Design of Chinese Morphological Analyzer

Huihsin Tseng
Institute of Information Science
Academia Sinica, Taipei
kaori@hp.iis.sinica.edu.tw

Keh-Jiann Chen
Institute of Information Science
Academia Sinica, Taipei
kchen@iis.sinica.edu.tw

Abstract

This is a pilot study which aims at the design of a Chinese morphological analyzer which is in state to predict the syntactic and semantic properties of nominal, verbal and adjectival compounds. Morphological structures of compound words contain the essential information of knowing their syntactic and semantic characteristics. In particular, morphological analysis is a primary step for predicting the syntactic and semantic categories of out-of-vocabulary (unknown) words. The designed Chinese morphological analyzer contains three major functions, 1) to segment a word into a sequence of morphemes, 2) to tag the part-of-speech of those morphemes, and 3) to identify the morpho-syntactic relation between morphemes. We propose a method of using associative strength among morphemes, morpho-syntactic patterns, and syntactic categories to solve the ambiguities of segmentation and part-of-speech. In our evaluation report, it is found that the accuracy of our analyzer is 81%. 5% errors are caused by the segmentation and 14% errors are due to part-of-speech. Once the internal information of a compound is known, it would be beneficial for the further researches of the prediction of a word meaning and its function.

1. Introduction

This is the first attempt to design a morphological analyzer to automatically analyze the morphological structures of Chinese compound words[1]. Morphological structures of compound words contain the essential information of knowing their syntactic and semantic characteristics. In particular, morphological analysis is a primary step for predicting the syntactic and semantic categories of out-of-vocabulary (unknown) words. The existence of unknown words is a major obstacle in Chinese natural language processing. Due to the fact that new words are easily coined by morphemes in Chinese text, the number of unknown words is increasingly large. As a result, we cannot collect all the unknown words and manually mark their syntactic categories and meanings. Our hypothesis to predict the category and the meaning of a word is basically based on Frege's principle: "The meaning of the whole is a function of the meanings of the parts". The meanings of morphemes are supposed to make up the meanings of the words. However, some words like idioms and proper nouns cannot be included in the principle. In general, unknown words could be divided into two different types: the type that has the property of semantic transparency, i.e. the words whose meanings can be derived from their morphemes and the type without meaning transparency, such as proper nouns. In this paper we are dealing with the compound words with semantic transparency only. For the type of compounds without semantic transparency, such as proper nouns, their morphemes and morphological structures do not provide useful information for predicting their syntactic and semantic categories. Therefore they are processed differently and independently. In addition, some regular types of compounds, such as numbers, dates, and determinant-measure compounds, are easily analyzed by matching their morphological structures with their regular expression grammars and the result can be used to predict their syntactic and semantic properties, so they will be handled by matching regular expressions at the stage of word segmentation. According to our observation, most Chinese compounds have semantic transparency except proper nouns, which means the meaning of an unknown word can be interpreted by their own morpheme components. The design of our morphological analyzer will focus on processing these compounds, but words without semantic transparency are excluded. It takes a compound word as input and produces the morphological structure of the word. The major functions are 1) to segment a word into a sequence of morphemes, 2) to tag the part-of-speech of those morphemes, and 3) to identify the

[1] Compound words here include compounds in traditional Chinese linguistics and morphological complex words.

morpho-syntactic relation between morphemes. Once the morpho-syntactic structure of a compound is known, the head morpheme provides the major clue for determining its syntactic and semantic category.

It seems that a Chinese morphological analyzer is similar to a tagging program. Indeed both systems have to resolve the segmentation and tagging ambiguities. However the major difference is that the morphological analyzer does not have contextual information of each target word. In other words, morphological structures of compounds are context independent. We cannot apply the same methods, such as n-gram language models, to resolve the ambiguities. We proposed a method of using the associative strength among morphemes, morpho-syntactic patterns, and syntactic categories to solve the ambiguities. Detail algorithms for morpheme segmentation, part-of-speech, and morpho-syntactic relation assignment are discussed in Section 2. In the final section, we will evaluate the morphological analyzer by comparing its results with those obtained from the analyses of 5 linguists and discussing the categorization of errors found.

