DEFINING AND AUTOMATICALLY IDENTIFYING WORDS IN CHINESE

by

Nianwen Xue

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy with a major in Linguistics

Fall 2001

DEFINING AND AUTOMATICALLY IDENTIFYING WORDS

IN CHINESE

by

Nianwen Xue

Approved: _____

William Idsardi, Ph.D.
Acting Chair of the Department of Linguistics

Approved: _____

Conrado M. Gempesaw II, Ph.D.
Vice Provost for Academic Programs and Planning

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Peter Cole, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

William Idsardi, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Gabriella Hermon, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Martha S. Palmer, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: _____

Rolf Noyer, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

First of all, I would like to thank my committee members. Many thanks go to my advisor Peter Cole, without whose encouragement and support this dissertation would not have been possible. I owe Peter more than just the dissertation: I first learned to make syntactic argumentation from Peter and benefited immensely from his insistence on making one's arguments as clear as possible. I would also like to thank Martha Palmer, who not only gave me the opportunity to work on the Chinese Treebank Project at University of Pennsylvania and got me started to think about the issues discussed in this dissertation, but also guided me on the computational aspect of this dissertation. Thanks also go to Rolf Noyer, whose comments on the previous draft of this dissertation lead to substantial improvements. His comments also corrected my misinterpretations of certain parts of the Distributed Morphology theory, the theoretical framework adopted in this dissertation. I learned a great deal from the advanced syntax seminars taught by Gaby Hermon. I benefited from Bill Idsardi's linguistic expertise in general and I would like to thank him for his patience with me in my earlier years in the program before I settled down with a research program.

Besides my committee members many people have contributed to my education here in the United States. I would like to thank Jane Creswell, our department coordinator, for help and numerous favors. I would like to thank Richard Venezky for helpful advice and making available a quiet and spacious lab for me to work in. I would like to thank Tony Kroch for stimulating discussions on issues in Chinese syntax and pointing me to a crucial article in the Distributed morphology framework. I would also like to thank my fellow graduate students at Delaware and my colleagues at the University of Pennsylvania. I benefited a great deal from discussions with James Huang, Fei Xia and Shizhe Huang, Fu-Dong Chiou and Shudong Huang.

Finally, I would like to thank my parents for their support of my education for all these years. This dissertation is dedicated to them.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

There are two important aspects of Chinese word formation that need to be addressed for in a theory of Chinese morphology. The first aspect is that the formation of complex words is highly regular and word formation is recursive. This seems to indicate that word formation is syntactic in nature. The second aspect of Chinese word formation is that Chinese words demonstrate lexical integrity effects. Components of words cannot be moved out of the word, cannot be deleted, are opaque to external reference and cannot take phrasal modifiers. This state of affairs seems to indicate that words are formed in the lexicon. There is thus a dilemma as to where words are formed in Chinese.

Work in the lexicalist framework either posits different notions of word (Dai 1992) or devises complicated word formation rules in the lexicon to account for this (Packard 2000). I have taken a radically different approach in this dissertation and insist that in Chinese complex words be formed in syntax, in the spirit of the Distributed Morphology Hypothesis (Halle and Marantz 1993; 1994 and others). In Chapter 2, I first examined the wordhood tests that have been proposed in the Chinese linguistics literature and conclude that some of the tests follow from the general X-bar theoretic framework and others follow from

locality conditions such as the LIH. I then showed how the LIH effects can be derived in a straightforward manner if words are formed in syntax in Chapter 3. In Chapter 4, I examined complex verbs and showed their formation provides further evidence for our theoretical position. In Chapter 5 I described an automatic word segmenter that implements our theoretical assumptions with the transformation-based error-driven algorithm (Brill 1993). Our working hypothesis is that if our theoretical assumptions are correct, we should see better results over "lexicalist" implementations. The results show that our implementation provides a significant improvement over a lexicalist implementation that uses the maximum matching algorithm in terms of overall accuracy and in dealing with new words. We take this to be a validation of our theoretical assumptions.

Chapter 1

INTRODUCTION

1.1 Overview

In a language like Chinese where there are no natural and practical clues to word boundaries, the notion of word can not be assumed, it must be argued for. Linguistic descriptions start from morphemes which are represented as characters in written form, as they are easier to identify, and combinations of morphemes either form words or phrases. The task of word identification in sentences thus involves determining when the combination of morphemes results in a word and when the combination of morphemes forms a phrase. To differentiate a word from a phrase, some "wordhood" tests are necessary. It is generally agreed among Chinese linguists that wordhood is a useful notion and the usefulness of this notion is supported by a series of wordhood tests which have been proposed in the literature (Huang 1984, Dai 1992, Duanmu 1997, Packard 2000). There is less agreement, however, on which tests are valid. A significant portion of the research in Chinese morphology is thus devoted to the motivation of wordhood and how words and phrases can be distinguished. It is only after the usefulness of

wordhood has been established and some means of word identification have been proposed that the theoretical modeling of Chinese morphology begins. The latest efforts in this regard in the framework of generative linguistics are represented by Dai (1992) and Packard (2000). In contrast, linguistic descriptions in English begin with words and wordhood can simply be assumed. Then one can either go "downward" to describe regularities in word formation or "upward" to capture regularities in syntax. In English and other Romance languages morphology is the place where word formation through affixation and compounding is described, characterized and theorized (Selkirk 1982, Di Sciullo and Williams 1987, Anderson 1992, Selkirk 1982, Lieber 1992, Zwicky 1990, to name a few).

The difference in how wordhood is approached does not necessarily mean the difference in the theoretical status of wordhood. In fact, except for the effort to motivate wordhood and find ways to identify words (from phrases) which is generally lacking or insignificant in morphology of Romance languages, there is a significant parallelism in how morphology is approached in Chinese and English-type of languages. Two lines of thinking particularly deserve reviewing. The first line of thinking is represented by Selkirk (1982), Sadock (1991), Lieber (1992), Ackema (1995) and others. The theme of this line of thinking is that morphology is an extension of syntax below the X^0 level, though the various authors differ in specific implementation of this general idea. This approach is adopted by Packard (2000) in his recent treatment of Chinese morphology. The

second line of thinking is represented by Zwicky (1990), which is extended to Chinese by Dai (1992). This line of thinking emphasizes on the co-existence of different notions of wordhood. There are at least notions of syntactic word, morphological word and phonological word, and the interactions between them explain the various phenomena surrounding word-like elements.

1.2 Dai's Work

Dai (1992) argues for the existence of the syntactic word (which he represents as W), the morphological word (which he represents as w) and the phonological word. A syntactic word, W is "a minimal constituent that syntactic rules may refer to" (Dai, 1992:20). Assuming Chinese has a syntactic rule in the form of VP --> V NP, the first immediate constituent V would thus constitute a slot which only Ws can fill. Thus, in (1),

(1) ta [$_{VP}$ [$_{V}$ xiu][$_{NP}$ qiche]]
he repair car
"He repairs cars."

(Dai:21)

xiu is a W by this definition. Morphological words, w, on the other hand, are the maximum domain to which morphological rules can refer. For Dai, w should not be understood as just a notion used by inflectional morphology, as is

traditionally assumed. Rather, it is motivated by the broader Lexical Integrity Hypothesis (LIH), first proposed in Jackendoff (1972). The LIH roughly states that components of a word (w for Dai) can not be manipulated by syntactic rules. Since inflectional morphology also obeys LIH, it can be subsumed under the latter. xiu in (1) vacuously obeys the LIH since it is not analyzable into multiple components and by definition it is a w. In this case there is no mismatch between w and W since xiu is both a W and a w. In other cases, however, the W and w may not coincide, this is illustrated in (2), from Dai (1992:27-28):

(2) a. ta lai-le liangci
he come ASP twice
"He came twice."

b. ta chang-le liangci
he sing ASP twice
"He sang twice."

c. ta lai chang-le liangci
he come sing ASP twice
"He came and sang twice."

d. *ta lai-le chang liangci
he come ASP sing twice

le is an aspect marker indicating perfect. (2a) and (2b) show that it can be attached to lai and chang respectively. (2c) shows that it can attach to the lai-chang sequence. However, it can not occur between lai and chang, as demonstrated (2d). Since lai and chang can occur in syntactic slots marked as

X^0s, as demonstrated in (2a) and (2b), they are Ws. However (2d) and (2d) show that lai-chang is a w instantiating two Ws. In this case there is no default one-to-one mapping between W and w in this situation. In fact, since lai and chang are also morphological words, lai-chang is also a morphological word composed of two morphological words. This can be represented schematically as (3):

(3) $[_{W/w}$ W/w W/w]

Dai intends to capture two basic sets of generalizations in Chinese morphology with this mechanism. The first set of generalizations is the phrase-like properties of complex words demonstrated in (2). For example such constructions are highly productive in the sense that lai can form complex morphological words this way with almost any other verb, as long as it does not violate any semantic or pragmatic constraints. Syntactic operations such as coordination can operate on the components of such complex words since it generally can operate on X^0 level elements. Word formation here often reflects syntactic relationships such as coordination, subject-predicate, head-complement and modifier-modifiee relationships. In some cases, the formation of complex words can be recursive. This set of generalizations can be captured by the notion of syntactic word, W. Since the components of such complex words are Ws, it is not surprising that they show the syntactic properties summarized above. On the

other hand, complex words like lai-chang also demonstrate another set of properties. They are not subject to syntactic operations such as alternative ordering (movement) and expansion. They demonstrate exocentric structures which phrases generally do not have. The set of properties can roughly be subsumed under the LIH. Dai captures this second set of properties by the notion of morphologic word, w.

It should be clear by now that for Dai, there is no systematic (derivational) correlation between W and w. W and w represent parallel grammatical modules. This is surprising since the main motivation for w is the LIH. In other words, morphological word is basically a domain where syntactic operations can not occur. More importantly, Dai's model is a static model and it does not explain how morphological words are derived. Therefore it is not satisfactory in that it does not explain why morphological words demonstrate the properties they demonstrate.

1.3 Packard's Work

Packard (2000) represents a different line of thinking in his work on Chinese morphology. He starts by demonstrating that it is possible to assign part-of-speech, which he terms as "form class" to word components or morphemes. This is made possible by his observation that words with unambiguous part-of-speech tend to retain their part-of-speech identities when they appear within

words, and that noun words have nominal constituents on the right and verb words have verbal constituents on the left. He called the second observation the "Headedness Principle", which he considers to be a language-specific principle in Chinese. Packard details how part-of-speech can be assigned to word components and readers should refer to his work for details. Given that it is possible to assign part-of-speech to word components, that complex words are formed independent of the form class they belong to and that complex words obey the Headedness Principle, the formation of complex words lends itself naturally to an X-bar theoretic analysis.

Packard further classifies morphemes in Chinese into four basic types: root words (X^{-0}), which are free content words, bound roots (X^{-1}) which are bound content words, word-forming affixes (X^{w}) which are bound function words that may change the part-of-speech of its host and grammatical affixes (G), which are bound function words that do not change the part-of-speech of its host. He proposes the following rules in the X-bar theoretic framework to account for word formation in Chinese:

(4) Rule 1: $X^{-0} \rightarrow X^{[-0,-1,\{w\}]}, X^{[-0,-1,\{w\}]}$

Rule 2: $X^{-0} \rightarrow X^{-0}$, G

Rule 1 means X^{-0}, X^{-1} and X^{w} can freely combine to form words except that X^{w}

can not combine with another X^{w}. Rule 2 means that X^{-0} can combine with G to form a word. Although Packard considers his focus of study to be syntactic words, it is clear from (4) that his syntactic word overlaps to a large extent with the morphological word of Dai.

Let us examine how Packard captures the two sets of generalizations in Chinese word formation. (4) captures the fact that complex words can be formed recursively by allowing X^{-0} to be a word component as well as the output. Packard rejects the seemingly grammatical relations between word components as only apparent, but since those are not primitive notions of X-bar theoretic syntax anyway, presumably those can be assigned to hierarchical structures provided by the X-bar theoretic formalism. Packard suggests that word-internal elements are potentially accessible to syntactic processes, subject to lexicalization.

Packard accounts for the second set of generalizations by appealing to lexicalization. Basically the more lexicalized a word is, the less likely that its components are accessible to syntactic processes. Packard also limits the productivity by allowing X^{-0} to be the only recursive node.

Packard may well be right in pointing out that the unavailability of word-internal components to syntactic processes is due to the high level of lexicalization and by lexicalization he means the cases in which material developed into or are recruited to form lexical items. However, lexicalization is generally considered to be a diachronic process and does not figure in synchronic

characterization of the grammar of a language.

It is clear that, Packard, like Dai, couched his analysis of word formation in a lexicalist framework in which there is a generative lexicon where complex words are formed. According to him, the lexicon is "a specialized linguistic module where all bound and free morphemes and all complex words known to the speaker (except for words containing grammatical affixes) reside and where the creation and comprehension of novel words takes place. Over time, the constituents of complex words in the lexicon may lose their individual identities, making them increasingly opaque to grammatical processes that refer to them" due to lexicalization. Specifically, Packard considers all morphemes to be listed in the lexicon. Also listed in the lexicon are all complex words in "precompiled" form with their morphological structure, except for complex words containing grammatical affixes. Rule 1 in (4), which composes all the listed complex words is a "redundancy rule" is the sense of Jackendoff (1972). This is in contrast with Rule 2, which composes grammatical words on-line. Although both words and morphemes are listed in the lexicon, only words are available for lexical insertion in the syntactic module.

1.4 Goal of the Present Work

The purpose of this dissertation is to provide a third alternative analysis of Chinese word formation. The proposed analysis will be in the spirit of

Distributed Morphology (DM) (Halle and Marantz 1993; 1994, Noyer 1997, Embick and Noyer 1999 and others). The analysis to be proposed intends to derive words in syntax or in the morphology module after syntax. In terms of empirical coverage, it intends to account for the fact that some word formation processes are highly regular and syntax-like, and at the same time, the resulting complex words obey the Lexical Integrity Hypothesis. Compared with Dai, the present analysis will not be a static module. Instead it will demonstrate how complex words are formed through a derivational process. In doing so I will show that Dai's morphological words are derived syntactically and there is no viable difference between syntactic word and morphological word in Chinese in Dai's sense. Compared with Packard, I will show that at least some of the complex words need to be formed in syntax or after syntax, feeding on the input of the syntactic structure. Since the mechanism used by Packard to compose complex words and those used to compose phrases are essentially the same, with only minimal differences, I will take it to be a disadvantage against that approach to have redundancy rules in the lexicon. It must be pointed out that Packard bases part of his arguments as to whether complex words are listed in the lexicon on experimental evidence, which I will not consider. My consideration will be from a pure formal perspective, where a simpler computational system that handles the same amount of empirical data will be the superior system.

This dissertation is organized as follows. In Chapter Two I will examine

the tests that have been used to motivate wordhood and differentiate words from phrases. I will argue some tests are more relevant than others to the present study. In Chapter Three I will outline my analysis of Chinese word formation. I will specify the underlying assumptions and the formal mechanisms I will use to account for facts in Chinese word formation. I will show how words are derived in this system and how the two sets of generalizations surrounding Chinese word formation are accounted for. In Chapter Four I will present cases in Chinese where word formation takes place in syntax and show those facts are easily accommodated in the present analysis but will present problems for lexicalist systems like those of Dai and Packard. I will also present cases where wordhood and phrasehood are dependent on the syntactic context and show these will also cause problems for lexicalist approaches. In Chapter Five I will consider the computational implications of this approach for automatic word identification for Chinese and show that the experimental results support the use of syntactic information in word identification. I will show the more syntactic information used, the better the result in terms of accuracy. I will reason that this supports the theoretic approach adopted in the present study. Finally, I will conclude this dissertation in Chapter Six.

Chapter 2

IDENTIFYING WORDS

2.1 Introduction

As we have suggested in the previous chapter, the study of word formation in Chinese begins with the identification of words. In this respect Chinese is very different from Romance languages such as English where one can find clues of words in written form by looking at the natural delimiters such as white space. Although such markers of word boundaries do not necessarily have any theoretical importance, they are nevertheless good indications of where a word starts and ends. Chinese is also very different from highly inflectional languages such as Korean and Japanese where one can detect word boundaries by looking at the inflectional patterns of the word in an isolated fashion. Instead, a set of wordhood tests have been proposed in Chinese linguistic literature over the years (Chao 1968, Lu 1979, Dai 1992, Duanmu 1997 and many others).

In this chapter, I will first review the non-syntactic tests that have been proposed to identify words. I will examine phonological, semantic and meta-linguistic approaches and show that their predictions are not consistent with the

predictions of the syntactic criteria. I will conclude that they are not useful in identifying words syntactically. Next I will examine the syntactic criteria that have been proposed previously in Chinese linguistics literature. I will show the expansion test and the XP-substitution test can be derived from the properties of the general X-bar theoretic framework that we will adopt. I will then show that the remaining tests are derivable from locality conditions that hold within the domain of words, such as the Lexical Integrity Hypothesis (Jakendoff 1972, Huang 1984). The purpose of this chapter is to evaluate these tests and filter out the core generalizations that need to be captured in the theoretical framework proposed in the next chapter.

2.2 Previous Non-syntactic Tests

2.2.1 Introduction

Phonological, semantic, meta-linguistic criteria for identifying Chinese words have been proposed in Chinese linguistics literature. I show that phonological and semantic criteria that have been proposed are not useful in identifying morphosyntactic words, which is the primary focus of study in this thesis.

2.2.2 Tests Based on Phonological Criteria

2.2.2.1 Syllable Count

The idea of identifying words with syllable count is first suggested by Lu (1979): "The word in the mind of the average speaker is a sound-meaning unit that is not too long and not too complicated, about the size of a word in the dictionary entry." The means that Chinese words should roughly be between one and four syllables/characters. Lu certainly does not mean to say that words can be determined by counting the number of characters from the beginning of the sentence to the end. His suggestion is limited to differentiating words from phrases in the case of Chinese nominal compounds. For instance, Chinese compounds can theoretically be arbitrarily long:

(1) a. ren-zao wei-xing
man-make satellite
"man-made satellite"

b. dian-xun guan-li-ju
communication administration
"communication administration"

c. hua-dong ke-ji da-xue
Hua-dong science-technology university
"Hua-dong science and technology university"

d. lian-he-guo jiao-yu ke-xue wen-hua zu-zi
United Nations education science culture organization
"United Nations Education Science Culture Organization"

Lu basically suggests that the compounds listed above are too long to be words. They must be phrases. There are several reasons why such a criterion is not very useful in identifying words. First, such a criterion obviously can not apply alone. In the above examples, it is possible to talk about syllable count only after words in the whole string are identified. It is until then syllable count can be applied to decide whether the whole string is a word or a phrase. Second, even if its predictions are correct, it is mysterious why it should work. Why is it that words can only have one to four characters / syllables? Assuming the predictions it makes are right, the fact that words can only have one to four characters sound more like a reflex of some other deeper criterion, rather than the criterion itself.

2.2.2.2 Bound / Free

The other phonological criterion that has been proposed is the distinction between bound and free forms. This criterion says that if an immediate component of an expression is a bound form then the whole expression is a word. For instance, gong-ye-hua ("industrialize") is correctly predicted to be a word since hua is bound. However it also predicts that the following particles together with the constituent it attaches to (the string in the square brackets) are words as these particles can not occur alone. This is a highly implausible conclusion from a syntactic point of view given that the constituents they attach to are highly analytical syntactically:

(2) a. Localizers:
[gai-ge he kai-fang hou], jingji dedao le fazhan.
[reform and open after] economy get LE development
"Economy has developed after reform and opening to the outside world."

b. Sentence-final particles:
[ta yao likai jia qu xuexiao ma]?
he will leave home go school MA?
"Will he leave home and go to school?"

c. ba/bei
ta [ba Zhangsan he Lisi] da le
he [BA Zhangsan and Lisi] hit LE
"He hit Zhangsan and Lisi."

ta [bei Zhangsan he Lisi] da le
he BEI Zhangsan and Lisi hit LE
"He was hit by Zhangsan and Lisi."

d. de
wo kanjian le [Zhangsan he Lisi de] penyou
I see LE [Zhangsan and Lisi DE] friend
"I saw Zhangsan and Lisi's Friend."

