequency with only one peak and also has very low substitution ratio with no altered order. Consequently, there are three identified collocations for headword 安全/an as annotated below,

S1: <bcolloc headword="W2" col_id="W2" type="3">
S2: <bcolloc headword="W7" col_id="W1" type="3">
S3: <bcolloc headword="W3" col_id="W4" type="1">

4.4. Quality assurance and internal association measure

4.4.1. *Quality assurance in manual annotation*

The annotators are three post-graduate students majoring in linguistics. To ensure annotation quality, the annotation is done in two phases. In the first annotation phrase, the collocations corresponding to 20% of the headwords were annotated by all annotators in duplicates. Their outputs were then checked by a program. Any identical annotations by either two or three annotators are saved separately as the standard while others are considered incorrect. The inconsistencies between different annotators were then discussed to clear any misunderstanding in order to come up with the most appropriate annotations and proper understanding by all annotators. In the second phase, the rest of the 80% headwords were then divided into three parts for annotation by different annotators in which 5% of headwords are distributed in duplicates to all three annotators so that the annotation agreement between different annotators can be estimated.

4.4.2. *Quantitative measures for internal associations*

The collocation classification proposed in Section 2 is based on the qualitative properties of collocations. This classification has good linguistic consistency. However, the significance of some properties is described by using *limited*, *very limited*, etc., are only qualitative which needs quantitative measures to justify. Following the existing works [5, 13, 23], quantitative measures are collected based on headwords. For a headword w_{hw}, any word within the context window of w_{hw} with a distance of d in the range of [−5, 5] is a co-word of w_{hw}, denoted as $w_{co\text{-}i}$ for $1 \le i \le k$, where k is the total number of unique co-words of w_{hw}. Each occurrence of a word pair is then recorded as (w_{hw}, $w_{co\text{-}i}$, d). A bigram table for the headword w_{hw} is then established, denoted as $BT(w_{hw})$. The "word" in this study refers to a word with a POS tag. Thus, a word with different POS tags, such as 安全/an (*safety*/*adnoun*), 安全/a (*safe*/*adjectives*), and 安全/d (*safely/adverbs*) are regarded as three different words.

Measures for substitutability

Synonym substitution ratio is a measure for estimating substitutability within a bigram pair. For each headword w_{hw} and its co-word $w_{co\text{-}i}$ which forms a bigram pair (w_{hw}, $w_{co\text{-}i}$), suppose there are two corresponding synonym sets for w_{hw} and $w_{co\text{-}i}$, denoted by $S_{syn}(w_{hw})$ and $S_{sym}(w_{co\text{-}i})$, respectively. For each word in $S_{syn}(w_{hw})$, labeled as $w_{syn\_hw\_j}$, (j = 1 to $|S_{syn}(w_{hw})|$) if the bigram pair ($w_{syn\_hw\_j}$, $w_{co\text{-}i}$) occur in $BT(w_{hw})$ with a frequency, labeled as $f(w_{syn\_hw\_j}, w_{co\text{-}i})$, greater than an empirical threshold (say 3 in a 5-million-word corpus), w_{hw} is regarded as it can be substituted by $w_{syn\_hw\_j}$ when collocated with $w_{co\text{-}i}$. The synonym substitution ratio of w_{hw} collocated with $w_{co\text{-}i}$, is then calculated by,

$$SubstitutionRatio(w_{hw}) = \sum_{j=1}^{|S_{syn}(w_{hw})||} v_j / |S_{syn}(w_{hw})|,$$

$$\text{where } v_j = \begin{cases} 1, & \text{if } f(w_{syn\_hw\_j}, w_{co\text{-}i}) > \text{threshold} \\ 0, & \text{otherwise} \end{cases}. \qquad (1)$$

The value of synonyms substitution ratio of w_{hw} ranges from 0 to 1. A smaller value means this word combination tends to be more non-substitutable. The synonym substitution ratio of $w_{co\text{-}i}$ can be measured in the same way.