2. The Morphological Analyzer

The morphological analyzer contains three functions: to segment a word, to tag the part-of-speech (POS) of morphemes, and to identify the relation between them.

2.1 Segmentation

The goal of this process is to segment a compound word into a sequence of morphemes. Since there are ambiguous segmentations, simple dictionary look-up methods may not work well. For instance, the compound of *meiguoren* (美國人) could be ambiguously segmented into either *mei-guoren* ([美[國人]] beautiful countryman) or *meiguo-ren* ([[美國]人] American people), but only *meiguo-ren* ([[美國]人] American-people) is the proper segmentation. The left-to-right longest matching method is commonly applied to segment either words or text. It works well, but there are still some small percent of compound words that cannot be properly segmented by such a simple algorithm. For instance, the word *xin-shenghuo* ([新[生活]] new life) will be segmented wrongly into *xinsheng-huo* ([[新生]活] the life of a new student) without considering the priority of segmenting the affix *xin* (新 new) first. In particular, words with multi-syllabic suffixes and words with reduplication constructions commonly cause segmentation errors. Those special types of words should be analyzed with other methods.

2.1.1 Affixes and reduplication

In order to remedy the segmentation error caused by the left-to-right longest matching, we observe the results of the algorithm and find that there are two useful clues to avoid segmentation errors, i.e. the information of affix and reduplication.

A word of a reduplication pattern cannot and need not be segmented by the longest left-to-right method, since it has special morphological structures and the reduplication patterns bring enough clues of knowing the syntactic functions of the word. Therefore we try to identify words that belong to reduplication patterns first. In general they fall into the following two types of patterns: reduplications and parallel words. Words, which do not conform to these patterns, will be segmented later.

Table 1 Special types of patterns and their examples

Patterns	Pattern Maker	Note and examples
Reduplication word	AA, AAA, ABB, AAB, AxA, ABAB AABB, AxAy, xByB,	*liang-liang* (亮亮), *dui-dui-dui* (對對對), *song-kua-kua* (鬆垮垮), *chi-chi-kan* (吃吃看), *xiang-yi-xiang* (想一想). *yan-jiu-yan-jiu*(研究研究), *chi-chi-he-he*(吃吃喝喝), *pao-shang-pao-xia*(跑上跑下), *yi-nian-zai-nian*(一唸再唸)
Parallel word	A-BC (AC, BC)	zhong-xiao-xue (中小學)

Reduplication means to duplicate the one or two character words into multi-character words. All reduplication patterns we used are listed in Table 1. If a word belongs to a reduplication pattern, the meaning of the word doesn't change too much. The reduplication word's category can be predicted by their patterns. For example, when B is not a noun, a word which belongs to the pattern AAB is intransitive verb. The category of a word that belongs to the pattern of "parallel word" is always a noun. The characteristic of parallel words is that both AC and BC are words with shared head word C.

At the next step of the morpheme segmentation, we will consider the compounds with affixes. The most productive compound construction is the structure of a morpheme plus an affix. Hence after the affix is identified, it would be easier to segment a word into two parts. The segmentation algorithm works as follows. A word is segmented immediately only if a prefix, infix or suffix morpheme is found. The affix table contains 186 prefixes, 2 infixes and 648 suffixes. Some affixes of the table are multi-syllabic. To segment an affix with higher priority will resolve most of the errors caused by the left-to-right longest matching algorithm. For instance, if *tiaoshangqu* (跳上去 to jump up) is segmented by the left-to-right longest matching method, and the result of the segmentation is *tiaoshang-qu*. The left-to-right longest matching method might cause error segmentation here. However, *shangqu* is one of the suffixes in the affix table, so in our morphological analyzer it would be segmented as *tiao-shangqu*. A word containing an infix is also not suitable for the general segmentation and it would be segmented into single character. There are some affixes examples in Table 2:

Table2 Types of affixes and their examples

Types of affix	Morpheme	Examples
Prefix	*xin*(新)	*xinsheng-huo* (新生活)
Infix	*de*(得)	*suan-de-shang*(算得上)
Suffix	*ju*(劇) *shangqu* (上去)	*feizao-ju*(肥皂劇) *tiao-shangqu* (跳上去)

2.1.2 Left-to-right longest matching

If a word is neither reduplication nor a compound with an affix, it should be segmented from left to right with longest matching. This general method can segment words into morphemes and also provide a possible part-of-speech of each morpheme by looking it up in the morpheme dictionary.

2.2 Tagging

The work here is to provide the part-of-speech for each morpheme and identify a morpho-syntactic relation between two morphemes based on the information of segmentation and their pos. This is the most difficult part of morphological analysis. In achieving the goal, we face two obstacles: the information insufficiency of morpheme categories and morphemes with the multiple categories. Since morpheme categories are not the same as word categories, it is necessary to assign each morpheme with appropriate categories and to compile a morpheme dictionary. Once the morpheme dictionary is built, the remaining job is to resolve the part-of-speech ambiguities of each morpheme. Since the part-of-speech of the morpheme is independent of its word level context, we cannot apply n-gram like language models to resolve part-of-speech ambiguity of morphemes. Even worse, there is no structure tagging training data available either. An EM-like unsupervised training on part-of-speech morphological structures is also not a sensible solution, since morpho-syntactic structure is more sensitive to the semantic combination than the syntactic combination of morphemes. Therefore we propose a method of using morphemes to predict the possible syntactic categories of the target compound word and selecting the most probable consistent result among the candidates of part-of-speech structures and the predicted categories.

2.2.1 Preparation of the morpheme dictionary

Before we start to tag morphemes, two steps are carried out to resolve the obstacles. That is the lack for a morpheme dictionary and morpheme ambiguity. First, in order to resolve the lack for morpheme categories, it is necessary to edit an affix table, as mentioned in Section 2.1, which contains prefixes, infixes, suffixes and their categories. Most frequently encountered 186 prefixes and 648 suffixes are listed in this table. Basically, if its morpheme has more than 2 characters, we adopt its categories in the CKIP Dictionary. Conversely, if it has fewer than 2 characters and it functions as a prefix or a suffix, we use the categories in the affix table.

Below we illustrate two examples to explain the need of morpheme categories. Both words *yu* (語 to speak) and *wu* (舞 to dance) are verbs. However they could also function as morphemes. When they function as morphemes, they are listed as nouns in their category. It is worth noticing that the categories of a morpheme are not the same as those of a word, even if they are in the same form. Therefore, it is important to assign morphemes categories properly.

Table 3 The categories of –*yu* and –*wu* as a suffix and as a word and their examples

Suffix	Category[2] as a suffix	Category as a word	Example
-*yu*(語)	Na	VE	*ying-yu*(英語), *de-yu*(德語)
-*wu*(舞)	Na	VC, Na	*jueshi-wu*(爵士舞)

[2] The category symbols here are based on CKIP(1993). The meaning of each category we adopt here is as following: A(non-predicative adjective), Na(common noun), Nb(proper noun), Nc(location noun), Nd(time noun), VA(active intransitive verb), VB(semi-transitive verb), VC(active transitive verb), VCL(active transitive verb with locative object), VD(ditransitive verb), VE(active transitive verb with sentential object), VG(classificatory verb), VH(stative intransitive verb), VHC(stative causative verb), VJ(stative transitive verb) and so on.