Such a conclusion is clearly incorrect. This criterion can be qualified to be "if one of the immediate constituent is bound and the other is at least a word, then the whole expression is a word." But this refinement results in circularity unless there are other criteria which can be used to test wordhood.

2.2.3 Tests Based on Semantic Criteria

2.2.3.1 Idiomaticity / Non-compositionality

A phonologically identified expression is a word if it is idiomatic or non-compositional, which means that its meaning is not derivable from the meaning of its parts in a well-defined manner. The underlying assumption of this criterion is that non-compositionality is a sufficient and necessary condition for wordhood. It correctly predicts (that is, its predictions are consistent with that of syntactic criteria), for example, da-yi is a word:

(3) da-yi
big-garment
"overcoat"

da-yi is not a garment which is big. Instead, it means "overcoat". However, many idioms are not words and they are syntactically analyzable. For instance,

(4) Gua yang tou, mai gou rou
hang goat head, sell dog meat
"Say one thing and do another."

It is obvious in the above example the structure can be analyzed as a coordination of verb phrases, with each verb phrase consisting of a verb taking an object. However, the meaning of the whole is not derivable from that of its

constituents. In fact, the meaning of such expressions can be described as ambiguous, with the literal meaning being compositional and the figurative meaning non-compositional. This makes it even more difficult to use non-compositionality as a test for wordhood. The special meaning attaches to syntactic structures of various sizes and is often culture-specific. It should be considered to be independent of the syntactic structure.

Idiomaticity is not a necessary condition for wordhood either. Some expressions which are generally recognized as words by other criteria are not idiomatic. For example, the meaning of shu-mu "tree-wood=trees" is arguably transparent yet it is generally considered to be a word.

In addition, the idiomaticity test does not readily apply to functional categories. For example, it is hard for the idiomaticity test to make a prediction as to whether the nominal modifier marker de is a word or not in Chinese. It is a syntactically important entity yet it does not appear to have a well-defined meaning.

2.2.4 Meta-linguistic Criteria

2.2.4.1 Productivity

Duanmu (1997), noting the difference in productivity between nominal phrases and compounds, concludes that the difference in productivity can be used to distinguish words from phrases. Specifically, phrasal rules are productive

while word-formation rules are subject to arbitrary gaps. The examples Duanmu uses to demonstrate the difference are listed below:

(5) a. *gui shou-juar
expensive handkerchief
"expensive handkerchief"

b. gui de shou-juar
expensive DE handkerchief
"expensive handkerchief"

(6) a. *bao hui-chen
thin dust
"thin dust"

b. bao de hui-chen
thin DE dust
"thin dust"

(7) a. *cong-ming dong-wu
clever animal
"clever animal"

b. cong-ming de dong-wu
clever DE animal
"clever animal"

(8) a. *hua-ji ren
funny person
"funny person"

b. hua-ji de ren
funny DE person
"funny person"

(9) a.*huang qi-chuan
yellow steam-boat
"yellow steam-boat"

b. huang de qi-chuan
yellow DE steam-boat
"yellow steam-boat"

(10) a.*shen shu
deep book
"difficult book"

b. shen de shu
deep DE book
"difficult book"

(11) a.*duan cheng-mo
short silence
"short silence"

b. duan de cheng-mo
short DE silence
"short silence"

(12) a. *bai shou
white hand
"white hand"

b. bai de shou
white DE hand
"white hand"

The above expressions form minimal pairs. The expressions with de, which are generally agreed to be phrases, are grammatical while lack of them leads to ungrammaticality. The de-less expressions are thus words since they are unproductive and have gaps, as the argument goes.

First of all the judgments in (5a), (7a), (9a), (12a) are problematic. For the

remaining cases, the ungrammaticality seems to be due more to prosodic constraint (Feng 1995) than anything else. For example, although hua-ji ren in (8a) sounds unnatural, but hua-ji ren-wu ("funny person") sounds perfectly grammatical.

Second, even if the judgments above are correct, to prove that productivity is a useful test for phrasehood, one would also have to show that where de-less expressions exist, their phrasal counterparts (with de) should also exist. The following shows this is not the case in Chinese:

(13) a. hei ban
black board
"blackboard"

b. ?? hei de ban
black DE board
"a board which is black"

(14) a. leng-she
cold shoot
"Shoot (soccer) suddenly"

b. (?) leng de she?

This means that productivity or lack thereof does not correlate with wordhood, as Di Scioullo and Williams (1987:10) concluded for English.

2.2.5 Summary

In this section, we have reviewed phonological, semantic and meta-linguistic criteria for wordhood in Chinese that have been proposed in the literature. It is concluded that the predictions of the proposed criteria are not plausible as tests of wordhood and none of the proposed criteria are adequate in defining words in Chinese, although they may be meaningful in their own right.

2.3 Syntactic Criteria

2.3.1 Introduction

In this section, we turn to the syntactic criteria for wordhood in Chinese. I will show some of the criteria are more relevant than others in determining morphosyntactic words in Chinese. I will also show that some of the criteria follow from the general properties of the X-bar theory while others can be subsumed under LIH.

2.3.2 Review of the Syntactic Criteria

2.3.2.1 Expandability

The expansion test was proposed by Wang (1944), Lu (1964) and was adopted by many others. This test says that if an expression allows an item to be inserted between its parts, then it is a phrase; otherwise it is a word. Therefore,

(15) a. bai zhi
white paper
"white paper"

b. *bai de zhi
white DE zhi

(16) a. xin zhi
letter paper
"letter paper"

b. *xin de zhi
letter DE paper
"letter paper"

bai zhi in (15a) is a phrase since it allows de to be inserted between its parts but xin-zhi in (16a) is a word since it does not. From a pure descriptive point of view, it basically says that if a nominal expression has a counterpart that has de, then it is a phrase. Otherwise it is a word.

However, in the following example, the test makes the wrong prediction that xin-zhi is a phrase since it allows a string of words to be inserted between its constituents:

(17) xin [yong bi xie zai zhi] shang de
letter [with pen write at paper] on DE
"The letter is written on paper with a pen."

It should then be concluded that xin-zhi is a phrase, contradicting what is predicted earlier. Clearly as described above the expansion test cannot be right.

This is like saying "writer" in English is a phrase since it is possible to say "writeING A LETTer".

It also should be noted that although some expressions allows other expressions to be inserted between its parts but the meaning has changed after the expansion:

(18) a. you-zui
oil-mouth
"glib talker"

b. you de zui
oil DE mouth
"greasy mouth !=glib talker"

Therefore the insertion test seems to falsely predict that you-zui is a phrase. Attempting to save the insertion test (which failed in the above two examples), proponents of the test qualify the test by imposing two conditions. One is that the resulting expressions should have the same structure as the original one. That will prevent the test from predicting that xin-zhi is a phrase. The other condition is that the resulting expressions will not change the meaning of the original expression. Presumably this will prevent the test from predicting that you-zui is a phrase.

For the qualified expansion test to work, there needs to be a clear idea of what constitutes change of structure and what constitutes change of meaning. For

example, why is that the insertion of de does not cause a change of structure? The answer to this question presupposes a clear understanding of the role of de, which seems to be the only material that can be inserted without causing a change in the structure.

The de-insertion test is largely a stipulation in the sense that it presumes the special status of de without attempting to put it in a larger context. For example, what is it about de that makes it so special that it can serve as a test for wordhood? Are there other elements in the language that can also play this role? I will show in section 4.4 that the de-insertion test bears on the wordhood test because it is a functional category that does not form words with other morphemes or words. In other words, it must form a phrase. It follows from the language-specific properties in Chinese that functional categories do not from larger functional categories by taking on additional material.

2.3.2.2 The XP-substitution Test

Another test that is proposed to account for the difference between words and phrases is called XP-substitution. As noted by Fan (1958), in [A de N] A can take an adverb that modifies it but in [A N] A cannot take such a modifier. Therefore [A de N] is a phrase while [A N] is a word. This test is also called the Adverbial Modification test by Duanmu. The following examples from Duanmu illustrates this point:

(19) a xin de shu
new DE book
"a new book"

b. hen xin de shu
very new DE book
"a very new book"

c. geng xin de shu
more new De book
"a more new book"

d. zui xin de shu
most new DE book
"the newest book"

e. zheme xin de shu
such new DE book
"such new books"

f. bu xin de shu
not new DE book
"a book that is not new"

(20) a. xin shu
new book
"a new book"

b. *hen xin shu
very new book
"a very new book"

c. *geng xin shu
more new book
"a more new book"

d. *zui xin shu
most new book
"the newest book"

e. *zheme xin shu
such new book
"such a new book"

f. *bu xin shu
not new book
"a book that is not new"

Replacing N with an NP is also forbidden (Duanmu, 1997:152). This is formalized as follows:

(21) a. [M de N]->[M de XP] where M is a modifier

b. xin de [san ben shu]
new DE three CL book
"three books that are new"

c. xin de [nei ben shu]
new DE that CL book
"the book that is new"

In contrast, when there is no de, XP substitution is not allowed:

(22) a. *[M N]-->[M XP]:

b. *xin [san ben shu]
new three CL book
"three books that are new"

c. *xin [nei ben shu]
new that CL book
"the book that is new"

It is concluded that [M N] is a word while [M de N] is a phrase. Both the

XP substitution test and the expansion test assume the special status of de. I will show in Section 4.4 that the XP substitution test works for the same reason that the de insertion test works, which is that de does not form large words by taking on additional material.

2.3.2.3 The Exocentricity Test

As I have shown in Chapter III exocentric structures occur when the internal category does not match the external category:

(23) kao-rou [V N]$_N$
roast-meat
"roast meat"

kao-rou in (23) it consists of a verb taking a noun phrase. Since the head is a verb and the entire structure is expected to be a verb phrase. However, the structure as a whole can be both verbal and nominal:

(24) Nominal
da-jia dou xi-huan kao rou
everybody all like roast meat
"Everybody likes roast lean meat."

(25) Verbal
a. da-jia dou zai kao xin-xian de rou
everybody all proceed roast fresh DE meat
"Everybody is roasting fresh meat."

b. rou, da-jia dou zai kao
meat everybody all proceed roast
"Meat, everybody is roasting."

When kao-rou is exocentric, as it is in (24), it is a word. However, when it is endocentric as it is in (25a), it is a phrase. This has been used as a test for wordhood by Duanmu (1997) and others.

The exocentricity seems to give the right predictions in these two cases. It correctly predicts that the nominal kao-rou is a word while its verbal counterpart is a phrase. Exocentricity coincides with wordhood in most cases and it can be used as a test for wordhood. However, this may not the end of the story. This reflects the general tendency that words, not phrases, can be used in categorical context (Marantz 1997).

2.3.2.4 Conjunction Reduction

Conjunction reduction is the result of deletion of parts of the conjoined structure. The deletion is context-dependent.

(27) jixu [gaijin he tigao]
continue [correct and improve]
"continue to correct and improve"

(28) *da [si he shang xuduo diren]
hit [dead and wounded many enemy]
"Kill and wound many enemies by hitting"

Conjunction reduction correctly predicts that jixu gaijin is a phrase while da-si is a word. Conjunction reduction also correctly predicts, localizers, sentence-final particles, ba and bei, de to be words.

(29) Localizers:
[gaige he kaifang] hou, zhongguo jingji dedao le fazhan.
reform and open after China economy get LE development
"China's economy has developed after reform and opening to the outside world."

(30) Sentence-final particles:
[ta yao likai jia qu xuexiao] ma?
he will leave home go school MA?
"Will he leave home and go to school?"

(31) ba/bei
ta ba [Zhangsan he Lisi] da le
he BA Zhangsan and Lisi hit LE
"He hit Zhangsan and Lisi."

ta bei [Zhangsan he Lisi] da le
he BEI Zhangsan and Lisi hit LE
"He was hit by Zhangsan and Lisi."

(32) de
wo kanjian le [Zhangsan he Lisi] de penyou
I see LE Zhangsan and Lisi DE friend
"I saw Zhangsan and Lisi's Friend."

The conjunction reduction test predicts that verb resultative compounds, verb potential forms are words since they do not allow expansion either:

(33) verb resultatives:

a. da-ying zhe chang bisai
play-win this CL match
"play and win this match"

b. *da [ying he shu] zhe chang bisai
play win and lose this CL match

c. *da da-ying zhe chang bisai
play big win this CL match

(34) verb potential forms:

a. da-de-ying zhe chang bi-sai
play-DE-win this CL match
"able to play and win this match"

b. *da de [ying he shu] zhe chang bisai
play DE win and lose this CL match

c. *da de da-ying zhe chang bisai
play DE big win this CL match

In contrast, V-de constructions are phrases and they allow conjunction reduction and expansion:

(35) a. da de duishou hen pilao
play DE opponent very tired
"play and cause the opponent very tired"

b. da de duishou hen pilao, ziji hen xinfen
play DE opponent very tired, self very excited
"play and cause the opponent very tired"

So conjunction reduction seems to be a sufficient condition (of course it is

not a necessary condition). In section 4.4 I will attempt to show why conjunction reduction is a possible test from a formal perspective.

2.3.2.5 Context-dependent Deletion

Context-dependent deletion predicts, among other things, that deleting material from inside a word is not allowed:

(36) a. fang-jian li zuo-zhe xu-duo [Shanghai ren]$_i$.
room inside sit-ASP many Shanghai people.
[pro$_i$] ge-ge shenqin jin-zhang
each look nervous
"Many Shanghai folks sat in the room. Everybody looks nervous."

b. *fang-jian li zuo-zhe xu-duo [Shanghai$_i$] ren.
room inside sit-ASP many Shanghai people.
[pro$_i$] shi yi-ge da chengshi
be one-CL big city
"Many Shanghai folks sat in the room. Shanghai is a big city."

In (36a) the deleted item co-refers with the entire phrase / word, so it is grammatical. In contrast, in (36b) since the deleted item co-refers with only a part of a word, it is bad. It therefore predicts that Shanghai is part of a word, not a word itself in this context.

2.3.2.6 Pronominalization

Still using the same example above, only the empty category becomes an

overt pronoun. The pronominalization test says that a proform can not refer to part of a word:

(37) a. fang-jian li zuo-zhe xuduo [Shanghai ren]$_i$.
room inside sit-CL many Shanghai people.
tamen$_i$ ge-ge shenqin jin-zhang
they each look nervous
"Many Shanghai folks sat in the room. Everybody looks nervous."

b. * fangjian li zuo-zhe xuduo [Shanghai$_i$] ren.
room inside sit-ASP many Shanghai people.
nei$_i$ shi yi ge da chengshi
that be one CL big city
"Many Shanghai folks sat in the room. Shanghai is a big city."

In (37a) the pronoun co-refers with the entire phrase/word, so it is grammatical. But in (37b) since the pronoun co-refers with only a part of a word, it is bad. It therefore predicts that Shanghai is part of a word, not a word itself in this context.

2.3.2.7 Movement

The movement test says subcomponents of words can not be moved out of the word. It correctly predicts that dan-xin is a phrase in the following example:

(38) a. wo dan xin
I carry heart
"I am concerned."

b. xin, wo yi-dian dou bu dan
heart I one-point all not carry
"I am not worried one bit."

The movement test is generally a reliable test. We will show in the next section that it follows from a more general locality condition, the Lexical Integrity Hypothesis.

2.3.3 Summary

To sum up, I have examined the various syntactic criteria proposed previously, namely, expansion (de-insertion), XP substitution, exocentricity, conjunction reduction, context-dependent deletion, pronominalization. I have suggested that the expansion test and the XP substitution test are useful because of the special status of de. The exocentricity test is useful because words, not phrases, tend to occur in different context, marked by functional categories. The movement test, the pronominalization test, the context-dependent deletion test and the conjunction reduction test are useful because that words obey locality conditions such as the LIH.

2.4 Deriving the Syntactic Tests

2.4.1 Introduction

What is undesirable with the syntactic criteria proposed previously is that they do not seem to follow from a coherent set of assumptions and a well-defined

linguistic model. As a result they appear to be no more than an unrelated set of observations that people use to get an idea of what a word is like in Chinese. For example, why should the XP-substitution test be related to the movement test? Why should they give consistent predictions as to what is a word and what is not? Huang (1984) attempts to unify these observations with the Lexical Integrity Hypothesis, first proposed in Jackendoff (1972). The syntactic operations such as movement is impossible from within words as a result of the Lexicalist Integrity Hypothesis.

In this section I will attempt to show that the facts predicted by the expansion test and the XP substitution test follow from the status of de as a functional category. The movement test, the conjunction reduction test, the context-dependent deletion test and the pronominalization test follow from a stringent notion of locality condition that holds in the domain of words.

2.4.2 The Validity of the Expansion Test and the XP Substitution Test

We will assume, with Packard, that the morphosyntactic word we will focus is basically the X^0 in the X-bar framework. We will maintain the position that X^0 is a theoretical primitive and the distinction between X^0 and XP is real.[1] In addition, it is reasonable to assume, as a language-specific condition in

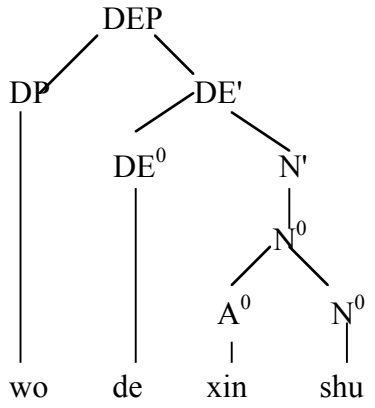
[1]See Chomsky (1995) for a different view.

Chinese, that not all X^0s, when they are the head, will form other X^0s by taking other elements as its complement, since they belong to a closed class and will never form derived entities of the same category.[2] For instance, it is reasonable to assume that a preposition will not form another preposition by taking a complement, either X^0 or X^{max}. Therefore the ability to form derived entities of the same category by taking complements belongs to open class words such as nouns and verbs. In Chinese, closed classes include de, localizers, particles, numerals, classifiers, etc.. Any constituent containing them are necessarily phrases.

The XP-substitution test makes crucial use of de. For these tests to be meaningful, it is crucial to understand the nature of de. This issue has been discussed extensively in the literature (Huang, 1982; Cheng, 1986; Ning, 1993; Xue, 1997 and many others). The exact nature of de is still a controversial matter, but it is generally agreed that de is some kind of functional category and is the head of a phrasal projection. We will assume with Xue (1997) that de is a determiner and projects a DEP, which is some kind of determiner phrase (39):

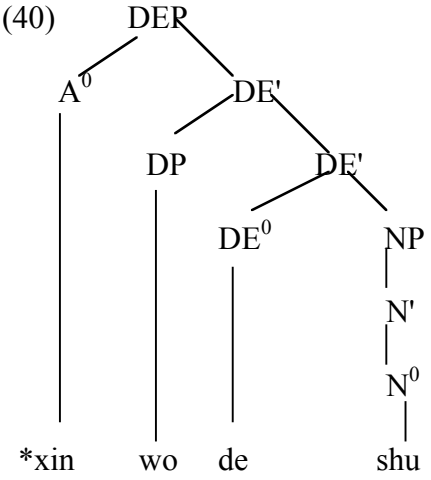
[2]I take this to be a language-specific property of Chinese. See Lieber (1992) for facts to the contrary in English.