Measures for modifiability

Modifiability is measured by two computational features. The first one is the *CrowdingLevel* which characterizes the distribution significance that $w_{co\text{-}i}$ around w_{hw}. The *CrowdingLevel* is defined as,

$$CrowdingLevel(w_{hw}, w_{co\text{-}i}) = \frac{\sum_{m=-5}^{5} \left| f(w_{hw}, w_{co\text{-}i}, m) - \frac{1}{10} \cdot \sum_{m=-5}^{5} f(w_{hw}, w_{co\text{-}i}, m) \right|}{\sum_{m=-5}^{5} f(w_{hw}, w_{co\text{-}i}, m)}. \quad (2)$$

The value of *CrowdingLevel*(w_{hw}, $w_{co\text{-}i}$) ranges from 0 to 1, and the larger value means w_{hw} and $w_{co\text{-}i}$ tends to co-occur in limited positions, i.e. w_{hw} and $w_{co\text{-}i}$ is less modifiable.

The second feature is the number of *Peak Co-occurrence*. For the word bigram (w_{hw}, $w_{co\text{-}i}$) having large value of *CrowdingLevel*(w_{hw}, $w_{co\text{-}i}$), its position of peak co-occurrence is observed. Assume that $f(w_{hw}, w_{co\text{-}i}, m)$ is the frequency $w_{co\text{-}i}$ co-occurs with w_{hw} at position $m(-5 \leqq m \leqq 5)$. If the co-occurrence frequency at position m fulfills the following condition,

$$f(w_{hw}, w_{co\text{-}i}, m) \geq \overline{f(w_{hw}, w_{co\text{-}i})} + \sqrt{\frac{1}{10} \cdot \sum_{j=-5}^{5} (f(w_{hw}, w_{co\text{-}i}, j) - \overline{f(w_{hw}, w_{co\text{-}i})})^2} \quad (3)$$

where, $\overline{f(w_{hw}, w_{co\text{-}i})} = \frac{1}{10} \cdot \sum_{j=-5}^{5} f(w_{hw}, w_{co\text{-}i}, j)$ is the average co-occurrence at each position. This bigram is considered to have a peak co-occurrence at position m. In other words, if a bigram co-occurs at one position over the average co-occurrence frequency in [–5, +5] positions for one standard deviation, the position is regarded as a peak. If a word bi-gram has only one peak, it is regarded as non-modifiable while two or more peaks indicate that it is more modifiable.

Measures for order altering

The ratio of the word bigram co-occurring in different order is used as the feature for estimating the property of order altering. For a bigram (w_{hw}, $w_{co\text{-}i}$), if w_{hw} occurs before $w_{co\text{-}i}$ within the same sentence for $f_{before}(w_{hw}, w_{co\text{-}i})$ times in the corpus while $f_{after}(w_{hw}, w_{co\text{-}i})$ times w_{hw} occurs after $w_{co\text{-}i}$, the order alerting ratio of w_{hw} and $w_{co\text{-}i}$ is defined as,

$$OrderAlteringRatio(w_{hd}, w_{co\text{-}i}) = \frac{|f_{before}(w_{hw}, w_{co\text{-}i}) - f_{after}(w_{hw}, w_{co\text{-}i})|}{f_{before}(w_{hw}, w_{co\text{-}i}) + f_{after}(w_{hw}, w_{co\text{-}i})}. \quad (4)$$

The value of $OrderAlteringRatio(w_{hw}, w_{co\text{-}i})$ ranges from 0 to 1, and the larger value means w_{hw} and $w_{co\text{-}i}$ tends to co-occur in one fixed order.