Second, in order to resolve the problem of morpheme ambiguity, we need a list of probabilities which contains all the possible combinations of categorical rules and their probabilities. For instance, in the list the probability P(Na+Na|Na) = 0.4692 means that the categorical rule of combining two common nouns (Na+Na) to form a common noun (Na) has the probability 0.4692. The probability values of each categorical rule were estimated from the set of 11,322 unknown words extracted from Sinica Corpus 3.0. The syntactic category of each extracted word is known but its structure annotation is unknown. Therefore the probability of each categorical rule is roughly estimated by assuming that every ambiguous categorical combination of a word have equal probability.

The process of computing the possibility of a combination is as follows:

1) We assign morphemes in a word with all their possible categories found in the dictionary and the affix table; for example, *sheyingzhan* ([[攝影]展] photography exhibition), which belongs to Na category means “photography exhibition”. *Sheyingzhan* could be segmented as *sheying* (攝影 photography) and *zhan* (展 exhibition). After segmentation, we found *sheying* with the categories Na and VA, and *zhan* with the category Na. The possible combinations of *sheying-zhan* are “Na+Na->Na” and “VA+Na->Na”. However, we don’t know which one is correct, so we presumably assign a frequency of 0.5 to each combination.

2) After we assign morphemes their categories and frequencies, we add up the frequencies of identical combinations. A list containing possible categorical rules and their probabilities is then established. Table 4 shows a part of the categorical rules of VHC.

Table 4 A partial list of categorical rules and their probabilities

Rule	Category	Probabilities
Na+VHC	VHC	0.4494
VH+VHC	VHC	0.2303
Nc+VHC	VHC	0.0674
VHC+VHC	VHC	0.0449
VA+VH	VHC	0.0280
VC+VHC	VHC	0.0224
VJ+VH	VHC	0.0224
Nd+VHC	VHC	0.0112
VC+Na	VHC	0.0112
VC+VC	VHC	0.0112
VC+VHC	VHC	0.0112

2.2.2 Part-of-speech

Once the affix table and the list of categorical rules are prepared, we can tackle the problems of the obstacles we mentioned in the beginning. After morpheme segmentation, each morpheme is assigned with their proper categories according to the morpheme dictionary and the affix table. However, morphemes might be ambiguous, so if the category of the target word is known, the most probable part-of-speech combination is chosen based on the list of categorical rules. However in the real implementation, it is assumed that the syntactic category of a target word is not known. The method mentioned above would not work, unless its syntactic category can be predicted. In our implementation, we adopted the method proposed by Chen, Bai and Chen (1997), by using the association strength between morphemes and categories to predict the syntactic categories of target words. By using the mutual information between affixes and categories, the top one prediction has the accuracy of 67.00% and the top three accuracy of the prediction can reach about 94.02%. We will then check the consistency between predicted the categories and their morpho-syntactic structures to make the final judgments on both the word category prediction and the morpheme category disambiguation. The final prediction is based on the maximal value of the combined probabilities of the category prediction and the categorical rule prediction.

Since P(Rule|compound) = P(Cat|compound) * P(Rule|Cat, compound) ≅ P(Cat|compound) * P(Rule|Cat), we try to find Cat and Rule which maximizes P(Cat$_i$|compound) * P(Rule$_j$|Cat$_i$), for all Cat$_i$ and Rule$_j$. The following is an example of *.she-ying-zhan.*

=====================================

sheying-zhan (攝影展 photography exhibition)

P(Na|*sheying-zhan*) *P(Na+Na|Na)= 0.6324*0.4692=0.2967
---max
P(Na|*sheying-zhan*) * P(VA+Na|Na)=0.6324 *0.0865=0.0547
P(VC|*sheying-zhan*)* P(Na+Na|VC)=0.3675* 0.0069=0.0025
P(VC|*sheying-zhan*)*P(VA+Na|VC)= 0.3675* 0.001=0.0003

sheying-zhan=(Na+Na)->Na

=====================================

The top1 accuracy of the original category prediction for unknown words is 67% by mutual information, but after the combination of the morphological analyzer, the accuracy of the word category prediction is raised to 71%. This is because the morphological analyzer will check if the categorical combination in a word is in its proper category. Therefore, when the original unknown word prediction system predicts a word in a category which the morphological analyzer finds the probability of its categorical combination in the category low, the morphological analyzer might reject the category and suggest the unknown word prediction system to choose the next highest-scoring category in which the categorical combination has higher probability.