(39)



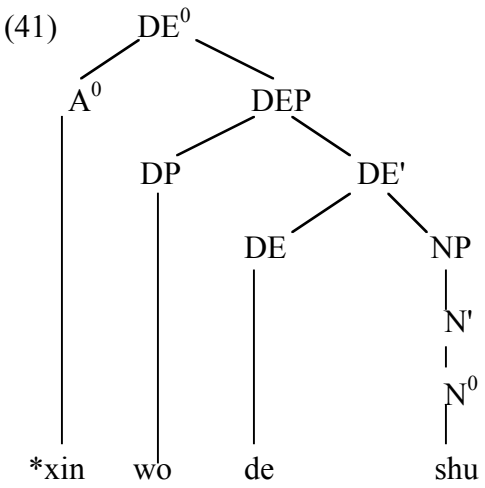
Two observations are in order here. First DE projects a DEP, not a DE^0, as it would be possible if de does not belong to a closed class. As a member of a closed class, the only option here is to form a phrasal category. Second, as a determiner, de does not take on multiple specifiers. This effectively rules out (40):

(40)



Note a structure like (41) is impossible either because X^0s can not dominate XPs:[3]

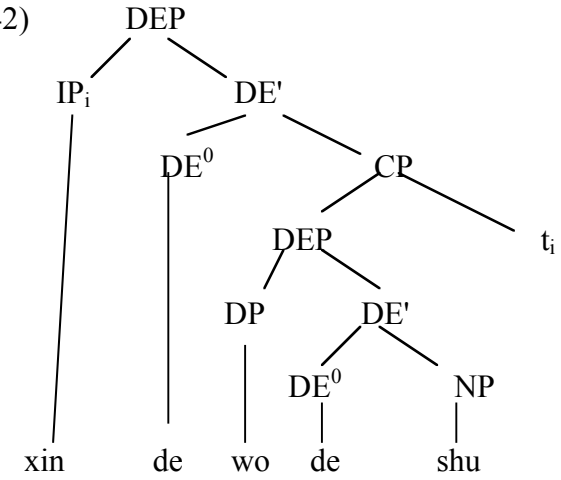
(41)



However, something like (42) is possible, as has been argued in Xue (1997), since there are two <u>de</u>s in this case and each licenses a specifier:

[3]Again, see Lieber (1992) for a different view.

(42)



If this line of analysis is on the right track, then when de can be inserted into a string is relevant as a test for wordhood because it can not occur within words. Whenever de occurs, it is hosting a phrase. Hence expandability is a useful test for wordhood. The XP substitution test is useful for the same reason: the insertion of a functional category de creates a situation where an XP occurs within the X^0, which is banned in Chinese.

2.4.3 Deriving the remaining tests from the LIH

I will assume with Huang (1984) that Chinese words obey the LIH and as a result, conjunction reduction of a component of a word, movement of a component of a word, deletion of a component of a word and pronominalization of a component of a word is impossible. I will recast the LIH as the Morphosyntactic Word Integrity in the DM framework in the next chapter,

following Embick and Noyer (1999).

2.5 Summary

In this Chapter, we examined the wordhood tests that have been proposed in the literature. We have shown that the predictions of the phonological and semantic tests do not converge with that of the syntactic tests. We have also shown that syntactic tests such as the expansion test and XP substitution test are relevant because of the special status of de as a functional category. The other tests follow from locality conditions such as the LIH, which holds within the domain of words.

Chapter 3

DISTRIBUTED MORPHOLOGY AND CHINESE WORD FORMATION

3.1 Introduction

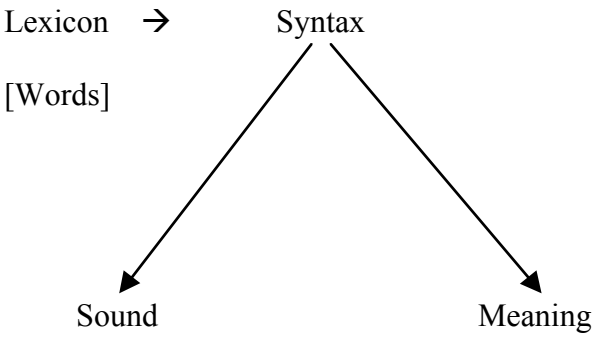
In this chapter I will first outline the theoretical assumptions of the Distributed Morphology (DM) hypothesis advanced in Halle and Marantz (1993; 1994), Marantz (1997a), Noyer (1997), Embick and Noyer (1999) and others. I will do this by first outlining the generally accepted lexicalist view of grammar and show how DM is different. I will not explore the full scope of the implications of DM; instead I will focus on how words are formed and where word formation takes place within this architecture of the grammar. Where necessary I will show how word formation in this theory of the grammar differs from that of the lexicalist hypothesis. Applying this theory to Chinese, I will show how the regularities of Chinese word formation are captured and how LIH is obeyed in word formation. I will discuss further implications of DM and show how DM might accommodate certain morphological phenomena in Chinese that are challenging for the lexicalist approaches.

3.2 The Distributed Morphology Hypothesis

3.2.1 Overview

It is convenient to begin by outlining the generally adopted architecture of the grammar that is consistent with the lexicalist hypothesis:

(1) The Lexicalist View of Grammar



For the lexicalist view of the grammar, the lexicon is much more than a list of lexical items that will feed syntax and serve as syntactic primitives. Instead it carries a significant burden of language description. Lexicalists may differ among themselves as to what the place called lexicon may contain but this much the lexicalists should agree. First of all it contains the list of words, with their morphological representations. Second, the lexicon also contains morphemes. Third, the lexicon also contains morphological operations that combine morphemes into words. For example, Packard (2000) proposes that all the

morphemes in Chinese should be contained in the lexicon. He also proposes all the words, except for what he calls grammatical words, should also be listed in the lexicon, in precompiled form. For many lexicalists, the lexicon contains even more. For example, for the Lexical Phonologists, the lexicon is also a storage house of special sounds. For others the lexicon is a storage house of special meanings and special sound-meaning correspondences. For them, idiomatic phrases should also be listed in the lexicon, although the formation of idiomatic phrases is clearly syntactic in nature.

Although the lexicalists may not agree on the extent to which morphology and syntax interact, most lexicalists would agree that morphology interacts with syntax in very limited ways. In fact, this is the primary purpose of positing a separate linguistic module called lexicon. Specifically, once a word is done with morphological processes, its internal structure is opaque to syntax. In other words, syntactic operations and processes can not make reference to or manipulate word-internal elements and structures. The lexicalist position is most explicitly articulated in Williams (1987:47):

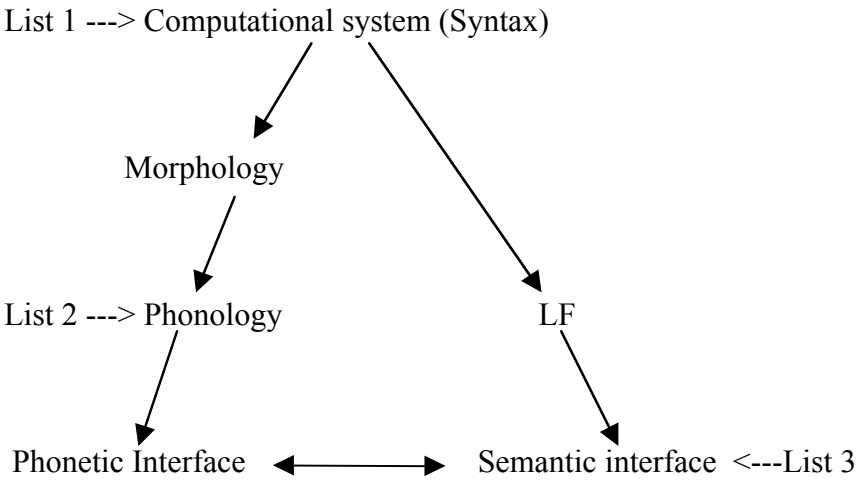
> We regard the need for the lexicalist hypothesis (especially the lexical integrity hypothesis) as arising from a fundamentally mistaken idea of what a grammar is. The hypothesis is true in that morphology and syntax are separate in the way they are, but ideally it should 'go without saying'. Morphology and syntax are different (though) similar sciences about different objects, so the idea that the derivations in one could get mixed up with those of the other should not arise in the first place.

> The lexicalist hypothesis is not so much a thesis of grammar (like an island condition) as it is a statement about the global architecture of grammar: the theory of grammar has two subtheories, morphology and syntax, each with its own atoms, rules of formation, and so on.

Although some lexicalists may not take this strong a position, but all lexicalists would agree that as a separate linguistic module, the lexicon is autonomous and has limited interaction with syntax. One of the most important functions of the lexicon is to do all the necessary morphological computation and provide input to syntax.

In one of the more explicit spellouts of the Distributed Morphology, following Halle and Marantz (1993, 1994), Marantz (1997) outlines the following alternative conception of grammar:

(3) DM view of grammar



The most notable departure of DM from a lexicalist view of the grammar is that it dissolves the all-encompassing lexicon. In DM the role of the lexicon is taken over by three separate lists (hence Distributed Morphology). List 1 is the narrow lexicon that contains morphemes that syntax operates with. Morphemes are roots or other functional elements containing bundles of semantic, syntactic and morphological features. These morphemes, rather than words, feed syntax and are thus syntactic atoms.

List 2, called Vocabulary by Marantz (1997), "provides the phonological forms for the terminal nodes from the syntax (for roots as well as bundles of grammatical features) unless roots come with phonological forms from the narrow lexicon). Vocabulary contains the connections between sets of grammatical features and phonological features, and thus determines the connections between terminal nodes from the syntax and their phonological realization." The vocabulary items are in the form of the correspondence between a set of semantic, syntactic and morphological features with a set of phonological features. During the Vocabulary Insertion (VI), the vocabulary item whose semantic, syntactic and morphological features matches that of a terminal node is inserted. In cases where there are multiple matches, the vocabulary items compete for insertion, with the vocabulary item with the most matches winning out. This implies that the vocabulary items do not need to have all the semantic, syntactic and

morphological features to be inserted. This feature is called Underspecification.

List 3 or the "Encyclopedia" is the list of special meanings. The Encyclopedia lists the special meanings of roots, relative to the context of other roots, within local domains. Special meanings are assigned to roots when they are in a special structural relation relative to other roots or bundles of grammatical features. For example, root "KICK" has a special meaning when it occurs in "kick the bucket" which is different from that of "KICK" in "kick the guy".

Comparing these three lists with the lexicon in lexicalist approaches, we will see that the lexicon's function of storing the morphemes is taken over by the narrow lexicon (List 1) and the vocabulary (List 2). The special meaning of the idiomatic expressions is accounted for by Encyclopedia, which is List 3. The words along with morphological operations that create them are nowhere to be found in these lists. In fact, DM insists that the morphological operations are not qualitatively different from the syntactic operations. Thus, words can either be formed in syntax, or in the Morphological component after syntax. Morphology is the place where certain morphological operations (different from the morphological operations in the sense of the lexicalist hypothesis) occur. The morphological operations in DM mediate the mapping between syntactic representations with pronunciation. The morphological operations for DM include addition of morphemes, Merger, Fusion, Fission, and Impoverishment (Halle and Marantz 1994). Merger adjoins the head (X^0) of one phrase to the

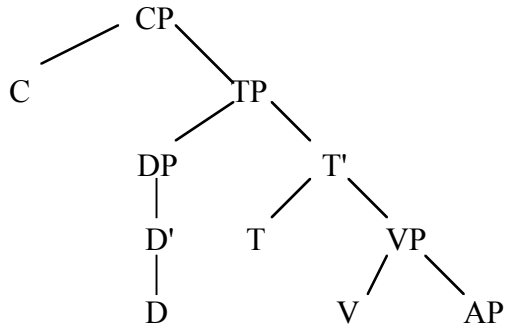
head of another phrase to form a complex X^0 element. This is first proposed in Marantz (1984) and formalized in Marantz (1988):

(4) Morphological Merger

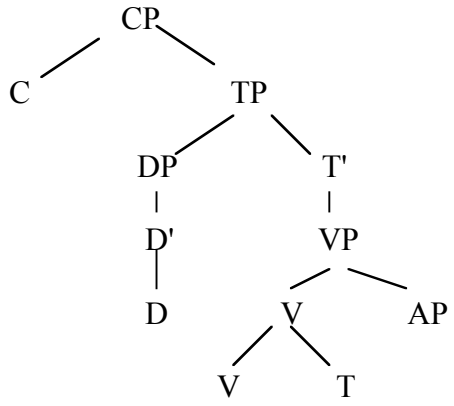
At any level of syntactic analysis (d-structure, s-structure, phonological structure), a relation between X and Y may be replaced by (expressed by) the affixation of the lexical head of X to the lexical head of Y.

Merger is illustrated in Halle and Marantz (1993:134-135):

(5) a.



b.



c. They sleep late

Halle and Marantz argued that syntactic "lowering" is an instance of Merger, as illustrated in (5b) for the sentence in (5c).

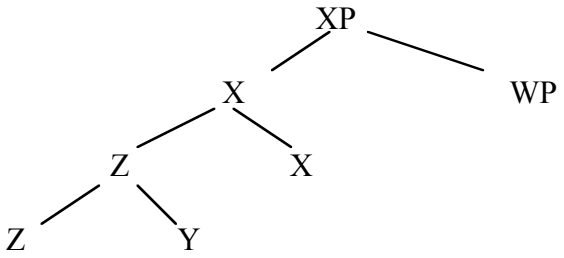
Morphemes that do not contribute to the syntactic structure prior to Morphology but are relevant to pronunciation are added in Morphology. For example, noun stems in English are augmented by case morphemes, which are not relevant to syntax prior to Morphology (Embick and Noyer 1999):

(6) Noun -> [Noun + Case Morpheme]

Another morphological operation, Fusion, fuses the morpho-syntactic features of two terminal nodes into one. Fission does the opposite: it divides the morpho-syntactic features of one terminal node and turns them into two terminal nodes. Lastly, impoverishment refers to the process in which certain mophosyntactic features in a terminal node are bleached (to affect Vocabulary Insertion) subject to certain structural conditions.

Prior to Morphology, (complex) words can also be formed via head-adjunction (Embick and Noyer 1999):

(7)



In this case Y adjoins to Z and Z+Y adjoins to X, with X, Y and Z being abstract morphemes which contain bundles of semantic, syntactic and morphological features. Following Embick and Noyer, I will call X a Morphosyntactic Word (MWd) since it is the highest segment of an X^0 not contained in another X^0. Z, the lower segment of Y, and the lower segment of X are Subwords since they are terminal nodes but not MWds. MWds and Subwords are formally defined as follows:

(8) a. At the input to Morphology, a node X^0 is (definitional) a Morphosyntactic Word (MWd) iff X^0 is the highest segment of an X^0 not contained in another X^0.

b. A node X^0 is a Subword if X^0 is a terminal node and not an MWd.

Taken together, there are at least three ways in which complex words are formed in DM, namely, head-adjunction in syntax (with or without head-

movement), addition of morphemes and Merger in Morphology. Merger may take on different forms before and after Vocabulary Insertion (Embick and Noyer 1999). Word formation in this model of grammar is strictly derivational. They are formed in syntax by syntactic operations such as head-adjunction, and the result is a MWd, with one or multiple Subwords arranged in a hierarchical order. The MWds will then undergo morphological operations such as Morphological Merger and the addition of morphemes. The MWds after the morphological operations will be complex words that are roughly Dai's morphological word and Packard's syntactic words.

I will follow Embick and Noyer in stating that a complex X^0 created in syntax can not be infixed within another X^0 in Morphology. MWds thus observes MWd Integrity.

Having described how words are formed in the DM framework, let us now turn to word-formation in Chinese. Let us assume that Packard is basically right in classifying word components in Chinese into four basic types: root words, bound roots, word-forming affixes and grammatical affixes. Root words are roots that can function independently as words, or in our terms, MWds, e.g., ma "horse", mai "buy". Bound roots are roots that must form words with some other word components, e.g., mu "wood", nao "brain". Word forming affixes are affixes that must attach to root words or bound roots to form new words, e.g.,. ke "-able" in ke-xing "feasible". Typical grammatical affixes are aspect markers,

e.g.,. zhuo "progressive", le "inchoative", guo "perfective". Some additional examples are given below:

(9) Affixes

a. Prefixes:

lao "lit. old", e.g.,. lao-wang "old Wang"

xiao "lit. young", e.g.,. xiao-wang "young Wang"

di "th?", e.g.,. di-si "fourth"

ke "-able", e.g.,. ke-xing "feasible"

b. Suffixes:

er "?" , e.g.,. yu-er "fishie?"

xue "study", e.g.,. wuli-xue "physics"

jia "lit. home, -ist", e.g.,. huaxue-jia "chemist"

hua "-ize", e.g.,. xiandai-hua "modernize"

(10) Roots

a. Root words: da "big", xiao "small", da "hit", qiao "bridge", etc.

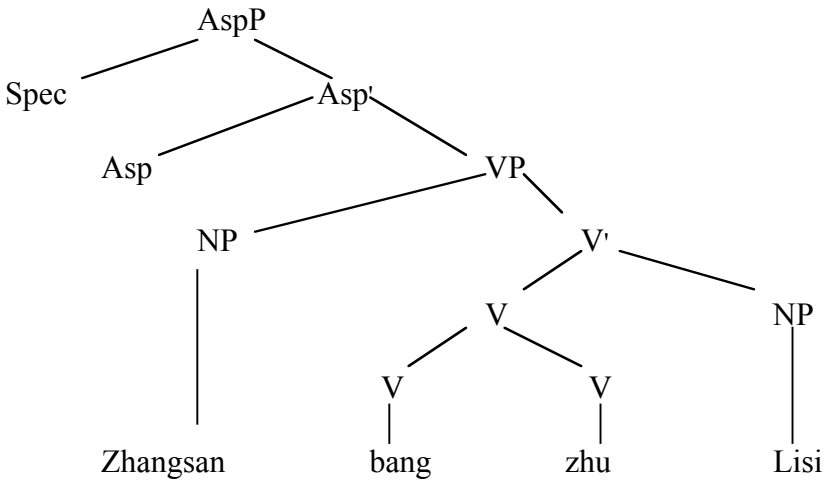
b. Bound roots: bi "wall", wu "matter", xing "go", gou "buy", etc.

Let us illustrate how Chinese words are formed under DM assumptions with the sentence in (11):

(11) Zhangsan bang-zhu-guo Lisi
Zhangsan help-help-PERF Lisi
"Zhangsan helped Lisi."