Measures for co-occurrence significance

For k unique co-words of the headword w_{hw}, labeled as $w_{co\text{-}i}$, (i = 1 to k) their average co-occurrence frequency is calculated as,

$$\overline{f(w_{hw})} = \frac{1}{k}\sum_{i=1}^{k} f(w_{hw}w_{co\text{-}i}) \quad (5)$$

and their standard deviation of the co-occurrence frequency, labeled as $\sigma(w_{hw})$ is calculated as,

$$\sigma(w_{hw}) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(f_{co\text{-}i} - \overline{f(w_{hw})})^2}. \quad (6)$$

Similarly, by taking $w_{co\text{-}i}$ as the headword, $\overline{f(w_{co\text{-}i})}$ and $\sigma_{co\text{-}i}$ can be obtained. $BI\text{-}Strength(w_{hw}, w_{co\text{-}i})$ is designed to bi-directionally measure the co-occurrence frequency significance between w_{hw} and co-word $w_{co\text{-}i}$ (i = 1 to k), by using both w_{hw} and $w_{co\text{-}i}$ as the headword, respectively. It is defined as

$$BI\text{-}Strength(w_{hw}, w_{co\text{-}i}) = 0.5 \cdot \frac{f(w_{hw}w_{co\text{-}i}) - \overline{f(w_{hw})}}{\sigma(w_{hw})} + 0.5 \cdot \frac{f(w_{hw}w_{co\text{-}i}) - \overline{f(w_{co\text{-}i})}}{\sigma(w_{co\text{-}i})}. \quad (7)$$

A larger value of $BI\text{-}Strength(w_{hw}, w_{co\text{-}i})$ indicates the co-occurrence of w_{hw} and $w_{co\text{-}i}$ is more significant than w_{hw} with its other co-words and also more than $w_{co\text{-}i}$ with its other co-words. Thus it indicates a stronger internal association between w_{hw} and $w_{co\text{-}i}$.

A small set of collocations are manually classified which is based purely on expert knowledge. The computational measurements, which are proposed above, corresponding to these classified collocations are analyzed to find out an automatic collocation classification strategy with the reference of Table 1 and discussion in Section 2. The use of computational measures and automatic classification strategy are helpful to collocation type determination. In this study, the same association measurement program is applied to obtain the values of

computational features based on two sets of data. The first set of data set is the annotated dependencies in the 5-million-word corpus which is obtained through Pass 1 annotation. Because the dependent word pairs are manually verified and annotated in Pass 1, the accurate statistical significance is helpful to obtain accurate annotation. However, data sparseness problem must be considered since 5-million-word is not large enough. Thus, another set of data is a 100-million-word segmented and tagged corpus [13]. With this large corpus, data sparseness is no longer a serious problem. But the collected statistics are quiet noisy since they are directly retrieved from the text without any verification. By observing the statistical features coming from these two sets of data, the annotators can use their professional judgment to determine whether a bigram is a collocation and its collocation type if applicable.

5. Current Status and Evaluations

5.1. Current status of *CCB*

The annotation of the first version of *CCB* is completed. A total of 23,581 bigram collocations and 2,752 n-gram collocations are identified corresponding to the 3,643 selected headwords. Their occurrences in the corpus are verified and annotated. With the help of the assistive tool, the bigram collocations are manually classified into three types. The numbers of annotated Type 0/1 collocations, Type 2 collocations and Type 3 collocations are 152, 3982 and 19447, respectively. It is obvious that there are much more Type 3 collocations than the other types.

The annotation quality is estimated by checking the 5% duplicate annotations. These annotations are compared among all annotated. The collocation identification accuracy of each annotator's work is given in Table 4.

It is shown that the average accuracy is 97.1%, with the low and high ranges from 95.6 to 98.2 showing that n-gram collocations are much easier to identify. This is because the restrictions to n-gram collocations is tight which effectively reduced difficulty for identification. Most annotation differences

Table 4. The annotation accuracies by different annotators.

	Annotator 1 (%)	Annotator 2 (%)	Annotator 3 (%)	Average (%)
n-gram	98.2	97.4	95.6	97.1
bigram	94.4	93.4	92.3	93.4

occurs when deciding the boundaries. For example, two annotators identified 中东/j 和平/n 进程/n (*Middle East Peace Process*) as an n-gram collocation while another annotator identified 推动/v 中东/j 和平/n 进程/n (*Push forward the Middle East Peace Process*) as an n-gram collocation. As for the bi-gram collocation identification, most differences attribute to the fact that some word bigrams have weak association significance. Some annotators regarded them as collocation while some did not. The bigram collocation identification and classification accuracies by different annotators are further evaluated. The achieved accuracies on bigram collocation classification are given in Table 5.