In the case that the syntactic category of the compound word is known, we will let P(Cat|compound) = 1 and the most probable part-of-speech combination will simply be the categorical rule Rule_j such that P(Rule_j|Cat) is maximized.

2.2.3 Morpho-syntactic relation between morphemes

Once the information of segmentation and part-of-speech is ready, the morpho-syntactic relation between morphemes can be identified. According to Chao (1968) and Li&Thompson (1981), there are relations between morphemes in compounds such as "modifier-hear", "predicate object" and so on. The purpose of knowing morpho-syntactic relation between morphemes is to help decide the meaning of the target word. The morpho-syntactic relation between morphemes is grouped into the types listed in Table 5. Generally, the relation between morphemes is highly related to the category of an unknown word. So the relation we assign to morphemes must be based on the category of the word. When the unknown word is a noun, the relation between its morphemes is "modifier-head". If it's a verb, it will be more complicated. There are five relation types in verbs. The first one is "verb-object", such as *chifan* (吃飯 to eat rice). The first morpheme must be a transitive one and the second one should be a noun. The second type of the relation is "modifier-head", and it means the second morpheme is the semantic head of the word. The third type is "resultative verb". The second morpheme in this type's word always expresses the result of the action. The forth type is "head-suffix". The appearance of the suffix changes the augment structure of the head verb, but the representing event remains the same. These suffixes are *ru* (如 to be similar to), *yu* (於 by), *wei* (為 to become), *gei* (給 to give), *chu* (出 to exit) and *cheng* (成 to become). The fifth type of the relation is "modifier-head", and there is only a morpheme *hua* (化 to transform) which belongs to this type. *Hua* is the head of a word. If a non-predicative adjective is an adjective, there are two kinds of structure. First, a non-predicative adjective has the same structure as a noun. The relation between its morphemes is also called "modifier-head". Second, the relation between morphemes for a non-predicative adjective which cannot be in the predicate position but has verbs structures can be "predicate-object" or "modifier head". This information will be helpful for predicting the core meaning of a new word.

Table 5 The morpho-syntactic relation between morphemes

	The morpho-syntactic relation between morphemes
Noun	Modifier-head
Verb	Verb-object Modifier-head Resultative Verb (RVC) Head-suffix Modifier-head(suffix)
Adjective	An: Modifier-head Av: verb-object, and modifier-head
Other	directional RVC and reduplication

Once the morpho-syntactic structure of a compound is identified, the head morpheme provides the major clue for determining its syntactic and semantic category. The compound word will inherit from the semantic and syntactic property of its head and the information will be beneficial for the semantic and syntactic categorization of new compound words in the future.

3. Evaluation and Discussion

The major functions of the morphological analyzer are to segment a word into a sequence of morphemes, to tag the part-of-speech of the morphemes, and to identify the morpho-syntactic relation between morphemes. The work in this section is to evaluate the quality of the word information which is processed by each function of the morphological analyzer. However, it is hard to evaluate the accuracy of the morphological analyzer automatically, so we compare the results generated by the morphological analyzer with results generated by human experts, which are made out of their language intuition. The answers agreed by the majority of the human experts are assumed to be the right answers. The closer the results of the morphological analyzer are to the human experts, the more accurate the morphological analyzer is.

The testing data is the set of unknown words extracted from the recently collected text by the system of Ma, Hsieh, Yang and Chen (2001). There is total 4,566 unknown words in our testing data. However, the validity of the morphological information is still uncertain; therefore five linguistic specialists have to manually verify the morphological structure of unknown words by filling out the survey. First, we randomly select 100 words as a testing set and the following three questions are answered by these five specialists.