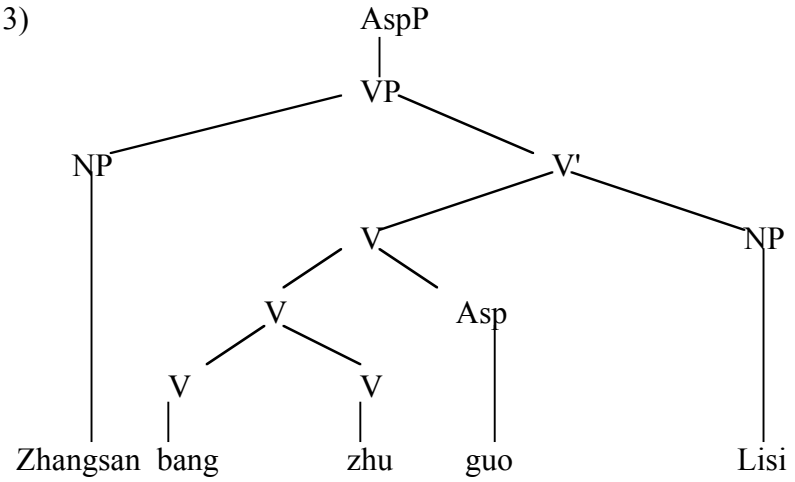
It is reasonable to assume that Chinese has a terminal node Aspect (Asp) which projects an Aspect Phrase (AspP), just like English has a Tense (T) which projects a TP. This Apsect Phrase dominates a VP:

(12)



For my purposes here I will not represent the morphemes as bundles of semantic, syntactic and morphological features. I will use the phonological realization instead. I will assume that bang "help" is head-adjoined to zhu "help" to form a complex X^0, or a MWd bang-zhu "help", in syntax. The structure in (12) then undergoes further derivation in Morphology, where Morphological Merger lowers the aspect marker guo "PERF" to merges it with the verb:

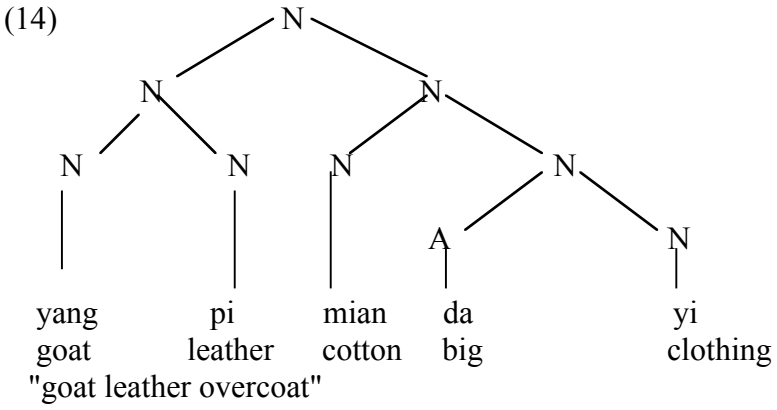
(13)



Let us consider how (12) and (13) capture generalizations of Chinese word-formation, especially those that are also captured by Packard's word-formation rules. His Rule 1 is captured by the head-adjunction in syntax, illustrated in (12). His Rule 2 is captured by Morphological Merger, which lowers the Asp guo "PERF" to adjoin it to the verb. In DM syntax feeds Morphology and therefore head-adjunction always precedes Morphological Merger, which is a morphological operation. As a result, the Grammatical affixes in the sense of Packard will always adjoin to the word after roots are adjoined. This effectively captures Packard's generalization that only root words can take Grammatical Affixes to form larger words. DM achieves this systematically by assuming that syntactic operations apply before morphological operations.

Just like Packard's Rule 1, head-adjunction can be applied recursively. The recursive nature of the word formation is thus captured:

(14)



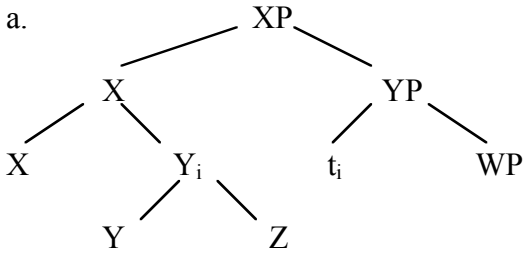
"goat leather overcoat"

Although word formation rules are recursive and thus are productive, their generative power is limited. Packard accounts for this limitation by stipulating that only root words can be formed recursively. That is, root words can form new root words by combining two roots words together but bound roots can not form new bound roots. Nor can affixes form new affixes by combining affixes together. In this the current framework, this can be accounted for by equating Bound Roots and Word-forming Affixes with Subwords. It is only natural that Subwords, as terminal nodes, can not be formed recursively.

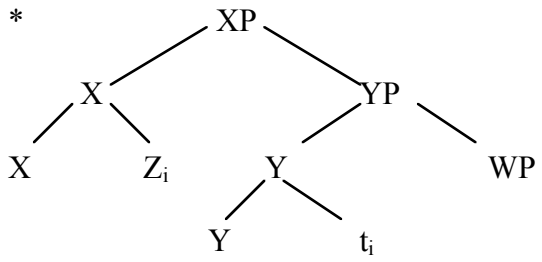
Having demonstrated how the current system captures the recursive nature of word formation, let us turn now to the facts that we have described in Chapter Two. Specifically, words generally demonstrate Lexical Integrity effects. What has been generally attributed to LIH means one of the two things. One is that

once a complex word is formed, its components can not be moved out of it. This is illustrated by the movement test. The other is that material can not be added to alter the structure of the word. This is illustrated by the expansion test and the XP substitution test. Let us see how these can be explained in the current framework. First, let us see why movement of a component is impossible. To do this, let us consider the operations before Morphology and the operations within Morphology respectively. In syntax, complex X^0s can be formed by head-adjunction. I assume that head-adjunction can occur with or without head-movement. If no movement takes place, naturally there will be no movement of anything out of a complex head. When there is head-movement, DM assumes that only the X^0 as a whole can be moved. This is illustrated in (15):

(15) a.



b. *



(15a) is allowed but (15b) is not because only a component of Y, Z is moved out of a complex word. In addition, the head can not move into a position that is occupied by an XP. For example, topicalization of a component of a word is banned since topicalization generally involves XP, not X^0. After syntax, morphological operations consider MWds to be islands and obey MWd Integrity. Morphological Merger only lowers a head as a whole to adjoin it to a lower head. It can not just lower a component of head, in the case of a complex head. As a result, Morphological Merger will treat a MWd as an island. Just as it is impossible to extract a component of a MWd, it is also impossible to delete a component of MWd. This is only natural if we consider movement to be a "copy and delete" operation. If we cannot copy and delete, of course we cannot delete. Similarly, we can tie pronominalization to movement. In the case of pronominalization, we can say a "copy and pronominalization" operation has taken place, in which the copy left behind is pronominalized. We would predict that it is impossible to pronominalize a component of a word.

The second part of the LIH is trickier. The basis on which to compare the MWd before and after expansion (by modification or otherwise) must be established. This comparison is different from the movement situation. In the case of movement, the comparison is "paradigmatic" in that it is between the possible spellouts of one underlying structure. In the case of expansion, we are

making syntagmatic comparisons in which we compare related structures with different underlying structures. Given the DM assumptions, we can have a general constraint banning XPs inside X^0s. If we do not allow XPs inside X^0s, it follows that expansion of a component of a word by assigning it a phrasal modifier or complement is impossible. However, it is possible for X^0s to take on morphemes as modifiers or complements. This has already been demonstrated by the recursive nature of word formation in Chinese.

We have shown that in DM the word formation function is taken out of the lexicon and implemented in syntax. By equating complex words with complex X^0s in syntax, we have shown that we are still able to account for the facts surrounding word formation in Chinese without loss of significant generalizations. We have demonstrated that the current system captures recursive nature of word formation. We have also shown that the wordhood tests that have motivated the notion of word in the first place can be captured in the present system by treating MWds as islands and by observing MWd integrity.

Another function of grammar that is often assigned to the lexicon in the lexicalist approaches is the ability to hold listed information, information that can not be derived in syntactic computation. In DM, such information is called idioms and a list called Encyclopedia is set aside to just hold these idioms. According to Marantz (1997), an idiom is a phonologically identified structure whose meaning is not predictable from the meaning of the subparts of the

structure along universal principles of interpretation of the structure. For example all roots are idioms since they are monomorphemic and have no subparts, e.g., da "big", xiao "small", etc.

A polymorphemic word whose meaning is not predictable from its subparts are also idiomatic, e.g., mao-dun "spear + shield = contradictory". However, not all words are idioms in this sense, e.g., tao-lun "discuss + discuss = discuss".

A phrase, which by definition has subparts, can either be idiomatic or not. For instance,

(16) Gua yang tou, mai gou rou
hang goat head, sell dog meat
"Say one thing and do another"

The above is idiomatic in the sense the meaning of the whole is not predictable from that of its subparts, although most phrases are not idiomatic. Since the meaning of the idiomatic expressions, phrases or otherwise, cannot be computed in syntax, they have to be listed somewhere, somehow. For the lexicalist approaches, the natural place is the lexicon. The lexicon thus becomes a heterogeneous linguistic module that not only holds morphemes, but also words, idioms, and word formation rules that look suspiciously syntactic. There is quite a bit of redundancy between syntax and lexicon. DM eliminates this redundancy

by forming words in syntax and setting aside a special place called Encyclopaedia to list idiomatic expressions. If a simpler model without redundancy is to be preferred, all other things being equal, this is an advantage for DM. Marantz (1997) insists that the argument should be made on empirical grounds but sometimes it can be difficult to tell an empirical argument from a conceptual argument. If there is some data Model A can account for but Model B cannot, this is not necessarily an empirical argument in favor of Model A, because Model B can always add a mechanism to account for it. In this case an empirical argument turns into a conceptual argument.

In the next chapter, I will explore further implications of the DM hypothesis. I will show that some of thefacts in Chinese word formation which cause problems for the lexicalist approaches like that of Dai and Packard can be accommodated easily in DM.

Chapter 4

ADDITIONAL ARGUMENTS FOR THE DM APPROACH

4.1 Introduction

In this chapter I will examine some interesting facts surrounding some complex verbs in Chinese. These are complex words that are formed by V+N, V+V and V+P. The V+N words are traditionally called "breakable compounds" as they have phrasal counterpart. I will show these pose problems for the lexicalist approaches as well for the Baker's incorporation-type of approaches. I will then show that the DM assumptions make a straightforward solution possible. Complex words like these thus provide the crucial evidence to differentiate the DM approach from Baker's incorporation approach. I will then show that the formation of the V+V and V+P words crucially needs the syntactic structure as their input and they provide arguments against the lexicalist approaches and for DM.

4.2 The V+ N Compounds

4.2.1 The Facts

In Chinese it is not uncommon to see expressions with the same phonological realizations to behave like words in one context and phrases in another. This is illustrated with the string dan-xin:

(1) a. ta hen dan-xin zhe jian shi
he very worry this CL matter
"He is very worried about this matter."

b. *ta dan-le san nian de xin zhejian shi
he carry LE three year DE heart this matter
"He worried about this matter for three years."

c. *xin, wo yi-dian dou bu dan zhe jian shi
heart, I one-bit all not carry this CL matter
(Lit. "-Ry, I don't wor- about this matter")

d. Q: ta dan-xin zhe jian shi ma?
he carry-heart this CL matter MA
"Is he worried about this?"

A: dan-xin
but
*dan

a'. ta dan xin.
he carry heart
"He was worried."

b'. ta dan le san nian de xin
he carry LE three year DE heart
"He worried for three years."

c'. xin, wo yi-dian dou bu dan.
heart, I one-bit all not carry
"I am not worried at all."

d'. Q: ta dan-xin ma?
he carry-heart MA
"Is he worried?"
A: dan-xin
or
dan

The above examples show that dan-xin in (1a), (1b), (1c) and (1d) should be considered to be a word. The example in (1b) shows that expansion of xin is not allowed. The example in (1c) shows movement of xin is forbidden. Finally, (1d) shows that it is not possible to answer the question with dan. These examples show that dan-xin should be considered a word.

In contrast, it is possible to expand xin in (1b'). (1c') shows that it is possible to move xin in a topicalization. (1d') shows that it is possible to answer the question with dan. These facts show that dan-xin here should be considered a phrase.

The only difference between the two sets of examples lies in the fact that in the first set of examples dan-xin is followed with a semantic theme as its object while in the second it is not. Other than that they mean exactly the same thing. It is clearly implausible to treat dan-xin in these two sets of examples as unrelated. Any theory would be inadequate without being able to account for this fact. It is important to point out that this is not an isolated phenomenon in Chinese: bang-

mang "help", bao-mi "keep secret", bao-xian "buy insurance for", cao-xin "worry", dao-luan "make trouble", fang-xin "stop worrying", fu-ze "be responsible for", fei-xin/shen "feel vexed about", guan-xin "care", ou-qi "sulk over", qi-hong "disturb", qing-shi "ask for instruction", tou-ji "opportunistic", zhu-yi "pay attention to", etc., all belong to this category.

In the next few sections we will first review how this phenomenon has been handled in previous works on this topic and show each of them have problems of their own. We will then provide an analysis under the DM assumptions and show that our approach is an improvement over the previous approaches.

4.2.2 Previous Approaches

4.2.2.1 Functional Approach

Li and Thompson (1981) consider constructions like dan-xin as breakable compounds. The term "breakable compounds" is not explanatory in any way from a formal perspective. Neither is it adequate descriptively. The expressions in question are not breakable compounds. Rather they are "breakable" in some syntactic context and "unbreakable" in others. In other words, they sometimes behave like words and sometimes like phrases. It is more appropriate to characterize them as a case of ambiguity between word and phrases.

4.2.2.2 Lexicalist Approaches

4.2.2.2.1 Isomorphism between Words and Phrases

For the lexicalist hypothesis where morphology and syntax are two subtheories each with its own atoms, what feed syntax are words. There are three possible ways to deal with this fact. The first possibility is to treat dan and xin as words in the lexicon, as pointed out by Huang (1984). There would be no problems for ta dan le san nian de xin but for ta hen dan-xin zhejian shi, given the inseparability that has been demonstrated above, the lexicalist hypothesis would have to invoke an *ad hoc* mechanism to "reanalyze" dan-xin as a word. This is the approach adopted by Huang (1984). Huang motivates this reanalysis mechanism with the Phrase Structure Condition (PSC) which he invokes to account for the well-formedness of phrase structures in Chinese:

The PSC: Within a given sentence in Chinese, the head (the verb or VP) may branch to the left only once, and only on the lowest level of expansion. (Huang 1984)

Assuming that the PSC (or something like it, A. Li's (1990) case theory approach) is a valid well-formed condition on Chinese phrase structures, the reanalysis is not necessarily the only logical fallout. For example, another way of satisfying the PSC is for xin to form a constituent with the noun phrase that

follows it. The "reanalysis" is not forced by the PSC, it is merely consistent with it. In addition, there are other cases of ambiguity between phrases and words that cannot possibly be accounted for by the PSC. For example, you-qian "have money" is shown by Fu (1999) to be a phrase when it has a verbal reading and a word when it has an adjectival reading:

(2) a. ta you qian
he have money
"He has money." or "He is rich."

b. ta you henduo qian
he have much money
"He has much money."

c. qian, ta you
money, he have
"Money, he has."

But

a'. ta hen you-qian
he very have money
"He is very rich."

b'. *ta hen you henduo qian
he hen have much money
"He has much money."

c'. *qian, ta hen you
money, he hen have
"Money, he has a lot."

In (2b) it is shown that qian can be expanded while in (2b') such expansion

is prohibited apparently due to the existence of hen. In (2c) it is shown that qian can be moved while in (2c') the movement of qian is banned. These facts cannot be explained by the PSC since it is not violated in (2b') and (2c'). Clearly explanation will have to be sought elsewhere. The reanalysis mechanism without any motivation like the PSC is nothing more than an *ad hoc* stipulation that is dubious in any theory. The second possibility would be to say that dan-xin is a word in the lexicon. This will not be a problem for ta hen dan-xin zhejian shi, but to explain ta dan le san nian de xin it will have to invoke a different, equally *ad hoc* mechanism to reanalyze dan-xin as two words. In fact, this is Packard's approach. The third possibility is to say that both dan, xin and dan-xin are words in the lexicon. In this case where there is a sentence like ta hen dan-xin, the syntax would have to decide what structure it is and which words to insert, dan-xin or dan and xin. There needs to be an independent principle to govern the choice of words to be inserted, as pointed out by Huang (1984). In addition, dan-xin would appear to be unrelated unless some add-on mechanism in the lexicon is invoked to relate them arbitrarily. In fact, this possibility is suggested by Jackendoff's redundancy rule in the lexicon. We will consider this possibility in Section 4.2.3.

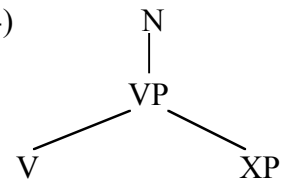
4.2.2.2.2 Phrase within Words

Di Sciullo & Williams (1987) proposes the concept of syntactic words,[4] which they consider to be phrases reanalyzed as words, to account for a similar phenomenon in French:

(3) essui glace
wipe windshield
"windshield wiper"

Taking a strong lexicalist position that contends "morphology and syntax are different (though) similar sciences about different objects," Di Sciullo and Williams (1987) proposes the following structure for expressions like the one above:

(4)



Di Sciullo & Williams (1987) justifies the dominating N node by observing that the words in question can be inserted into X^0 positions and display syntactic opacity. No syntactic rule can insert or move a category in the structure:

[4]Notice they use this term in a sense different from that of ours.

(5) a.*essui-between glace
wipe well windshield

b. *[glace essui e]

c. *Glace a été r'eparé cet [essui e] par Jean
windshield has been repaired this wipe by Jean

The dominated phrasal node can be justified by noting that the right-hand noun can be analyzed as an internal argument to the verb and together they form a verb phrase.

To the extent that this analysis is relevant to the Chinese examples under consideration, there are three problems with this analysis. First, the postulation of the dominated verb phrase implies that the verb phrase can undergo syntactic operations, if this VP is not any different from other VPs. This defeats the very purpose of positing the dominating N, which is to indicate the lexical integrity of the words. Second, the structure suggests that the "phrase to word" and the category switch (from V to N in this case) are the same process. This is inappropriate given the category switch and the "phrase to word" processes do not always co-occur. For example, in Chinese dan-xin is verbal either as a word or as a phrase. Also, there is a set of words in Chinese which can occur both as a verb or a noun and no switch from phrase to word is involved in these words. Third, the structure fails to account for the fact that the "word" always has a phrasal counterpart and thus fails to relate them as the same expression occurring in

different contexts. To establish that they are related, an adequate theory should be able to recognize that they are the same expression occurring in different contexts rather than simply different expressions and be able to identify the different contexts.

4.2.2.2.3 Co-licensing between Syntax and Morphology

Dai (1992) criticizes Huang's PSC as being a language-specific stipulation which makes wrong predictions. He points out that dan-xin can be a word where the PSC does not apply. For example, the following example show that dan-xin should be a word even if it is not followed by an NP:

(6) a. ta hen dan-xin
he very worry
"He was very worried."

b. *ta hen dan-le bantian de xin
he very carry-ASP half-day DE mind
"He was very worried for a while."

c. *xin, ta yizhi hen dan
heart, he continuous very carry
"He has been worried."

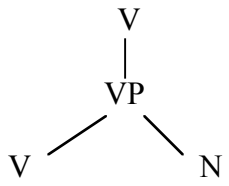
(Dai 1992:84-85)

Dai intended to show with the above examples that dan-xin should be treated as a word even if dan-xin is not followed by a NP. Dai argues that the lexical status of dan-xin in the above examples follow from syntactic and

morphological co-determination. Specifically, a phrase-like constituent can be analyzed as a lexical item if some syntactic construction refers to it as a syntactic atom (X^0 in the X-bar theory) and if the internal structure of this constituent observes the lexical integrity. In this case, since the syntactic construction above requires that the position be filled with an lexical item instead of a phrase (syntactic determination), and its parts (dan and xin) are unextractable and unexpandable (morphological determination), dan-xin should be reanalyzed as a word.

Dai (1992) implicitly adopts the same structure as proposed in Di Sciullo and Williams (1987):

(7)



and therefore suffers the same problems: In assuming a VP it allows the possibility that the VP takes adjuncts, the very possibility it is designed to avoid.

However I believe Dai is right in taking into account both the internal structure as well as the external structure. The problems with Dai's analysis, as well as other lexicalist approaches, stem from a word-based approach, thus necessitates reanalysis, an *ad hoc* move at best.