It shown that the accuracy of annotation decreases as the restrictions on collocations are looser. Both Type 0/1 and Type 2 annotation have achieved very good accuracy with Type 0/1 having the best accuracy. Further observations show that many Type 2 bigram collocations are wrongly classified as Type 3. Type 3 collocations is more difficult as reflected in the accuracy of two annotators. It is more difficult to distinguish Type 3 bigram collocations because they have high-frequency co-occurrences yet are weak in semantic association.

Corresponding to 3,643 headwords, 23,581 bigram collocations are annotated in the collocation bank. The collection of bigram collocations in the collocation bank is named as *Chinese Collocation Bank Bigram part* (*CCB_B*). For the same 3,643 headwords, *The Dictionary of Modern Chinese Collocations* [9] provided 35,742 bigram collocations in which 20,035 appear in the People's Daily corpus. This collection of bigram collocations is named *Mei's Collocation Collection Bigram part* (*MCC_B*). There are 19,967 common entries in *MCC_B* and *CCB_B*, which means 99.7% collocations in *MCC_B* appear in *CCB_B* indicating good linguistic consistency between *MCC_B* and *CCB_B*. Furthermore, 3,614 additional collocations are found in *CCB_B* which means 20% of the collocations are new. The new knowledge on Chinese collocation provided by *CCB* is helpful to the research related to collocation. Meanwhile, this result shows that a static collocation lexicon is not enough, the effective automatic collocation extraction algorithm is desired.

Table 5. The bigram collocation annotation accuracies by different annotators.

	Annotator 1 (%)	Annotator 2 (%)	Annotator 3 (%)	Average (%)
Type 0/1	96.1	98.2	97.6	97.3
Type 2	95.2	96.1	94.2	95.2
Type 3	94.2	92.8	91.9	93.0

5.2. Dependencies statistics of collocations

Through statistical analysis on *CCB*, some characteristics of Chinese bigram collocations are observed. It is interesting to know that corresponding to the 23,581 collocations, 25,793 dependencies are identified. That is, a collocation pair can appear in the text under different dependency types and the diversity is more apparent for loose collocations as shown in Table 6 which lists the numbers of dependency types with respect to different collocation types.

It is observed that about 90% bigram collocations have only one dependency type. This indicates that a collocation normally has only one fixed syntactic relation. The other 10% bigram collocations have more than one dependency type and are mostly Type 3 collocations. For example, two types of dependencies are identified in the bigram collocation 安全/an-国家/n which can either be 安全/an-AN-国家/n (*a safe nation*) to indicate the dependency of a noun and its adjective modifier where 国家/n serves as the head, or 国家/n-NN-安全/an (*national security*) to indicate the dependency of a noun and its nominal modifier where 安全/an serves as the head. It is because the use of Chinese words is flexible. For example, an adnoun sometimes serves as an adjective and sometimes as a noun. Furthermore, a collocation with different dependencies leads to different distribution trends and thus, most of these collocations are classified as Type 3.

More detailed analysis of the 10 types of dependencies with respect to different types of collocations are given in Table 7. Each row in the table shows statistics of a particular dependency type. For each type of collocation *Count* is the total number of collocations. *P_T* shows its percentage with respect to this type of collocation and *P_D* shows its percentage with respect to this type of dependency.

Among all the collocations, about 82% belongs to five major dependency types. They are *AN* (*a noun and an adjective modifier*), *VO* (*predicate and its*

Table 6. Collocation classification versus number of dependency types.

Collocation type	One type of dependency	Two types of dependencies	More than two types of dependencies	Total
Type 0/1	152	0	0	152
Type 2	3,970	12	0	3,982
Type 3	17,282	2,130	35	19,447
Total	21,404	2,142	35	23,581

Table 7. The statistics of collocations with different collocation type and dependency.