1) What's the category of the unknown word?
2) What are the morpheme segmentations of the testing words?
3) What is the syntactic tag of each morpheme?

The definition of our "standard answer" is the answer the majority of the subjects give. For example, if three out of the five subjects consider the category of an unknown word X as VG, the standard answer of X would be VG. If five subjects think unknown word X belongs to five different categories, we would ask one more language specialist for opinions to determine the category of this unknown word. The standard answer we obtained from this survey will be the standard answer of the morphological analyzer.

The morphological analyzer contains three functions: to segment a word into morphemes, to tag pos, and to identify the relation between morphemes. The accuracy we mention here is the result from comparing the morphological result with the majority answer.

==

T=the total number of test set

R=the total number of being the same with the “Standard answer” of X

X=Subject1, Subject2, Subject3, Subject4, Subject5, Morphological Analyzer

Accuracy(X)= R(X)/T

==

Table 6 The accuracy of five subjects and morphological analyzer (MA)

	S 1	S 2	S 3	S 4	S 5	Average of 5 Ss	MA
Accuracy	89%	94%	94%	86%	83%	89.2%	81%

After comparing the result of the morphological analyzer with the standard answers obtained from the five linguists, we come to the conclusion that the accuracy of the morphological analyzer is 81%. Out of all the errors, 5% is caused by segmentation on proper nouns and loanwords, such as *bilinshan* (比林山 a name of a mountain), *dingwan (*鼎灣 a name of a place), *yanyou* (延祐 a name of a dynasty), *maniuda* (馬紐大 a name of a place), and *hongburang* (紅不讓 home run). These words cannot be segmented because they only make sense when they are treated as a unit. The remaining 14% is caused in the tagging process produced by a morpheme table which lacks in accuracy. For example, in some cases the suffix *zhou* (週 week) is supposed to be listed as Nd but is instead listed as Nc. Next, there are no proper categories for certain morphemes, such as the morpheme *lie* (列 to list as a verb, a row as a noun) in the word *qinglie* (清列). In the suffix table, the category of *lie* is only Na, but the morpheme *li*e should have a VC category when the meaning of *qinglie* "to list completely" is adopted. Another possible error-causing factor would be the choice made by following the combination rule. When there is more than one possible combination, errors might appear. For example, there are two possible combination for *waizhan* (外展 to stretch out), “Ng+Na” and “Ng+VC”. Comparing the score of the two combinations, the combination of “Ng+Na” is chosen. However, it is not the correct category of *zhan* (展 to stretch).

The best way to resolve these problems mentioned above is to revise the morpheme table more often. Since the category of the suffix and prefix is fixed, it might cause a reduction in morpheme ambiguity. We are also interested in the similarity (or the range of agreement with language intuition of each individual) between those subjects. Since the standard answers are the answer of the majority, we can compare the standard answer with each individual. The average rate of the similarity rate is 89.2%. The ten-percentage puzzle might be due to the ambiguity of the word and can be interpreted that there are indeed some words that are not only difficult for a machine to analyze but also difficult for human beings to categorize.

Table 7 The error rate and examples

	Percent-age	Examples
Segmentation Error	5%	*hongburang*(紅不讓), *maniuda*(馬紐大)
		yanyou(延祐), *dingwan*(鼎灣), *bilinsha*n (比林山)
Tagging Error	14%	*miho*u(獼猴) ***tao*(桃)(Nc),** *zixun(*資訊) ***zhou(*週)(Nc),** *wa*i(外) ***zhan*(展)(Na,VC)**

The evaluation of the identification of the morpho-syntactic relation is separated from the evaluation of segmentation and tagging, because the relation between morphemes is identified based on previous information, such as the category of a word, segmentation, and the pos. Once the essential information is clear, the morpho-syntactic relation is known. Nine out of a hundred examples are marked by linguists as errors of the morpho-syntactic identifier.