4.2.2.3 Syntactic Derivation

Fu et al (1999) observes that a number of you-constructions in Chinese demonstrate the same type of ambiguity between words and phrases as dan-xin. When the following expressions have the verbal reading, they tend to be phrases, but when they have the adjectival reading, they tend to behave like phrases:

(8) a. you-qian
have-money
"have money"
"rich"

b. you-wenti
have problem / question
"have problems / questions"
"problematic"

c. you-yunqi
have-luck
"have luck"
"lucky"

d. you-xingqu
have-interest
"have interest"
"interested"

e. you-kanfa
have-opinion
"have opinion"
"opinionated"

f. you-shuiping
have-level
"have a high level"
"highly competent"

All the expressions in (8) can be modified by adverbs of degree and they all have verbal readings as well as adjectival readings.

When they are modified by hen they have only adjectival readings and they behave like words. The phrase vs word contrast has been demonstrated above, and they are replicated in (9):

(9) a. ta you qian
he have money
"He has money." or "He is rich."

b. ta you henduo qian
he have much money
"He has much money."

c. qian, ta you
money, he have
"Money, he has."

But

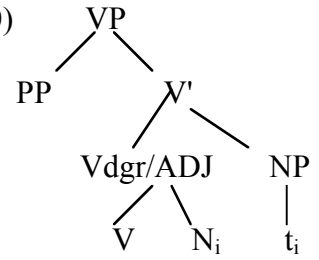
a'. ta hen you-qian
he very have money
"He is very rich."

b'. *ta hen you henduo qian
he hen have much money
"He has much money."

c'. *qian, ta hen you
money, he hen have
"Money, he has a lot."

Noting the problems with the lexicalist approach, Fu (1999) proposed the analysis below to account for the "lexicalization" of certain you constructions:

(10)

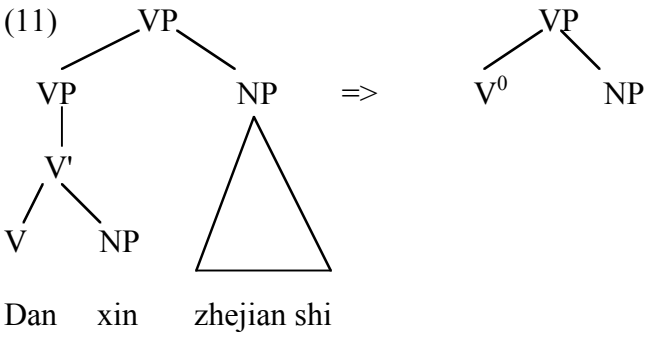


Specifically, Fu (1999) proposes that such lexicalization effect is induced by head movement of the N to a position that adjoins to V to form a degree verb. Given the analysis that V together with the moved N forms another V, assuming that head movement is a "word-formation" rule rather than a phrasal formation rule, this correctly predicts that no nominal modifier can occur to the left of N. However, it incorrectly predicts that this movement can leave possible NP adjuncts within the NP that is the complement to the verb.

Another problem with Fu's analysis is that Fu does not specify when such head movement can occur. Without specifying the exact context where such rules should be invoked, Fu does not make a distinction between the phrasal context and the word context and specify when the head movement should take place. Obviously, since these expressions behave like words only in some environment but not in others, an explanation is not adequate without noting the context in which head movement applies.

The third problem with Fu's approach is underscored when it is extended to account for dan-xin in the word context. Presumably a reanalysis like what is

illustrated in (11) will be necessary. They are two serious problems with (11). First, it will be necessary to allow VP to take further complement, which would be a significant extension of the X-bar framework. Second, massive reanalysis will be needed in order to derive the word structure from the phrasal structure:



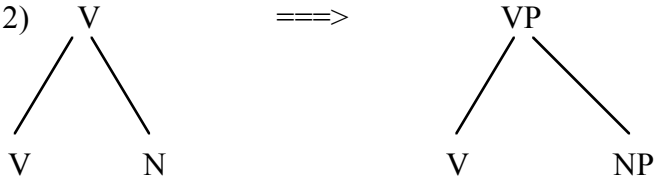
Technically Fu's approach is the same as Baker's incorporation analysis of American Indian languages. While an incorporation analysis is appropriate for Baker, the same does not readily apply in the analysis of the V-N compounds in Chinese. Unlike Baker's examples in Onondaga, Chinese V-N compounds do not leave N modifiers behind. Also, the noun does not have the referential transparency displayed there.

4.2.2.4 Packard's Reanalysis Approach

Packard (2000) suggests a reanalysis that is the opposite of that of Huang

(1984). He suggests that while dan-xin has the dual status of a phrase and a word, underlyingly it is always a word listed in the lexicon. It can be subjected to limited reanalysis as a phrase in syntax. Schematically, this can be represented as follows:

(12)

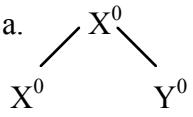


While this approach avoids the problem of dangling modifiers with Fu's incorporation approach, it is just as *ad hoc* as Huang's reanalysis approach. Moreover, it violates the LIH that is generally assumed to hold within words.

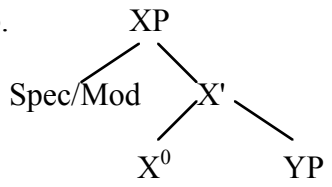
4.2.3 Distributed Morphology Approach

As we have discussed in the previous chapter, the crucial difference between Distributed Morphology and the lexicalist approaches lies in whether words can be formed by syntactic operations and where complex words are formed. Since for DM it is possible to form words in syntax, that means DM allows structures such as those in (13). Note we assume that head is initial with respect to the complement, but none of our arguments will hinge on this.

(13) a.



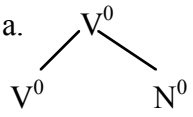
b.



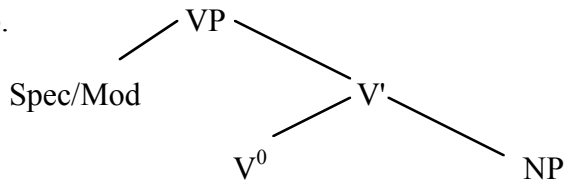
On one hand, we will have structures that observe the standard X-bar theoretical assumptions, such as (13b). On the other hand we would have structures like (13a) which resulted from the extension of the X-bar theory so that X^0 recursion is allowed.

Suppose X^0 is a verb of some kind and Y^0 is a noun, then we would have structures such as those in (14):

(14) a.

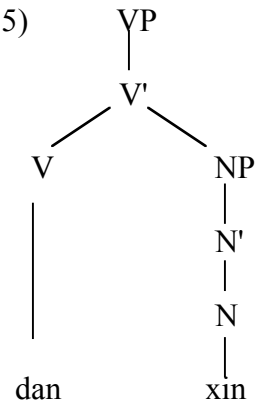


b.



Now let us look at the dan-xin. When it occurs in a phrasal context, the structure will look like (15):

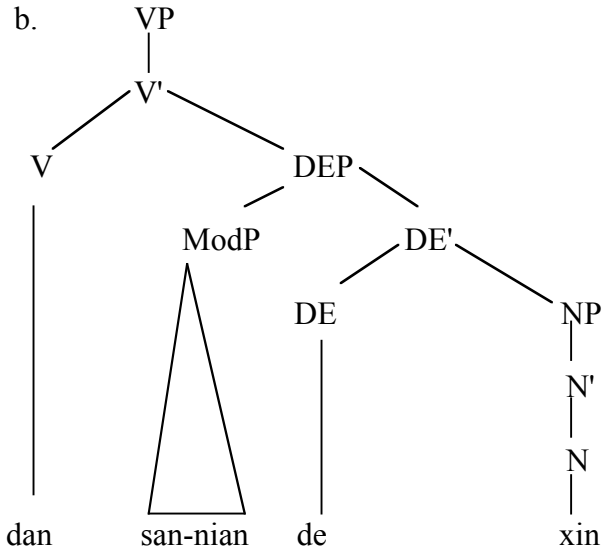
(15)



This structure predicts that it is possible for xin to occur in phrasal context, as in (16), in which case the structure becomes (16b):

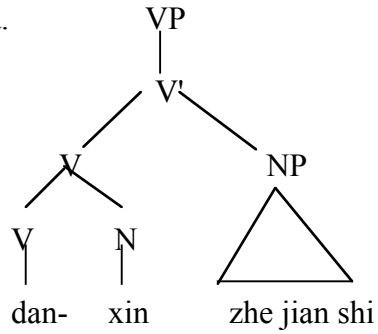
(16) a. ta dan le san nian de xin
he carry ASP three year DE heart
"He was worried for three years."

b.



When it occurs in a word context, the structure should look like (17a):

(17) a.



b. ta dan-xin zhe jian shi
he carry-heart this CL matter
"He was very worried about this matter."

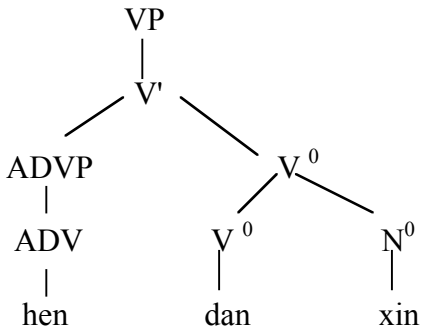
The other scenario where dan-xin occurs in a word context is when it has an adjectival reading. The adjectival reading is forced by hen, an adverb which modifies adjectives. This is illustrated in (18):

(18) a. ta hen dan-xin

he very carry-heart
"He was very worried."

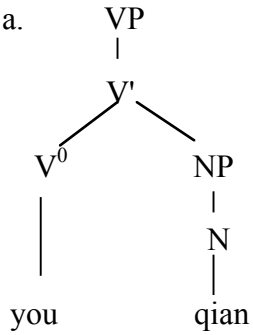
b. *xin, ta hen dan
Heart, he very carry
"He was very worried."

c.

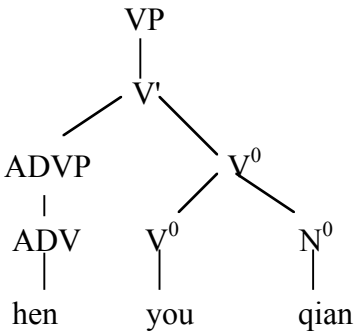


Here I still mark the higher V^0 as a verb since in Chinese there is no strong evidence for the category adjective. Given the analysis of the adjectival reading of dan-xin an analysis of the you construction follows in a straightforward manner. you-qian, when it has a verb reading, has the structure (19a), where it is a phrase. In contrast, when it takes on an adjectival reading, as in (19b), it is a word.

(19) a.



b.



Our analysis of the dan-xin and you-construction basically allows dan and xin (similarly you and qian) to form a phrase or a word in syntax depending on the context. The phrasal and verbal dan-xin are related in that they are formed via the same morphemes. Such an analysis is not available to the lexicalist approaches without cost because words are formed in the lexicon. In order for the lexicalist approaches to have a similar analysis, they will have to say dan, xin and dan-xin are all words in the lexicon. In order to relate the phrasal dan xin and the word dan-xin, the lexicalist approaches will have to resort to lexical redundancy rules in the sense of Jackendoff, which is not necessary in the DM framework. If

a simpler model is the superior model, this would constitute an argument for the present approach.

Now let us consider the relative strengths and weaknesses of this approach against the various the lexicalist approaches and Fu's "semi-lexical" approach in detail. For Huang (1984), Dai (1992) and other lexicalist approaches, the option of V and N forming another V is not available in syntax. Neither Huang (1984) nor Dai (1992) spells out in formal terms how the separable compounds should be achieved, but a "reanalysis" along the lines of Di Sciullo and Williams (1987) seems to be inevitable. As we have already demonstrated in previous sections of this chapter, the Di Sciullo and Williams style of structure incorrectly predicts that it is possible to have full-fledged phrases reanalyzed as words. Our analysis here is free of this problem since it only allows X^0-level complements or modifiers. Second, for any reanalysis to be viable, it will have to specify the triggering environment, otherwise it will not be able to tell when the reanalysis should or should not happen. Towards this end, Huang (1984) offers PSC while Dai (1992) offers syntactic licensing. As we have explained in previous sections, Huang's proposal is inadequate in that it only explains why the verb reading of dan-xin should be a word when it is followed by a complement and a phrase when it is not. It says nothing about the adjectival reading of dan-xin and you constructions. Dai correctly pointed out the inadequacies in Huang's formulation of the PSC and proposed to replace it with syntactic licensing, but he did not go

far enough to provide a formally workable formulation. In contrast, our assumption of transitivity is highly specific and workable and at the same time free of the problems with the PSC. Third, a Di Sciullo and Williams style of analysis would certainly be contradicting the assumptions of the X-bar theory with regard to bar-levels and therefore is not derivable from the X-bar assumptions. In this sense it is *ad hoc*. In contrast, our analysis stems from minor adjustment of the standard X-bar theory and no further stipulations are needed. Therefore, it must be concluded that the existence of this kind of ambiguity between phrasal and word structures is an argument in favor of the present analysis and against the reanalysis approach of the lexicalist hypothesis.

Now let us turn to a comparison between the present approach with Fu's (1999) "semi-lexicalist" approach. First, as we have demonstrated earlier, by attempting to derive words from phrases via head movement, Fu's analysis moves the head N in the complement NP and adjoins it to the head verb. This leaves open the possibility that there might be some dangling phrasal modifiers inside the complement NP. This is due to the fact that although Fu's approach allows the formation of Vdgr of V and N in syntax, the underlying structure only allows the head taking a phrasal complement. The present analysis, by allowing V to take an N complement in the syntax, and deriving the phrase from the word, does not have this problem. Second, Fu (1999) says nothing about when this type of head movement might take place. An explanation is incomplete without specifying the

triggering environment since it is not true that such head movement applies across the board, as we have demonstrated. Third, it is unclear as to how Fu's approach can be extended to explain the verbal reading of dan-xin. If dan-xin is base-generated as a phrase, how can it take on another complement, when it is transitive? Presumably one would have to reanalyze the VP that consists of dan and xin into an X^0-level element. This will revert to a position similar to the lexicalist approaches. The present analysis, by allowing V to take N as a complement, avoids this problem. Therefore, it safe to conclude that the present analysis therefore compares favorably against Fu's approach.

The crux of the present analysis is that it allows word formation to be done in syntax. This maximizes the combinatorial possibilities, which renders the reanalysis unnecessary. All other things being equal, this approach should be adopted.

4.3 Verb Resultative Compounds (V+V)

4.3.1 The Facts

Verb resultative compounds have been the topic of voluminous literature in Chinese linguistics (Chao 1967, Li and Thompson 1981, Li 1990; 1997, Dai 1992 and others). The verb resultative compounds are composed of a verb head followed by another verb (or preposition, which we discuss in the next section). The first verb, which is the head, generally denotes an action and the second verb

indicates the result due to the action of the previous verb. This is exemplified in (20):

(20) ta da-po-le chuangzi
he hit-break ASP window
"He broke the window."

In (20), the first verb da "hit" denotes an action as a result of which the window is broken, which is indicated by the second verb po "break". Note that although po is glossed as "break" it does not have an action denotation. There are at least two reasons for treating da-po as a word and not as a phrase. First, if an aspect marker is present, it can only attached to the whole word, as illustrated in (20), not to the first verb:

(21) *ta da-le-po chuangzi
he hit-ASP-break window
"He broke the window."

Second, nothing can occur between the two verbs, although the second verb can take a modifier when it occurs alone:

(22) a *ta da-quan-po chuangzi
he hit-totally-break window
"He broke the window."

b. chuangzi quan po le
window totally break ASP
"The window is completely broken."

The status of the verb resultative compounds as words is often contrasted with the phrasal status of the V-de constructions (Huang 1988, Li 1997). Like verb resultative compounds, the V-de constructions have a verb head followed by a result portion. Unlike the verb resultative compounds, the result portion of the V-de construction is not a verb. Instead it is a clause introduced by de, which is then incorporated into the verb, as we will show in Section 4.4.

(23) ta da-de chuangzi po-le
he hit-DE window break-ASP
"He broke the window."

Although almost all the authors touching on the subject assume that the resultative verbs are formed in the lexicon while the V-de constructions are formed in syntax, Li (1997) provides the most explicit arguments for such a dichotomous analysis. Li cites various differences between the V-de construction and the verb resultatives and argues that these differences are best explained by forming the verb resultatives in the lexicon and the V-de construction in syntax. In the sections that follow, I will first review Li's arguments. I will show that although Li's analysis of the V-de construction is basically correct but it does not necessarily lead to the conclusion that verb resultatives are formed in the lexicon.

I will show that the differences pointed out by Li can be equally attributed to a difference in structure between the verb resultative compounds and the V-de construction. After that I will show it is difficult to form some verb resultative compounds in the lexicon and they should be formed in syntax, as DM predicts. I will then try to derive both the verb resultative compounds and V-de constructions under the DM assumptions.

4.3.2 Li's Analysis

Li (1997) provided three arguments for the position that verb resultative compounds should be formed in the lexicon. Li's first argument is that the verb resultative compounds and their corresponding V-de constructions are ambiguous in different ways. For example, (24) shows that the verb resultative compound is three-way ambiguous while its corresponding V-de construction in (25) are only two-way ambiguous:

(24) Youyou zhui-lei-le Taotao le
Youyou chase-tired-ASP Taotao le
a. "Youyou chased Tao and as a result Taotao became tired."
b. "Taotao chased Youyou and as a result Taotao became tired."
c. "Youyou chased Taotao and as result Youyou became tired."

(25) Youyou zhui-de Taotao tai-bu-dong tui le
Youyou chase-de Taotao can't lift leg le
a. "Youyou chased Taotao and as a result Taotao couldn't move his (Taotao's) legs."
b. "Taotao chased Youyou and as a result Taotao couldn't move his (Taotao's) legs."

The notable difference between (24) and (25) is that (25) lacks a corresponding third reading which is that Youyou chased Taotao and as a result Youyou couldn't move his (Youyou's) legs. Li explains this by positing the following structure for (25):

(26) Youyou zhui-de Taotao$_i$ [pro$_i$ tai-bu-dong tui le].

According to Huang's (1989) generalized control theory, only the closest c-commanding NP, Taotao, can control (bind) the *pro* in the embedded clause. As a result, there is no way that the matrix subject Youyou can be identified with the pro and the sentence can only mean that Taotao is tired. Therefore, (24) can not be possibly be derived from an underlying structure like (26) via Baker-type incorporation otherwise the c reading of (24) cannot be explained. The verb resultatives should be explained along the lines of Li's (1990) theta identification analysis in which the arguments are identified freely with the each verb in the compound, subject to the thematic hierarchy.

I agree with Li that the verb resultative compounds should not be derived from an underlying structure like (26) and his theta identification analysis is correct. However, this does not automatically lead to the conclusion that verb resultatives should be formed in the lexicon. It is still possible that both the V-de construction and the verb resultative compounds are formed in syntax. It seems to

me that there is no reason why Li's theta-identification algorithm cannot be implemented in syntax. Whether the verb resultatives should be formed in the lexicon or syntax should be independently motivated.

Li's second argument is based on the analysis of ba. Compare (24), which was reduplicated as here as (27), and (28):

(27) Youyou zhui-lei-le Taotao le
Youyou chase-tired-ASP Taotao le
a. "Youyou chased Tao and as a result Taotao became tired."
b. "Taotao chased Youyou and as a result Taotao became tired."
c. "Youyou chased Taotao and as result Youyou became tired."

(28) Youyou ba Taotao zhui-lei-le
Youyou ba Taobao chase-tired-ASP
a. "Youyou chased Tao and as a result Taotao became tired."
b. "Taotao chased Youyou and as a result Taotao became tired."
c. *"Youyou chased Taotao and as result Youyou became tired."

The c reading of (28) is impossible because of ba, as Li assumes that ba must introduce a CAUSEE argument of a resultative construction and the CAUSEE must participate in the argument structure of the result portion of the resultative construction. Since the CAUSEE is Youyou in this reading and ba introduces Taotao, it is impossible. This would have been possible if (28) is biclausal and ba only operates on the first clause, since ba also introduces the object of a non-resultative construction:

(29) Youyou ba Taotao da-le
Youyou ba Taotao hit-ASP
"Youyou hit Taotao."