	Type 0/1 collocations			Type 2 collocations			Type 3 collocations			Total	
	Count	*P_T*	*P_D*	*Count*	*P_T*	*P_D*	*Count*	*P_T*	*P_D*	*Count*	*P_T*
AN	20	13.2	0.4	871	21.8	15.4	4771	22	84.3	5662	22
VO	26	17.1	0.5	652	16.3	12.5	4545	21	87	5223	20.2
NN	44	28.9	0.9	1036	25.9	21.6	3722	17.2	77.5	4802	18.6
ADV	9	5.9	0.3	322	8.1	11.2	2555	11.8	88.5	2886	11.2
SBV	4	2.6	0.2	285	7.1	11.1	2279	10.5	88.7	2568	10
ADA	1	0.7	0.1	212	5.3	11.5	1637	7.6	88.5	1850	7.2
VV	3	2	0.2	227	5.7	13.4	1464	6.8	86.4	1694	6.6
CMP	12	7.9	2.2	144	3.6	26.9	379	1.8	70.8	535	2.1
OT	25	16.4	7.7	203	5.1	62.5	97	0.4	29.8	325	1.3
NJX	8	5.3	3.2	42	1.1	16.9	198	0.9	79.8	248	1
Total	152	100.0	0.6	3994	100.0	15.5	21647	100.0	83.9	25793	100.0

object), *NN* (*a noun and its nominal modifier*), *ADV* (*predicate and its adverbial modifier*) and *SBV* (*predicate and its subject*). It is noteworthy that there are a lot of *NN* collocations because nouns are often used in parallel to serve as one syntactic component in Chinese sentences which is different from English.

The collocations with different types of dependencies have shown their own characteristics with respect to different collocation types. The collocations with *CMP* (*Predicate and its complement*), *NJX* (*Juxtaposition structure*) and *NN* dependencies on average have higher percentage to be classified into Type 0/1 and Type 2 collocations. This is because *CMP*, *NJX* and *NN* collocations in Chinese are always used in fixed patterns and these kinds of collocations are not freely modifiable and substitutable. On 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 the auxiliary word 的/u and *ADV* collocations by the auxiliary word 地/u, which means that they always have two peak co-occurrences. They are thus classified as Type 3 collocations. Furthermore, 7.7% and 62.5% of the collocations with dependency *OT* (*others*) are classified as Type 0/1 and Type 2 collocations, respectively. This attributes to the fact that many Type 0/1 and Type 2 collocations have strong semantic relations rather than syntactic dependency (their dependencies are difficult to be labeled by any scoped syntactic dependency and thus, it is labeled as *OT*).

Table 8. Chunking distances of Type 3 collocations with different dependencies.

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

5.3. Chunking statistics of collocations

As mentioned before, Type 0/1 and Type 2 collocations have only one or two peak occurrences with very low order altering ratio and significant *CrowdingLevel*. Therefore, they either co-occur within one chunk or between neighboring chunks. It is more interesting to give some detailed analysis for Type 3 collocations since collocation pairs can often co-occur in long distant chunks as shown in Table 8. Chunking distance refers to the distance between collocated words with respect to chunks where they are located. The value of 0 is within the same chunk; 1 is in adjacent chunks, 2 or more means that they are in distant chunks.

It is shown that the number of Type 3 collocations decreases when chunking distance increases. Yet, the patterns for different dependency types are different. It is observed that roughly more than 94% of *OT* (99.9%), *ADV* (96.8%), *ADA* (95.0%), and *AN* (94.2%) collocations co-occur within the same chunk or neighboring chunks. On the other hand, more than 10% of *VO* (23.2%), *NJX* (14.4%), *CMP* (14.3%), *SBV* (12.3%) and *VV* (11.7%) collocations have chunk distance of 2 or more. Especially, the chunking type distribution of *VO* collocations is much flatter which is consistent with the understanding of the verb/predicate relationships which tends to have longer distance than others.