Furthermore, the reasons causing the error of relation identification are 1) the category predication's error, 2) the part-of-speech error, and 3) the lack of the relation type. Firstly, since the relation identifier is based on the result of the segmentation and pos, it is understandable that the error here is caused by previous functions. The category of *qipai-jia* ([[起拍]價] initial bidding price) is Na, but the system predicts it as an intransitive verb VA. So the identifier guesses the relation between *qipai and jia* as "verb-object" based on the previous information. The error of the category prediction system might result in errors of the relation identifier. Secondly, the relation of *qing-lie* (清列 to clearly list) should be "modifier-head", but the identifier marks it as "verb-object" relation because *lie*(列) is tagged as Na. When the suffix is a Na, the prefix is a verb, and since the category of *qing-lie* is predicted as a verb, the identifier can only predict the relation of *qing-lie* as "Verb-object". Therefore, the error of part-of-speech might cause the identifier errors. Thirdly, the linguists suggest the relation between morphemes in *nian-song* (念誦 to read) is "conjunction relation". That means that both the semantic meaning and syntactic function of *nian* (念 to read) and *song* (誦 to read) are the same. However, we don't have the "conjunction relation", because we think the number of words which belong to the kind of the relation is very limited, and since both morphemes the bring same information, there is no difference that enables us to mark both of them as heads or only one of them as a head for the application of predicating the semantic and syntactic property of a word. Therefore, in the morphological analyzer the words which belong to the "conjunction" relation are identified as "head-final" relations.

4. Conclusion and future work

This is a pilot study to design a morphological analyzer to analyze the morphological structures of Chinese compound words automatically. The major functions are 1) to segment a word into a sequence of morphemes, 2) to tag the part-of-speech of those morphemes, and 3) to identify the morpho-syntactic relation between morphemes. We evaluate the morphological analyzer by comparing 5 linguists' research results and discuss the type of errors we find. The more similar the results of the morphological analyzer compared with the human results, the better the morphological analyzer is. It is found that the accuracy of our analyzer is 81%. In comparison with the performance of human experts resulting in an accuracy of 89%, the performance of the current morphological analyzer is not bad, but still has room for improvement. More, the types and the identification of relations of morphemes still have much room to be improved. It is also worth noticing that the syntactic category prediction for general compounds can also be improved by the morphological analyzer. Once the internal information of a compound is known, it can provide clues for prediction of a word meaning and its function. The prediction of a word's meaning is very hard and will be one of the main themes in our future researches.

5 Reference

Bosch, Antal van den, Walter Daelemans and Ton Weijters. (1996) Morphological Analysis Classification: an Inductive-Learning Approach. *NeMLaP*.

Chao, Yuen Ren. (1968) *A grammar of spoken Chinese*. Berkeley:University of California Press.

Chen, Chao-jan, Ming-hung Bai and Keh-jiann Chen. (1997) Category Guessing for Chinese Unknown Words. *Proceedings of the Natural Language Processing Pacific Rim Symposium 1997*, 35-40.

Chen Yun-chai. (2001) *Corpus Analysis of Reduplication in Mandarin Chinese.* National Kaohsiung Normal University: English Department.

CKIP. (1993) *Technical Report no. 93-05: The analysis of Chinese category.* [中文詞類分析] CKIP:Nankang

Creutz, Mathias and Krista Lagus. (2002) Unsupervised Discovery of Morphemes. *Proceedings of Morphological and Phonological Learning Workshop of ACL'02.*

Beaney, Michael.(editor) (1997) *The Frege Reader.* Oxford: Blackwell.

Li, Charles and Sandra A. Thompson. (1981) *Mandarin Chinese*. Berkeley: University of California Press.

Ma, Weiyun, Youming Hsieh, Changhua Yang, and Keh-jiann Chen. (2001) "Chinese Corpus Development and Management System " [中文語料庫構建及管理系統設計]. *Proceedings of Research on Computational Linguistics Conference XIV*, 175-191.