Thus Li reasoned that (27) must be mono-clausal and verb resultative compounds are formed in the lexicon.

However, as we have illustrated in Chapter Three, complex heads can be formed in syntax and therefore the fact that (27) is mono-clausal does not necessarily mean that verb resultative compounds are formed in the lexicon. Compounds like this can be formed through head adjunction, a point to which we will return.

Li's third argument concerns the interaction of anaphors with the resultative constructions. First Li tried to establish that the use of the anaphor ta-ziji makes a special "inversion" reading possible in the V-de construction:

(30) Youyou zhui-de Taotao$_i$ [lian ta-ziji$_i$ dou tai-bu-dong tui le].
Youyou chase-de Taotao even himself all can't lift leg le
Can mean: "Taotao chased Youyou and as a result even he (Taotao) himself couldn't move his legs."

This would not have been possible if there was no coreferentiality between the object in the matrix clause and the subject of the embedded clause, in which case the CAUSEE reading of the matrix object would not be possible:

(31) Youyou zhui-de Taotao [lian laoshi dou bu gaoxing le].
Youyou chase-de Taotao even teacher all not happy le
Cannot mean: "Taotao chased Youyou and as a result even the teacher became unhappy."

Since the inversion reading is possible in (30) and the object Taotao can be the CAUSEE, ba should be able to introduce the CAUSEE, given that ba must introduce the CAUSEE:

(32) Youyou ba Taotao$_i$ zhui-de [lian ta-ziji$_i$ dou tai-bu-dong tui le].
Youyou ba Taotao chase-de even himself all can't lift leg le
Can mean: "Taotao chased Youyou and as a result even he (Taotao) himself couldn't move his legs."

In contrast, Li showed that coreferentiality does not license a corresponding resultative compound:

(33) *Youyou ba Taotao$_i$ shuo-sao-le taziji$_i$.
Youyou ba Taotao scold-embarrassed-ASP himself
"Youyou scolded Taotao and as a result Taotao became embarrassed."
"Taotao scolded Youyou and as a result Taotao became embarrassed."

Taotao can be the CAUSEE so ba should be able to introduce it. There is no violation of case filter either since ba was showed by Li to be a case assigner:

(34) a. Youyou ba Taotao chang-wang-le xin-li de fannao.
Youyou ba Taotao sing-forget-ASP heart-inside de worry
"Youyou sang and as a result Taotao forgot his worries."

b. *Youyou chang-wang-le Taotao xin-li de fannao.
Youyou sing-forget-ASP Taotao heart-inside de worry

(34b) is bad because the case filter is violated and since ba is a case assigner (34a) is OK. The only reason (33) is bad must be that verb resultative compounds are formed in the lexicon and therefore when they are formed, the binding relation can not be established since binding is only relevant in syntax. If it is formed in syntax, there is not reason why (33) is bad.

The problem here is whether contrast between the two sentences in (34) is the result of the case assigning ability of ba and whether (34a) is OK because ba assigns an extra case that (34b) is lacking. If this were correct, we would also expect (35) to be also OK, which turns out to be a wrong prediction:

(35) *Youyou ba Taotao chang-wang-le xiao gou
Youyou ba Taotao sing-forget-ASP little dog
"Youyou sang and as a result Taotao forgot his (Taotao's) little dog."

Therefore ungrammaticality of (34b) as well as (33) cannot be due to case violation. A reasonable explanation is that it is due to the fact that verb resultative compounds do not allow more than two arguments. If this is the case we still have (34a) to explain.

I suggest that in (34a) Taotao and xin le de fanno is licensed through a topic-comment (Li 1976) relationship in a clause introduced by ba. In this analysis ba is considered to be a verb that takes a clause as its complement.

Readers are referred to Bender (2000) for arguments for a similar position. The topic occupies the clause-initial position and the rest of the clause is a comment about the topic. The topic-comment structure is very common in Chinese. For example, (36) is another frequently cited example:

(36) ta [$_{VP}$ ba [$_{CP}$ [$_{topic}$ juzi] [$_{IP\text{-}comment}$ bo-le pi]]]
he ba orange peel off-ASP skin
"He skinned an orange."

In (36) juzi is the topic and the rest of the clause is the comment. The comment has to be about to topic and such "aboutness" can be implemented in a number of ways. In this case this "aboutness" is crucially licensed by a whole-part relationship between the topic NP juzi and the object NP pi. If there is no relationship between them, the sentence will be bad and uninterpretable:

(37) *ta ba juzi bo-le pingguo
he ba orange peel off-ASP apple

If this is correct, then the grammaticality of (34a) can be accounted for since Taotao and xin le de fanno are related and the sentence is grammatical. In contrast, since Taotao and xiaogou are not related this way, the ungrammaticality of (35) is also expected. Therefore the ungrammaticality of (33) does not necessarily lead to the conclusion that binding relation does not apply here and verb resultative compounds are formed in the lexicon.

Taken all together, none of Li's arguments supports the position that verb resultative compounds are formed in the lexicon. In fact, I will show in the next section that some of V-V compounds have to be formed in syntax. I will then show how verb resultative compounds in general can be formed in syntax.

4.3.3 V-V Compounds that should be Formed in Syntax

In Chinese there is a group of V-V compounds which can only occur in some special syntactic constructions, specifically ba and bei constructions:

(38) meiguoren xuan Bush zuo zongtong
Americans elect Bush act as President
"Americans elected Bush to be the President."

(39) meiguoren ba Bush xuan-zuo zongtong
American ba Bush elect-act as President
"Americans elected Bush to be the President."

(40) Bush bei meiguoren xuan-zuo zongtong.
Bush bei Americans elect-act as President
"Bush was elected to be the President by the Americans ."

xuan-zuo can only occur in a ba construction (39) or a bei construction (40). When neither ba nor bei is present, the verbs that form the compounds have to occur separately, as in (38). Such words are not isolated phenomenon and there is a whole list of them:

(41) shi-wei "consider to be", dan-zuo "regard as", lie-ru "list-in", wuzhuang-

chengwei "arm and become", kanzuo-shi "consider to be", shen-ru "stick into", ding-wei "label as", fazhan-chengwei "develop into", lie-wei "list as", jianshe-cheng "build into", ji-wei "calculate as", hua-wei "transform into", ronghe-cheng "integrate as"

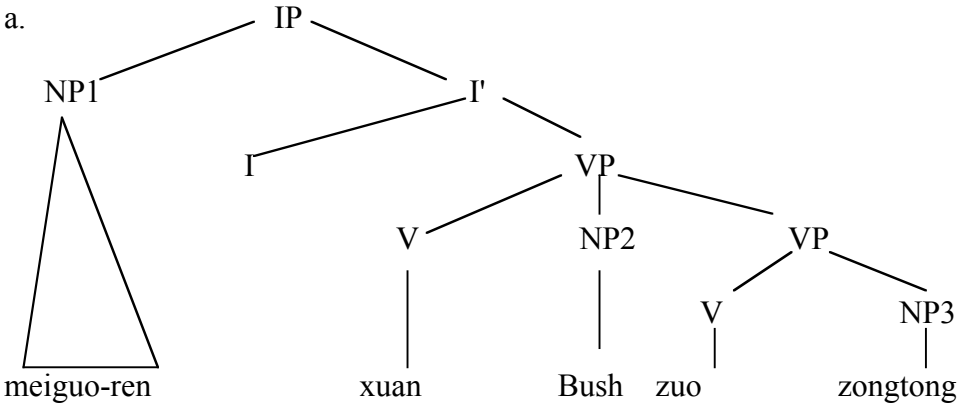
What is unique about these V-V compounds is that they cannot occur without the presence of ba or bei. If we assume with Baker (1988) that identical thematic relations should be assigned with identical structural relations, (38), (39) and (40) should be assigned identical structures with the underlying structure being something like (38). V-V compounds like xuan-zuo should then be formed in syntax. Forming words like this has another advantage. If they were formed in the lexicon, we would need some mechanism to record this dependency and guarantee that they will occur together with ba or bei. This can not be recorded as the sort of selectional restrictions such as those between a verb and its complements that are familiar for lexical items. Adding additional mechanisms will necessarily further complicate the already heterogeneous lexicon, as proposed in Packard (2000). Moreover, such words are highly regular and are likely to be infinite in number. Listing them will be impossible. They may be formed "on-line" with rules, as Packard suggests. The problem with that is that such rules, if formulatable, will be syntactic in nature. So, neither of the lexicalist approaches are attractive.

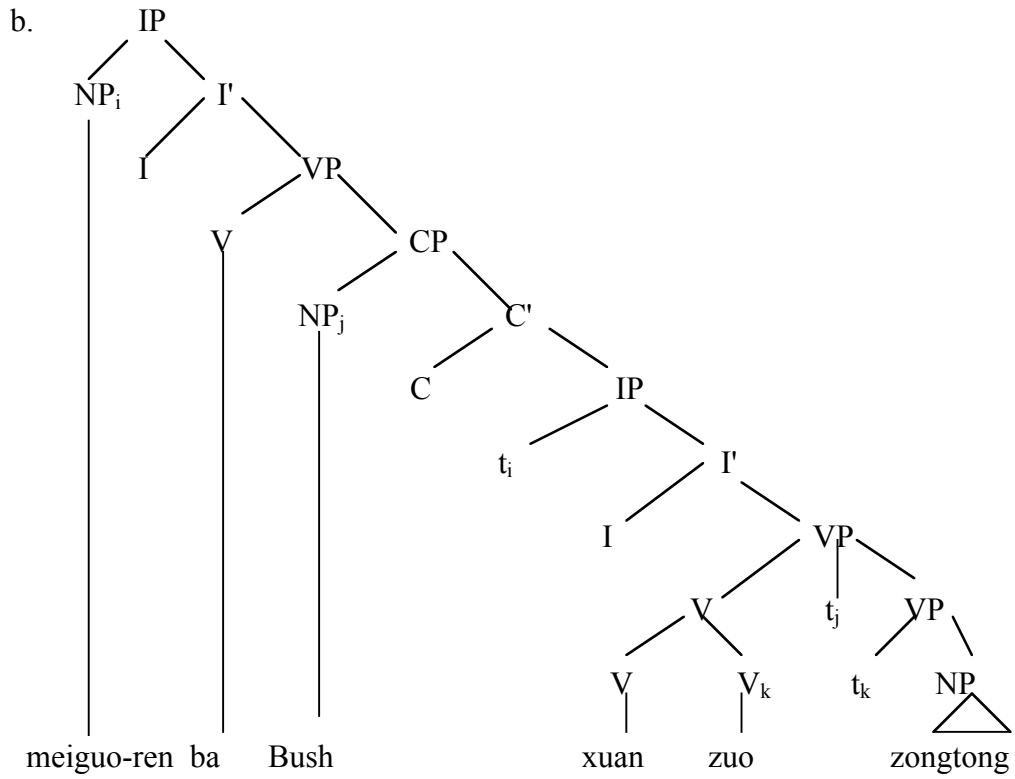
Let us see how such words might be formed in syntax. There are two obvious choices, one is Baker-type of head movement, and the other is

Morphological Merger in the sense of Marantz (1984; 1987; 1988).

First let us look at head-movement. Assuming the structure of (38) is (42a), applying head-movement in the ba-construction would give us a configuration like (42b):

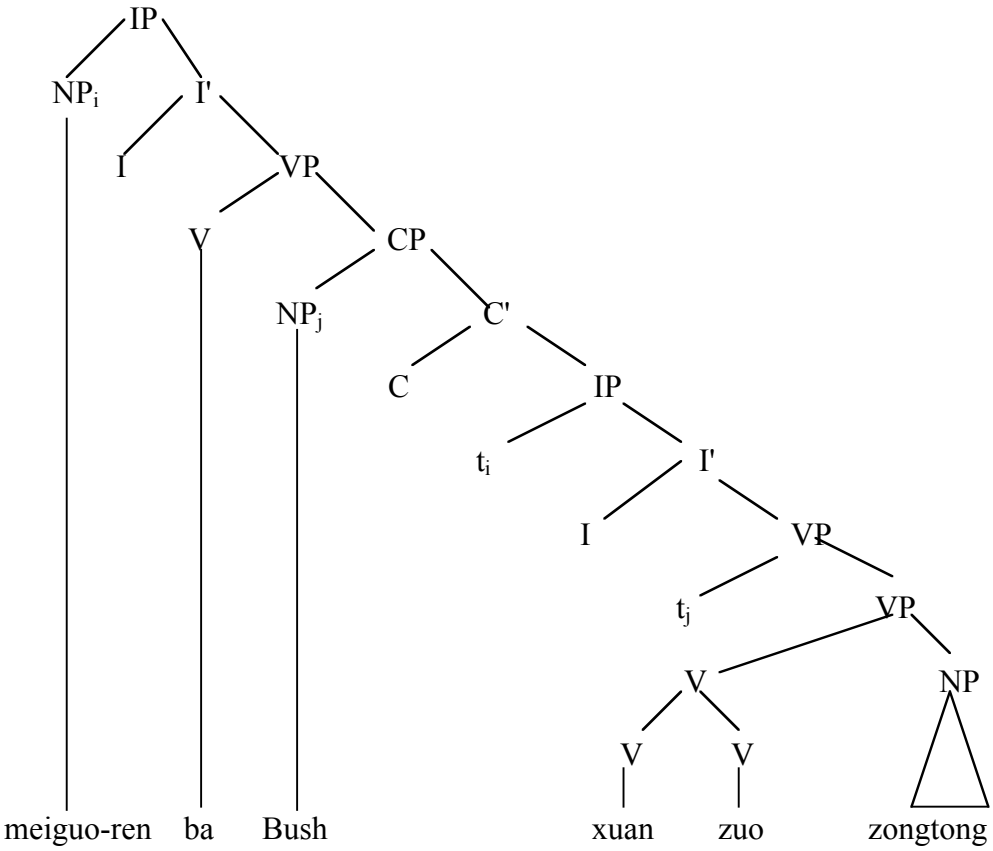
(42) a.





The verb in the lower VP moves up and adjoins to the higher verb. Head movement thus gives us the correct word order. Let us now look at Morphological Merger (lowering). The result of lowering the first verb will give us (43), which is also correct in terms of word order:

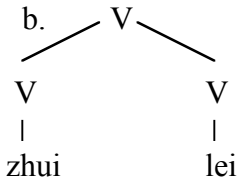
(43)



Since we assume that other complex verbs such as verb resultative compounds which do not have to occur in the context of ba are formed in syntax through head adjunction, as illustrated in (44), we will prefer head movement over Morphological Merger in this case so that we can have a unified account.

(44) a. Youyou zhui-lei-le Taotao

b.



The head adjunction in syntax has implications for the argument structure of each individual verb. Head-adjunction leads to thematic identification in the sense of Li (1990, 1997).

To summarize, I have shown that compound verbs can be formed in syntax and in fact some compound verbs have to be formed in syntax. Some complex verbs are formed by head-adjunction and when this happens, thematic identification occurs. Other complex verbs are formed via head movement. When head movement occurs, the underlying structure is not the so-called V-de construction. In fact, The V-de constructions themselves are formed via head movement of preposition, which I will discuss in the next section. If our analysis is on the right track, this will be an argument for our approach and against the lexicalist position that word formation takes place in the lexicon.

4.4 Preposition Incorporation

In the previous section, I showed how V-V compounds are formed. I also suggested that V-de constructions are formed by head movement. Let us see how

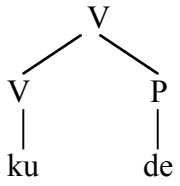
this can be implemented. I will assume with Li (1997) that de in the V-de construction is a preposition. In Chinese, it is possible for a preposition to take a clause as its complement. This is illustrated in (45):

(45) ta [$_{PP}$ [$_{P}$ wei] [$_{IP}$ chuli zhe jian shi]] qu-le xianggang
he for handle this CL matter go-ASP Hong Kong
"To deal with this matter, he went to Hong Kong."

Since de is phonologically close to dao, which is a preposition denoting goal, it is only natural that its meaning is extended to mean result, as suggested by Li (1997). When the verb is intransitive, the incorporation is vacuous and no word order change is observed:

(46) a. ta ku-de [IP shoupa dou shi-le]
he cry-de handkerchief all wet-ASP
"He cried (so badly) that his handkerchief was wet."

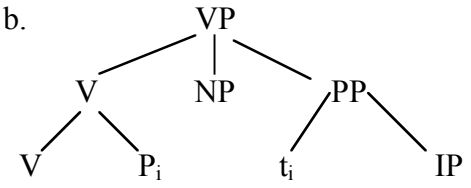
b.



However, when the verb is transitive, incorporation of the P will place the P in a position preceding the object of the verb:

(47) a. ta kua-de Taotao [IP pro hen gaoxing]
he praise-de Taotao very happy
"He praised Taotao and Taotao is very happy."

b.



Notice that preposition incorporation is also not an isolated phenomenon:

(48) kuoda-dao "expand to", tigao-dao "improve to", jizhong-dao "concentrate to", ronghe-yu "mix in", xiu-dao "build to", song-dao "send to", zhuan-xiang "turn to", tui-xiang "push to", tui-shang "push onto", kongzhi-zai "control at", yi-zhi "move to", jianshao-dao "decrease to", gao-shangqu "do up", qian-zhi "move to", dingwei-yu "designate as"

And like the V-V compounds, some of V-P compounds are also structure-dependent. Some of these complex words formed by a verb followed by a preposition can only occur together with ba:

(49) ta ba zhuyili zhuang-xiang-le xuexi shang
he turns focus turn to ASP study on
"He turned his focus to study."

Notice that the aspect marker, which generally attaches to verbs, can only occur after the preposition, an indication that incorporation has taken place:

(50) *ta ba zhuyili zhuang-le-xiang xuexi shang

If the V-P compounds were formed in the lexicon, there is no reason why they cannot occur alone without the presence of ba in the matrix clause. It is hard to state this property as a selectional restriction of these words since it is not typical of words to "select" a lexical item that is higher up in the matrix clause. Therefore, like V-V compounds, these V-P compounds must be formed in syntax.

4.5 Summary

In this chapter, I have shown that how V-N, V-V, and V-P compounds can be derived under the DM assumptions. I have shown that complex words in some cases must be formed in syntax. Taken together, these can be viewed as arguments for the DM approach and against the lexicalist approaches such as those of Dai (1992) and Packard (2000).

Chapter 5

AN AUTOMATIC WORD SEGMENTER

5.1 Introduction

In Chapters 3 and 4 I have argued for the position that Chinese word-formation is a syntactic process and should be dealt with in syntax proper rather than in the lexicon. I have argued against the various lexicalist approaches. The position that syntax is all the way down to morphemes is mainly motivated by the formation of complex words in Chinese, specifically, the formation of complex verbs. I have showed that word-formation in Chinese feeds on syntactic structure.

If our analysis is on the right track, we would expect that this has computational implications, since computational models of natural language processing (NLP) often mimic theoretic models in theoretical linguistics. For example, most current NLP models are "lexicalist" in that they use a computational dictionary that contains a list of words, with various levels of structural complexity. Some lexicalist models have more complicated dictionaries than others. It must be pointed out that there are also differences between computational dictionaries and theoretical lexicons of the kind described

in Packard (2000). For example, Computational dictionaries do not contain word-formation rules of the type Packard proposes for Chinese. The words that are listed are necessarily finite in number. A "lexicalist" approach in NLP also means that the processing of low-level information such as morphology does not interact with the high-level information such as syntactic structure except that the latter uses the former as its input. This is basically a pipeline approach where the low-level processing feeds the high-level processing.