6. Applications of Collocation Bank

An experiment is conducted to evaluate the effectiveness of using *CCB* to improve automatic collocation extractions. Xu and Lu has developed a multi-stage collocation extraction system to identify collocations of different types using a combination of discriminative features and criteria associated with different types [13]. In this system, Type 3 collocations are identified by a linear multinomial classifier based on the combination of six statistical features and

its weights were optimized by Perceptron training. *CCB_B* can be directly used as training data where annotated bi-grams are considered as positive examples. All other word pairs not being annotated are considered as negative examples. The threshold values of the six internal strength measures (proposed in [13]) are optimized based on Perceptron classifier. As the training data used in [13] does not have dependency verification, it is regarded as a baseline system labeled as $T(0)$ where 0 indicates it has zero percentage of verified training samples. *CCB_B* is divided into five equal parts. One part is reserved for open test while the other four parts were incrementally used to update $T(0)$ by using the verified co-occurrence of collocations, labelled $T(20)$, $T(40)$, $T(60)$ and $T(80)$, respectively. Figure 1 shows the effect of training iterations $w(i)$ with respect to precision for the 5 sets of training data with verified training samples from 0% to 80%. The learning rate is set to 0.05 which is an empirically obtained learning rate. It can be seen that training started to converge after 400 training circles.

Figure 1 shows that the difference in performance of the algorithms gives 68.1% precision using verified data, $T(80)$, yet the same algorithm using a completely unverified data, $T(0)$, only achieves 76.3% precision. In other words, the system using annotated collocation bank can give (76.3−68.1)/68.1% = 12.0% improvement in performance for a collocation extraction algorithm. This is a good indication that an annotated collocation bank can be helpful for collocation extraction algorithms.

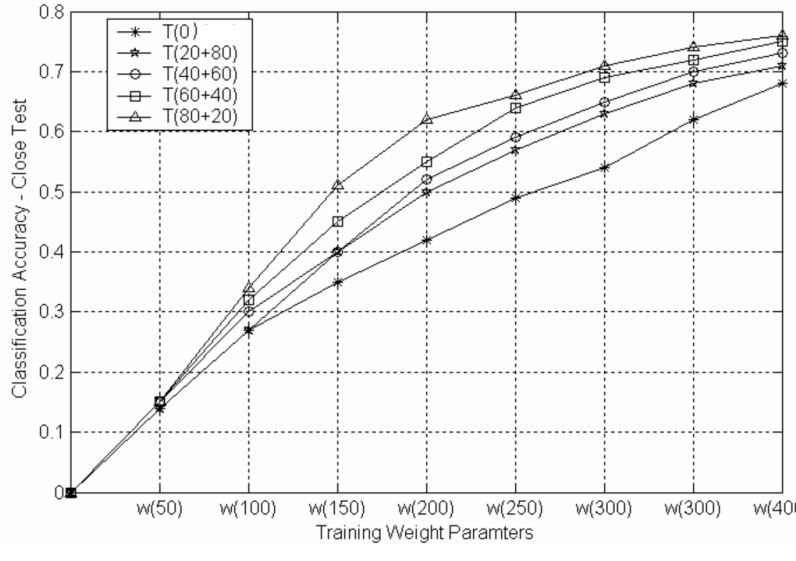

Figure 1. The classification accuracies with different weights.

7. Conclusions

This paper describes the design and construction of a Chinese collocation bank. This is the first attempt to construct such a linguistic resource. Following a well-designed guideline, collocations corresponding to 3,643 selected headwords are identified from a chunked 5-million-word corpus. A total of 2,752 n-gram collocations and 23,581 bigram collocations are annotated. For each annotated bi-gram collocation, three kinds of information are provided including syntactic dependency information which describes a direct relationship between two words at the word level, the chunking information which describes the relationship between the collocated words and their context words at the chunk level, and the classification information which describes the strength of internal associations of the collocations. Characteristics of Chinese collocations with different types and different dependencies in the collocation bank are observed.

The annotated collocation bank will be used to analyze the characteristics for different types of collocations, and to select the appropriate discriminative features for automatic collocation extraction algorithms. Furthermore, the collocation bank can also be used as a standard answer set for evaluating the performance of collocation extraction algorithms.