In contrast, few computational models use a lexicon that contains morphemes, mimicking the Distributed Morphology approach we have discussed in previous chapters. What implications will the DM approach have for computational models if it is correct? Limiting ourselves to Chinese word-formation, what implications will the discoveries in the previous chapters have on automatic word identification in Chinese?

First of all, since what are listed are morphemes in the computational lexicon, not precompiled words, and the morphemes are finite in number, there will not be the "new word" problems that are often encountered in the "lexicalist" approaches (Sproat 1996, Wu and Jiang 1998; 2000). The "new word" problem arises because computational dictionaries can not possibly list all the words that are encountered in NLP tasks. When unlisted words occur, the "lexicalist" systems will have to come up with a*d hoc* ways to deal with them.

Second, since word-formation rules are syntactic in nature and they

manipulate syntactic objects which are generally classes of morphemes, we predict that syntactic information such as part-of-speech information is useful in automatic word identification. Given such syntactic information, one would be able to predict, to some extent, which words are possible and which are not. For example, we will see words formed by a verb followed by a preposition in Chinese, but we will not see words formed by a preposition followed by a verb.

Finally, our position that word-formation is an integral part of syntax predicts that the knowledge of syntactic structure helps with automatic word identification. We have showed that dan-xin can either be a word or a phrase depending on the syntactic context. Dan-xin attracts the attention of pure theoretic linguists because it poses a challenge to the lexicalist assumption that words are listed in the lexicon in a precompiled fashion and the lexicalist theorists have to come up with a way to characterize it. However, there are other cases where the syntactic structure is so obvious that it escapes the attention of theoretical linguists. For example, in (1)

(1) a. jing-cha [$_{VP}$ [$_{PP}$ [$_{P}$ yong] [$_{NP}$ qiang]] [$_{V}$ sha-le] [$_{NP}$ tao-fan]]
police with gun kill LE escapee
"Police killed the escapee with a gun."

b. *jing-cha [$_{VP}$ [PP[P yong]] [$_{V}$ qiang-sha-le] [$_{NP}$ tao-fan]]
police use gun kill LE escapee

c. jing-cha [$_{VP}$ [$_{V}$ qiang-sha-le] [$_{NP}$ tao-fan]]
police gun-kill LE escapee
"Police killed the escapee with a gun."

The presence of yong in (1a) makes it impossible to have qiang as a part of the verb qiang-sha. In other words, we can not have a structure like (1b). This is in contrast with (1c) where qiang is part of the verb qiang-sha. This is because in Chinese preposition stranding is impossible and it has to take an object as its complement.

It must be pointed out that linguistic information that is non-syntactic in nature also affects word identification. For example, in (2), correct word identification depends on world knowledge in general:

(2) a. ri-wen zhang-yu zeng-ma shuo?
Japanese octopus how say
"How do you say octopus in Japanese?"

b. ri wen-zhang yu zheng-ma shuo?
Japan essay fish how say

--adapted from Sproat (1996)

Given the possible words in Chinese, it is possible to segment the sentence in two ways, (2a) and (2b). (2a) is the obvious interpretation for humans given (2b) makes very little sense.

In the remainder of the chapter I will describe an automatic word segmenter that implements our theoretical approach by (i) using morphemes (Chinese characters) as input, (ii) taking limited syntactic structure into consideration and (iii) using part-of-speech information. The segmenter works by

learning a set of word-formation rules and applying those rules to form words. It is worth noting that these rules are not of the kind that are familiar in lexicalist morphology where word-formation rules typically combine stems and affixes into words. They are simple operations that combine morphemes into words, just like syntactic operations that combine words into phrases. I will evaluate this implementation by comparing its results with that of a "lexicalist" implementation that uses a computational dictionary which contains a list of words. I will first review the current approaches in Chinese word segmentation in Section 5.2. In Section 5.3 I will discuss various aspects of the transformation-based error-driven algorithm proposed by Brill (1993), which we will use to implement our morpheme-based segmenter. In Section 5.4 I will evaluate this implementation by comparing our results with that of previous implementations that also use this algorithm and also by comparing the results of this implementation with that of the "lexicalist" implementations using the maximum matching algorithm. Finally, in Section 5.6 I will summarize this chapter.

5.2 Previous Work

"Lexicalist" approaches that use a computational dictionary face two outstanding problems in Chinese automatic word identification. The first problem is ambiguity. Given a sentence, there is more than one way to compose the sentence with the words found in the dictionary, although not all of the possible

interpretations are syntactically or semantically feasible. This has been illustrated in (1) and (2). For example, (1a) is grammatical but (1b) is not. Similarly the segmentation in (2a) makes sense but (2b) does not. The second problem is the so-called "new word" problem. As we have pointed out, there is no way for any computational dictionary to list all possible words. Various methods have been proposed to tackle these problems in previous work on Chinese word segmentation. These fall into three main categories: pure statistical approaches (Sproat and Chih 1990), non-statistical dictionary-based approaches (Liang 1986, Gu and Mao 1994 and many others), statistical and dictionary-based approaches (Sproat and Chih 1996). More recently work on Chinese word segmentation also includes transformation-based error-driven approaches (Palmer 1997, Hockenmaier and Brew 1998).

As a representative of pure statistical approaches, Sproat and Shih (1990) relies on the mutual information of two adjacent characters to decide whether they form a two-character word. Although this approach has the advantage of not needing a dictionary, such an approach generally does not work very well in terms of accuracy of the segmentation. This result is expected to a large extent. It is hard to imagine how ambiguity could be handled successfully in this approach. Still using dan-xin as an example, these two characters should be treated the same in terms of mutual information no matter where it occurs, and whether they should be grouped as a word or not really depends on the syntactic context.

Statistical dictionary-based approaches generally try to exhaust the possible segmentations of a sentence using words listed in the dictionary and then calculate the most probable segmentation. The most probable segmentation is then chosen as the preferred segmentation. While this approach produces satisfactory results, it is not easily comprehensible by humans. The results do not lend themselves easily to linguistic analysis. In addition, there is no straightforward mechanism built on the statistical approach which can be used to predict new words. Pure dictionary-based approaches generally ignore the inherent ambiguities of the Chinese language by using some heuristics such as the maximum matching method. A segmenter that uses the maximum matching algorithm walks through a sentence trying to find the word that has the longest string of characters that is listed in a dictionary. In effect, this approach makes an arbitrary decision as to what should be considered to be a word. For instance, assuming dan, xin, dan-xin are all listed in the dictionary, the maximum matching algorithm will always favor dan-xin as a word, over dan-xin as a phrase. This is because dan-xin is a longer string than dan. When the segmenter finds dan, it will continue to see if there is a possible extension. When it finds there is another word dan-xin in the dictionary it will decide against inserting a word boundary between dan and xin. It is obvious that none of the above approaches have a way of dealing with words that are not listed in the dictionary.

In general, most current implementations rely on a dictionary of words and

are therefore rooted in the lexicalist assumptions. The completeness of the dictionaries often affects the degree of success of the segmenters. Recent work on Chinese word segmentation has used the transformation-based error-driven algorithm (Brill 1993) and has achieved various degrees of success (Palmer 1997, Hockenmaier and Brew 1998). Although the actual implementation of this algorithm may differ slightly, in general the transformation-based error-driven approaches try to learn a set of rules from the training corpus and apply them to segment new text. The use of the transformation-based error-driven algorithm provides an attractive alternative that can easily accommodate our theoretical assumptions. It is consistent with our theoretical assumptions because (i) it does not use a computational dictionary, (ii) it provides a natural way to take syntactic context into account and (iii) it provides a natural way to use part-of-speech information. It is attractive because, like statistical approaches, this approach provides a trainable method to learn the rules from a corpus and it is not labor intensive. In addition, these rules are transparent compared with the statistical approaches. That is, they are easily comprehensible by the humans and readily lend themselves to linguistic analysis.

5.3 Transformation-Based Error-Driven Approach

5.3.1 Background

The transformation-based error-driven algorithm is a machine learning

routine first proposed by Brill (1993) and initially used in POS tagging as well as parsing with varying degrees of success. It has been used in Chinese word segmentation by Palmer (1997), Hockenmaier and Brew (1998). The core of this approach is a learning routine that learns a set of rules that can be used to segment new text. Below I briefly describe the four important aspects of this learning routine as it is used in word segmentation, namely the type of input, the learning algorithm, the rule templates and the evaluation functions. I will then show how we can use this algorithm to implement our theoretical assumptions.

5.3.2 Types of Input

The input to the learning routine is a (manually or automatically) segmented corpus as the reference and its unsegmenteded (or undersegmented) counterpart. While the segmented corpus that serves as the reference should remain constant, the unsegmented or undersegmented corpus can be adjusted along two parameters. First it can either be POS-tagged or untagged. In general, words / characters in the POS-tagged corpus are classified based on a predefined set of categories and a set of rule templates can be designed to exploit these categories. The result will be more general rules that are defined over classes of characters rather than individual characters. Second, the level of segmentation of the Chinese text can vary from no segmentation at all, characters as words, or the output of another (less accurate) segmenter which serves as the preprocessor.

5.3.3 The Learning Algorithm

The learning algorithm compares the segmented corpus and the undersegmented (and tagged) dummy corpus at each iteration and finds the rule that achieves the maximum gain if applied. The rule with the maximum gain is the one that makes the dummy corpus most like the reference corpus. The maximum gain is calculated with an evaluation function which quantifies the gain and takes the largest value. The rules are instantiations of a set of pre-defined templates. After the rule with the maximum gain is found, it is applied to the dummy corpus, which will better resemble the reference corpus as a result. This process is repeated until the maximum gain drops below a pre-defined threshold, which indicates improvement achieved through further training will be no longer significant. The training will be then terminated.

The output of the training process would be a ranked set of rules instantiating the predefined set of templates. The rules will then be used to segment new text.

5.3.4 Designing the Rule Templates

The rule templates predetermine the possible rules that can be learned. The designing of the templates should target the most useful information that helps word segmentation. For instance, if a two Chinese character sequence C1 C2 never forms a word together in that order then splitting the two characters

would be a useful rule. A rule template that can capture this can be something like "split C2 from C1". C1 and C2 would be variables that range over all the Chinese characters and all instantiations of these templates would be possible rules. The rule templates can also be defined over tags which indicate classes of characters. This is possible because most Chinese characters are also words in some context and therefore can be assigned part-of-speech tags.

5.3.5 Adjusting the Evaluation Function

The learning routine needs to use an evaluation function to quantify the gain of each rule and determine which rule is the best. The rules, when applied, can affect the unsegmented (or undersegmented) corpus positively (make it more like the reference corpus) in some cases and negatively (make it less like the reference corpus) in others. The evaluation function can either take into account the positive effect only, or both. One way of calculating the positive effect only is to count the number of cases where the rule makes the right corrections and one way of calculating both the positive and the negative effect is to use the cases where the rule makes the right corrections minus the cases where the rule makes the wrong corrections and divide the difference by the total number of corrections for that rule.

5.3.6 Implementing our Theoretical Aassumptions

Having described how this algorithm works when used in Chinese word identification, let us now see how we can use this algorithm to implement our theoretical assumptions. First of all, let us see how to make our implementation morpheme-based. It turns out this is very easy since in Chinese each character is roughly a morpheme. We can make our implementation morpheme-based by using characters as input and group them into words with word-formation operations. These word-formation operations will be in the form of simple rules that combine the morphemes together. Word-formation in our sense is different from the way it is commonly used and it is not a rule that combines stems and affixes, since in this implementation we do not classify the morphemes into stems and affixes. Compared with the dictionary-based approaches, our word-formation rules provide a much richer mechanism to capture linguistic information than the words in a computational dictionary. In a sense, the lexical entries in a Chinese dictionary specify a special type of word formation rule. If "A", "B" are characters and "AB" is an entry in the dictionary, then this entry is equivalent to a word formation rule "merge 'A' and 'B' when they are next to each other". If this is all there is in Chinese word segmentation, then there is no difference between a dictionary and a set of word formation rules. It is safe to say that word formations rules can capture all the benefits of a dictionary. However, when more complicated scenarios are considered, the word formation rules become more

useful than a dictionary.

Second, we can design the rule templates in such a way that syntactic information is taken into consideration. As I have shown previously, strings of Chinese characters can be ambiguous between phrases and words and this underscores the difficulty of the dictionary look-up approach. Suppose "A", "B" are Chinese characters and "A", "B", and "AB" are all dictionary entries. When there is a sequence of "A" and "B" in the text, the segmenter will have to decide whether "A" and "B" should be combined to form a single word or left alone as two separate words. In this case the segmenter cannot rely on the dictionary: both are possibilities. In this situation humans would rely on the context in which "A" and "B" occur to decide whether or not they should be combined. For example,

(3) ta [tou tong]
he head ache
"He is having a headache."

(4) ta de [tou] [tong] (shou bu tong)
he DE head ache hand not ache
"His head aches but his hand does not."

In (3) tou-tong is one word since tou is no longer referential. In contrast [tou] [tong] in (4) are two words. Clearly this is an impossible situation for a pure dictionary look-up approach. Most current word segmenters deal with this using some kind of heuristic. The maximum matching approach would always combine

when "A" and "B" are next to each other and "AB" is a lexical entry in the dictionary. While it is almost impossible to make all the syntactic context accessible to the segmenter -- to do so would require a full parse of the sentence, which is a difficult task itself -- it is possible to make some contextual cues available to the segmenter by adding a conditional statement to the word formation rules. One such rule could be "combine tou and tong when they are not preceded by de". In this way we will correctly predict that tou tong are two words in (4).

Third, it is easy to design rule templates that refer to word classes. We can do this by first assigning a part-of-speech tag to each character. This is possible, since, as we have discussed in previous chapters, it is possible to assign word categories to morphemes in Chinese as they are often words themselves in other contexts. Of course if we tag the characters in the text they better be tagged correctly so that the right rules can be triggered. This is possible because a substantial number of characters in Chinese are not ambiguous and they can be guaranteed to be tagged correctly. Suppose we have the rule "combine a character that is a verb and a character that is a preposition", we will be able to combine xiang "think" dao "to" into a word xiang-dao "think of".

Since our implementation will be morpheme-based and uses word-formation rules to form words and since the morphemes (characters) in Chinese can be exhaustively listed, if our word-formation rules are correct, we should not

encounter new words. Properly constructed word formation rules can capture words that are generally not listed in any dictionary but are derivable from productive syntactic processes. For example, although it may not be possible to find the word leng-she "shoot abruptly" in any dictionary, one can predict such possibilities with a word formation rule V-> ADV + V; however, it is likely that our approach will generate words that are not attested since we use rules that are over-productive. We deal with this with a different set of rules to undo over-generation. If in some context the merge operation generates a string that is too large, split rules can be used to undo the merge. For example, if merge mistakenly forms "AB C" when it should generate "A BC", an split operation can then undo this and get it right. The split operations can be viewed as context-conditioned adjustments.

It is a property of this learning algorithm that these word-formation rules can be ranked to achieve the best effect. For example, given {A, B, C} are characters in Chinese and {A, C, AB, BC} are words, if in a particular context a sequence of A, B, and C should be correctly segmented as "AB C", it is possible for a segmenter using the word formation rules to get it right by ranking the rule "merging A and B" higher than the rule "merging B and C". Generally, the more general rules are ranked higher than the less general rules.

To summarize, it is possible to implement all of our theoretical assumptions with this algorithm.

5.4 Evaluation

5.4.1 Previous Work in Transformation-Based Approach

Palmer (1997) used a corpus of 2,000 sentences (roughly 60,187 words) for training and 560 sentences (roughly 18,783 words) for testing. The corpus was taken from the Xinhua newswire and he performed four experiments, which differ only in the choice of the initial segmentation. The initial segmentation was respectively done by a character-as-word algorithm, a maximum matching algorithm, a maximum matching algorithm with unknown character sequences as individual words, and the NMSU CHSEG segmenter. The results are presented in Table 1:

Table 1. Palmer's results

Initial algorithm	Character as word	Max-match	Max-match2	NMSU
Transformations	5903	1897	2450	1755
F-measure (initial)	40.3	64.4	82.9	87.9
f-measure (final)	78.1	84.9	87.7	89.6

Hockenmaier and Brew (1998) conducted three experiments using 100,000 words from Guo Jin's Xinhua News Agency corpus as training data and another 25,000 words from the same corpus as testing data. They used a character-as-word segmentation as the initial input. The experiments differ in the

use of rule templates. Experiments 1 and 2 used simple bigram rules while Experiment 3 used more elaborate trigram rule templates. The results are listed in Table 2:

Table 2. Hockenmaier and Brew's results

templates	bigram	bigram corrected	trigram
Transformations	7635	7523	7442
F-easure(initial)	42.3	42.02	42.2
F-measure(final training)	97.33	97.89	98.59
F-measure (final testing)	87.1	87.39	87.86

Hockenmaier and Brew also conducted preliminary experiments to determine whether part-of-speech tags can be used to help segmentation. They used the Concise Oxford English-Chinese Dictionary as a reference to assign initially the most common part-of-speech tag to the characters. After no further improvement is possible, the initial tagging is changed for less common tags where segmentation failed. This experiment was restricted to a corpus of 1700 characters (1110 words). For these experiments, in addition to using rules that are defined over characters, there are also rules that are defined over tags which indicate classes of characters. For all three experiments they achieved complete accuracy (100% accuracy) with their segmentation of the training data. However

they did not try the rules with new data, due to their small set of rules and their small training corpus.

In order to determine whether part-of-speech information is useful in segmenting new text, one obviously needs to conduct further experiments. We will explore this possibility in the next section.

5.4.2 Our Experiments

Hockenmaier and Brew's experiments showed that even with simple character-as-word segmentation as the initial state, the transformation-based error-driven approach can work very well with Chinese word segmentation. However there are two aspects in which their implementation can be improved. One is that their rules are generally triggered by the presence or absence of individual characters and there are no rules which are triggered by strings of characters or words. For instance, the rule "delete the segmentation sign between C_i, C_{i+1}, C_{i+2} if $C_i = char_m$, $C_{i+1} = char_n$ and $C_{i+2} = char_0$" is conditional on individual characters $char_m$, $char_n$ and $char_0$ rather than on a string of characters $char_m$, $char_n$ and $char_0$ for instance. There is no straightforward way of capturing cases where the segmentation is decided upon a continuous string of characters without segmentation signs between them. The other aspect of Hockenmaier and Brew's implementation that needs improvement is the use of part-of-speech tags as a means of improving segmentation accuracy and the compactness of the

training models. Since Hockenmaier and Brew used a very small corpus of 1700 characters (1100 words), and did not test their models on new text, there is serious question as to whether their models can scale up and be used in segmentation of new Chinese text, as they pointed out themselves.

For all our experiments we used data from the Penn Chinese Treebank (Xia *et al*, 2000) which contains 100,000 manually annotated words. The source of the data is also the Xinhua newswire. We used 78,674 words of the corpus as training data and the remaining 22,273 words as testing data. The training data are less than Hockenmaier and Brew's 100,000 words in their first three experiments.

5.4.2.1 Experiment One

The first experiment used a left-to-right maximum matching algorithm and our only interest in it is that it could be used as a benchmark to evaluate the performances of the models in the second and third experiments. Although there are other Chinese word segmenters that use the maximum matching algorithm, it is still hard to use them to evaluate our results since the words are annotated with different standards and the corpora they use are not available to us.

The maximum matching algorithm uses the dictionary compiled from the training data. Therefore there were no new words in the training corpus.