Acknowledgments

This research is partially supported by the Hong Kong Polytechnic University (Project Code A-P203), a CERG Grant (Project code 5087/01E) , The Chinese University of Hong Kong under the Direct Grant Scheme project (2050330) and Strategic Grant Scheme project (4410001), and Post-doctoral research funding from the City University of Hong Kong.

References

[1] J. R. Firth, *Papers in Linguistics*, Oxford University Press, 1957, pp. 55–62.

[2] K. W. Church, A stochastic parts program and noun phrase parser for unrestricted text, in *Proc. 2nd Conf. on Applied Natural Language Processing,* 1988, pp. 136–143.

[3] J. Sinclair, *Corpus*, *Concordance*, *Collocation*, Oxford University Press, 1991.

[4] M. Mitra, C. Buckley, A. Singhal and C. Cardie, An analysis of statistical and syntactic phrases, in *Proc. 5th Int. Conf. "Recherche d'Information Assistee par Ordinateur"* (*RIAO*), 1997, pp. 200–214.

[5] F. Smadja, Retrieving collocations from text: Xtract, *Computational Linguistics,* 19, 1993, 143–177.
[6] M. Benson, Collocations and general-purpose dictionaries, *International Journal of Lexicography*, 3, 1990, 23–35.
[7] K. W. Church and P. W. Hanks, Word association norms, mutual information and lexicography, *Computational Linguistics*, 16, 1990, 22–29.
[8] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.
[9] J. J. Mei, *Dictionary of Modern Chinese Collocations*, Hanyu Dictionary Press, 1999.
[10] D. K. Lin, Extracting collocations from text corpora, in *Proc. 1st Workshop on Computational Terminology*, 1998.
[11] R. F. Xu and Q. Lu, Improving collocation extraction by using syntactic patterns*,* in *Proc. IEEE Int. Conf. on Natural Language Processing and Knowledge Engineering*, 2005, pp. 52–57.
[12] P. Pecina, An extensive empirical study of collocation extraction methods*,* in *Proc. ACL 2005 Student Research Workshop*, 2005, pp. 13–18.
[13] R. F. Xu and Q. Lu, *A Multi-stage Chinese Collocation Extraction System*. Lecture Notes in Computer Science, Vol. 3930, Springer-Verlag, 2006, pp. 740–749.
[14] S. Kosho, T. Masahito, K. Yasuo and Y. Kenji, Collocations as word co-occurrence restriction data — An application to Japanese word processor, in *Proc. 2nd Int. Conf. on Language Resources and Evaluation* (*LREC* 2000), 2000.
[15] N. B. Atalay, K. Oflazer and B. Say, The annotation process in the Turkish treebank, in *Proc. 11th Conf. EACL — 4th Linguistically Interpreted Corpora Workshop*, 2003.
[16] M. Gross, *Méthode en syntaxe*, Hermann, Paris, 1975.
[17] A. Tutin, Annotating lexical functions in corpora: Showing collocations in context, in *Proc. 2nd Int. Conf. on the Meaning — Text Theory,* 2005.
[18] J. C. Richards and R. Schmidt, *Longman Dictionary of Language Teaching and Applied Linguistics*, Longman Group UK Limited, 1992.
[19] D. J. Allerton, Three or four levels of co-occurrence relations, *Linguistics*, 63, 1984, 17–40.
[20] K. McKeown and D. Radev, Collocations, in R. Dale, H. Moisl and H. Somers (eds.), *Handbook of Natural Language Processing*, Marcel Dekker, New York, 2000, pp. 385–401.

[21] J. Brundage, M. Kresse, U. Schwall and A. Storrer, *Multiword Lexemes: A Monolingual and Contrastive Typology for Natural Language Processing and Machine Translation*, Technical Report 232 — IBM, 1992.
[22] S. W. Yu *et al.*, *Guideline of People's Daily Corpus Annotation*, Technical Report, Peking University, 2001.
[23] M. S. Sun, J. Fang and C. N. Huang, A preliminary study on the quantitative analysis on Chinese collocations, *Chinese Linguistics*, 1, 1997, 29–38.