5.4.2.2 Experiment Two

For our second experiment we used the character-as-word segmentation as the initial state. In addition we used only the rules defined over the characters as well as strings of characters. These rule templates are listed in below.

mergeleft

Merge the current character with whatever character is on the left

mergeright

Merge the current character with whatever character is on the right

splitfromleft

Split the character from whatever character is on the left

splitfromright

Split the current character from whatever character is on the right

mergeleft if p1=x

Merge the current character with the previous character if the previous character is x

mergeright if p1=x

Merge the current character with the next character if the previous character is x

mergeleft if n1=x

Merge the current character with the previous character if the next

character is x

mergeright if n1=x

Merge the current character with the next character if the next character is x

splitfromleft if p1=x

Split the current character from the previous character if the previous character is

x

splitfromright if p1=x

Split the current character from the next character if the previous character is x

splitfromleft if n1=x

Split the current character from the previous character if the next character is x

splitfromright if n1=x

Split the current character from the next character if the next character is x

mergeleft if p2p1=xy

Merge the current character with the previous character if the previous two characters are xy

mergeright if p2p1=xy

Merge the current character with the next character if the previous two characters are xy

mergeleft if n1n2=xy

Merge the current character with the previous character if the next two characters are xy

mergeright if n1n2=xy

Merge the current character with the next character if the next two characters are xy

splitfromleft if p1p2=xy

Split the current character from the previous character if the previous two

characters are xy

splitfromright if p1p2=xy

Split the current character from the next character if the previous two characters are xy

splitfromleft if n1n2=xy

Split the current character from the previous character if the next two characters are xy

splitfromright if n1n2=xy

Split the current character from the next character if the next two characters are xy

5.4.2.3 Experiment Three

For Experiment Three, in addition to using the rule templates that are used in experiment two, we also used three rule templates that are defined over tags which indicate classes of characters. These are listed below:

Merge the two characters if the previous character is tagged as T1 and the current character is tagged as T2

Merge the two characters if the previous character is tagged as T and the current character is C

Merge the two characters if the previous character is C and the current character is tagged as T

In order to avoid generating the massive strings of characters during the transformation that has been reported in Hockenmaier and Brew (1998), we made sure that these rule templates are only applicable when the adjacent characters have segmentation signs on both sides.

The third experiment also differs from the second experiment in the evaluation function used to rank the rules. As we have discussed above, the rules, when applied, can affect the unsegmented (or undersegmented) corpus positively (make it more like the reference corpus) in some cases and negatively (make it less like the gold standard) in others. Choosing the right evaluation function is therefore very important. In Experiment Two, we used an evaluation function which calculates the total positive changes minus the negative changes. This evaluation function effectively ranks the more general rules higher than the more specific ones. However, this evaluation function becomes undesirable for Experiment Three since the rules defined over tags are more general and therefore are triggered much more often than the rules that are defined over characters or

strings of characters. In effect this would prevent the rule templates defined over characters from being instantiated. To remedy this situation, in Experiment Three we used a slightly different evaluation function. We used the function that calculates the number of positive changes minus the number of negative changes divided by the total number of corrections. By normalizing the evaluation function used in Experiment Two, we were able to rank the most "correct" rules higher, that is, rules which make the maximum positive corrections and at the same time make the minimum negative corrections.

Instead of getting the tagging information from a lexical source as Hockenmaier and Brew did in their experiments we tagged the characters with the Brill tagger (Brill, 1993) retrained on Chinese. The correct tagging as a by-product of segmentation was not of particular concern to us, however it is perhaps reasonable to assume that accurate tagging would help segmentation. Still we did not retag the corpus during the training process, except to get rid of the tag of a character / string when it is merged with the following character / string, assign the tag of the whole string to all the substrings where there is a split operation.

The results of the three experiments are summarized in Table 3:

Table 3. The results of our first three experiments

setup	Max-match	Character only	Character + tag
transformations	Not applicable	3292	6530
F-measure (initial)	31.6	31.6	31.6
F-measure (final training)	95.65	90.85	95.22
F-measure (final testing)	85.06	88.33	90.24
Total new words	1745	1745	1745
correctly segmented new words	55	718	754

5.4.2.4 Experiment Four

The setup of the fourth experiment is similar to Experiment Three except that the input is not character-as-word segmentation. Rather the output of a pure statistical segmenter is used. The following steps are followed when combining the present segmenter with the statistical segmenter:

a. Divide the Penn Chinese Treebank data into the training data (80k) and the testing data (the other 20k)

b. Segment the training data with the statistical segmenter and tag it with the Brill tagger

c. Segment the testing data with the statistical segmenter and tag

it with the Brill tagger.

d. Use the output in (b) to learn a set of rules with the training routine

e. Use the output in (d) to segment the testing data

Results achieved with just the statistical segmenter (output of b):

Precision: 71.59%

Recall: 77.14%

F-score: 74.26%

Results achieved through combining the statistical segmenter with the rule-based segmenter (output of e)

Precision: 89.27%

Recall: 92.24%

F-score: 90.73%

Number of transformations: 4703

The results show that there is an improvement in accuracy. More importantly, the trained model is more compact, cutting the number of transformations by a third.

5.4.3 Discussion

The results of our Experiment Two show that when no part-of-speech information was used, the balanced F-score is 88.33%, slightly higher than

Hockenmaier and Brew's 87.86% when they used trigram models. However, the model for our Experiment Two with 3293 rules is much more compact than that of Hockenmaier and Brew's 7442 rules when they used the trigram model. Our Experiment Three shows when part-of-speech information is used, the results are significantly better, with the F-score being 90.24%.

Comparing the three experiments of our own we found that Experiment Three produced the best results. Either result compares favorably with the 85.06% produced by the maximum matching algorithm. The results also show that Experiment Three handles new words (words not found in the training data) the most effectively. Of the 1745 new words in the testing data, 754 of them were segmented correctly, significantly better than the 55 that the maximum matching algorithm gets right. It is reasonable to assume that the better accuracy overall and the success in dealing with new words is a result of our morpheme-based approach. If this reasoning is correct, it can be considered to be a validation of our theoretical approach.

5.5 Summary of the Chapter

In this chapter, we described a segmenter that implements our theoretical assumptions. First of all, it is morpheme-based and does not use a computational dictionary. Instead it uses rules learned with the transformation-based error-driven algorithm first proposed by Brill (1993). Second, it takes syntactic context

into consideration in identifying words in Chinese. Third, it uses word class (part-of-speech) information. The results show that it provides a significant improvement over a word-based approach that uses the maximum matching algorithm in terms of the overall accuracy measured by the balanced F-score. It compares even more favorably against the dictionary-based approach in dealing with new words, as we expected. Taken together, these results can be viewed a validation of our theoretical approach in understanding Chinese word formation.

Chapter 6

CONCLUSIONS

There are two important aspects of Chinese word formation that need to be accounted for in a theory of Chinese morphology. The first aspect is that the formation of complex words is highly regular and word formation is recursive. This seems to indicate that word formation is syntactic in nature. The second aspect of Chinese word formation is that Chinese words demonstrate lexical integrity effects. Components of words cannot be moved out of the word, can not be deleted, are opaque to external reference and cannot take phrasal modifiers. This state of affairs seems to indicate that words are formed outside of syntax, in a linguistic module of their own. There is thus a dilemma as to where words are formed in Chinese, in syntax or in the lexicon. If there are formed in syntax, while accounting for the first aspect is straightforward, the lexical integrity effects are an apparent challenge. If they are formed in the lexicon, then the first aspect has to be accounted for.

Work in the lexicalist framework either posits different notions of word (Dai 1992) or devise complicated word formation rules in the lexicon to account

for this (Packard 2000). As a representative for the first approach, Dai posits Syntactic Word to account for the first set of facts and Morphological Word to account for the second set of facts. While his approach covers the empirical facts, it is unsatisfying in that it does not attempt to tie these two aspects of word formation together. Packard's approach displays a different problem, which is that his word-formation rules overlaps with the syntactic operations to a large extent.

I have taken a different road from that of the lexicalist approaches and insist that in Chinese complex words are formed in syntax, in the spirit of the Distributed Morphology Hypothesis (Halle and Marantz 1993; 1994, Marantz 1997, Noyer 1997, Embick and Noyer 1999 and others). In Chapter 2, I first examined the wordhood tests that have been proposed in the Chinese linguistics literature and conclude some of the tests follow from the general X-bar theoretic framework and others follow from locality conditions such as the LIH. I then showed how the LIH effects can be derived in a straightforward manner if words are formed in syntax in Chapter 3. In Chapter 4 I examined complex verbs and showed their formation provides further evidence for our theoretical position. I have showed under our theoretical assumptions, the so-called "breakable compounds" (Li and Thompson 1981) can be explained without resorting to redundancy rules in the lexicon. I have also showed the V+V and V+P compounds have to be formed in syntax since the formation of some of these

words feeds on the syntactic structure.

In Chapter 5 I described an automatic word segmenter that implements our theoretical assumptions with the transformation-based error-driven algorithm (Brill 1993). Our working hypothesis is that if our theoretical assumptions are correct, we should see better results over "lexicalist" implementations. The results show that our implementation is a significant improvement over a lexicalist implementation that uses the maximum matching algorithm in terms of overall accuracy and in dealing with new words. We take this to be a validation of our theoretical assumptions.

References

Ackema, Peter. 1995. *Syntax below Zero.* OTS Dissertation Series.

Anderson, S. R. 1992. *A-Morphous Morphology*. Cambridge: Cambridge University Press.

Aronoff, Mark. 1994. *Morphology by Itself*. Cambridge, Massachusetts: The MIT Press.

Baker, Mark C. (1988*) Incorporation: a theory of grammatical function changing*. Chicago:University of Chicago Press.

Bender, Emily. 2000. "The syntax of Mandarin -ba". *JEAL* 9:2, 105-145.

Brill, Eric. 1993. *A Corpus-Based Approach to Language Learning*. U. of Penn. Dissertation.

Chang, Claire Hsun-huei. 1997. "V-V Compounds in Mandarin Chinese: Argument Structure and Semantics". In Jerome L. Packard ed. *New Approaches to Chinese Word Formation*. Berlin, Germany: Mouton de Gruyter.

Chao, Yuen Ren. 1968. *Grammar of Spoken Chinese*. Berkeley and Los Angeles, California: University of California Press.

Cheng, Lisa L.S.. 1986. De in Mandarin. Canadian Journal of Linguistics, 31(4):313-326.

Chomsky, Noam. (1970) "Remarks on nominalization." In R. Jacobs and P. Rosenbaum, eds., *Readings in English Transformation Grammar*. Waltham, MA:Ginn

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht:Foris.

Chomsky, Noam. 1995. *The Minimalist Program.* Cambridge, Massachusetts: MIT Press.

Dai, John Xiang-Ling. 1997. "Syntactic, phonological, and morphological words in Chinese". In Jerome L. Packard ed. *New Approaches to Chinese Word Formation*. Berlin, Germany: Mouton de Gruyter.

Dai, Xiang-Ling. 1992. *Chinese Morphology and its Interface with the Syntax*. Ohio State University Dissertation

Di Sciullo, Anna-Maria and Edwin Williams. 1987. *On the Definition of Word*. Cambridge, Massachusetts: The MIT Press

Duanmu, San. 1997. "Wordhood in Chinese", in Jerome J. Packard ed. *New Approaches to Chinese Word Formation*, New York: Mouton de Gruyter.

Embick and Noyer. 1999. Locality in Post-Syntactic Operations. *MIT Working Papers in Linguistics* 34,265-317.

Feng, Shengli. 1995. *Prosodic structure and prosodically constrained syntax in Chinese*. Ph.D. dissertation, University of Pennsylvania.

Fu, Jingqi. 1994. *On Deriving Chinese Derived Nominals: Evidence for V-to-N Raising.* University of Massachusetts (Amherst) Dissertation.

Fu, Jingqi. 1999. "From Phrase to Word: Syntactic and Semantic Derivation of the "HAVE+noun" pattern in Chinese". *NACCL*-11, Harvard University, Cambridge, Massachusetts.

Gan, Kok Wee. 1993. *Integrating Word Boundary Disambiguation with Sentence Understandin*g. National University of Singapore Dissertation.

Gan, Kok-Wee, Martha Palmer and Kim-Teng Lua. 1996. "A statistically Emergent Approach for language processing: application to modeling context effects in ambiguous Chinese word boundary perception". *Computational-Linguistics*, 22:4, 531-53.

Gu, Ping and Yuhang Mao. 1994. Hanyu zidong fenci de jinlin pipei suanfa jiqi zai QHFY hanying jiqi fanyi xitong zhong de shixian. [The adjacent matching algorithm of Chinese automatic word segmentation and its implementation in the QHFY Chinese-English system]. In *International Conference on Chinese Computing*, Singapore.

Halle, Morris and Alec Marantz. 1993. "Distributed Morphology and the pieces of inflection", in Hale, Kenneth and Samuel Jay Keyser eds. *The View from Building 20. Cambridge, Massachusetts*. The MIT Press

Halle M. and Alec Marantz. 1994. "Some key features of Distributed Morphology." *MIT Working Papers in Linguistics* 21, 275-288

Hockenmaier, Julia and Chris Brew, 1998a. "Error-Driven Learning of Chinese Word Segmentation". In J. Guo, K. T. Lua, and J. Xu. eds, *12th Pacific Conference on Language and Information*, Pages 218-229, Singapore. Chinese and Oriental Languages Processing Society.

Huang, James C. T.. 1982. *Logical Relations in Chinese and the Theory of Grammar*. MIT dissertation.

Huang, James C. T. 1984. "Phrase structure, lexical integrity, and Chinese compounds", *Journal of the Chinese Language Teachers Association* 19.2:53-78

Huang, James C.T. 1988. "Wo Pao De Kuai and Chinese phrase structure". *Language* 64: 274-311.

Huang, James C.T. 1989. "pro-drop in Chinese: a generalized control theory". In Osvaldo Jaeggli-Kenneth Safir (eds.) 185-214.

Huang, Shuanfan. 1997. "Chinese as a headless language in compounding morphology". In Jerome L. Packard ed. *New Approaches to Chinese Word Formation*. Berlin, Germany: Mouton de Gruyter.

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar*. Cambridge, MA: MIT Press.

Jackendoff, R. 1977. *X" Syntax: A study of Phrase Structure*. Cambridge: MIT Press.

Jin, Wangying and Lei Chen 1998. "Identifying Unknown Words in Chinese Corpora*". The First Workshop on Chinese Language Processing*. Philadelphia, University of Pennsylvania.

Kitagava, Chisato and Ross, Claudia. 1982. "Prenominal Modification in Chinese and Japanese". Linguistics Analysis 9:1, 19-53.

Li, Charles N. 1976. "Subject and Topic: A New Typology of Language." In Li (1976:457-489).

Li, Charles N. and Sandra A. Thompson. 1981. *Mandarin Chinese, a Functional Reference Grammar. Berkeley*: University of California Press.

Li, Yafei. 1990. "On V-V compounds in Chinese". *Natural Language and Linguistic Theory* 8:177-207.

Li, Yafei. 1997. "Chinese resultative constructions and the Uniformity of Theta Assignment Hypothesis". In Jerome L. Packard ed. *New Approaches to Chinese Word Formation*. Berlin, Germany:Mouton de Gruyter.

Liang, Nanyuan. 1986. Shumian hanyu zidong fenci xitong-CDWS]. *Journal of Chinese Information Processing*, 1(1):44-52.

Lieber, Rochelle. (1992). *Deconstructing Morphology*. Chicago: The University of Chicago Press.

Lu, Shuxiang. 1979. *Hanyu yufa fenxi wenti* [Problems in the analysis of Chinese grammar]. Beijing: Shangwu Yinshuguan.

Marantz, Alec. 1984. *On the Nature of Grammatical Relations. Cambridge, Massachusetts*: MIT Press.

Marantz, Alec. 1988. Clitics, morphological merger, and the mapping to phonological structure. In *Theoretical Morphology*, ed. M. Hammond and M. Noonan, 253-70. San Diego, California: Academic Press.

Marantz, Alec. 1989. Clitics and phrase structure. In *Alternative conceptions of phrase structure*, ed. M. Baltin and A. Kroch, 99-116. Chicago: University of Chicago Press.

Marantz, Alec. 1997. "No escape from syntax: Don't try morphological analysis in the privacy of your own Lexicon." *Proceedings of the 21st Annual Penn Linguistics Colloquium: Penn Working Papers in Linguistics* 4:2, ed. Alexis Dimitriadis et.al. 201-225.

Ning, Chunyan. 1993. *The Overt syntax of topicalization and relativization in Chinese*. UC Irvine Dissertation.

Noyer, Rolf. 1997. *Features, Positions and Affixes in Autonomous Morphological Structure*, Garland, New York.

Noyer, Rolf. 1999. "Vietnamese 'morphology' and the definition of word." *University of Pennsylvania Working Papers in Linguistics 5:2: Current Work in Linguistics*, ed. y A. Dimitriadis, H. Lee, C. Moisset & A. Williams. University of Pennsylvania, Philadelphia, 65-89.

Palmer, David. 1997. A trainable rule-based algorithm for Word Segmentation. *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*. (ACL "97), Madrid, 1997.

Ross, Claudia. 1997. "Cognate Objects and the Realization of Thematic Structure in Mandarin Chinese". In Packard-Jerome-L.ed. *New Approaches to Chinese Word Formation: Morphology and the Lexicon in Modern and Ancient Chinese.* Berlin, Germany : Mouton de Gruyter

Ross, claudia. 1995. "Temporal and Aspectual Reference in Mandarin Chinese". *Journal-of-Chinese-Linguistics*, 23:1, 87-136.

Ross, Claudia. 1991. "Coverbs and Category Distinctions in Mandarin Chinese". *Journal-of-Chinese-Linguistics* 19:1, 79-115.

Ross, Claudia. 1985. "Compound Nouns in Mandarin". *Journal of the -Chinese Language Teachers Association*, 20:3, 1-22.

Ross, Claudia. 1984. "Adverbial Modification in Mandarin". *Journal of Chinese Linguistics*, 12:2, 207-234.

Ross, Claudia. 1983. "The function of Mandarin de". *Journal of Chinese Linguistics*, 11:2, 214-246.

Sadock, J. M. 1991. Autolexical Syntax. Chicago: University of Chicago Press.

Selkirk, Elisabeth O.. 1982. *The Syntax of Words*. Cambridge, Massachusetts: The MIT Press

Sproat, Richard and Chilin Shih. 1990. A statistical method for finding word boundaries in *Chinese text. Computer Processing of Chinese and Oriental Languages*, 4:336-351.

Sproat, Richard and Chilin Shih. (1996). "A Corpus-Based Analysis of Mandarin Nominal Root Compound". *JEAL* 5:1,49-71

Sproat, Richard; Chilin Shih; William Gale and Nancy Chang. (1996) "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese". *Computational Linguistics*, 22:3, 377-404

Starosta, Stanley, Koenraad Kuiper, Siew-ai Ng and Zhi-qian Wu. 1997. "On defining the Chinese compound word: Headedness in Chinese compounding and Chinese VR compounds". In Jerome L. Packard ed. *New Approaches to Chinese Word Formation*. Berlin, Germany: Mouton de Gruyter.

Stowell, T. 1981. *Origins of Phrase Structure*. Ph.D. dissertation, MIT, Cambridge.

Wu, Andi and Zixin Jiang. 1998. Word Segmentation in Sentence Analysis. In *Proceedings of the 1998 International Conference on Chinese Information Processing*, Nov. 1998, Beijing, pp. 167-180.

Wu, Andi and Zixin Jian. 2000. "Statistically Enhanced New Word Identification in a Rule-Based Chinese System". In *Proceedings of the Second Chinese Language Processing Workshop* (in conjunction with ACL), HKUST, Hong Kong, pp. 46- 51.

Wu, Dekai and Pascale Fung. 1994. Improving Chinese tokenization with linguistic filters on statistical lexical acquisition. In *Proceedings of the Fourth Conference on Applied Natural Language Processing*, pages 180-181, Stuttgart, October.

Xia, Fei, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. 2000. "Developing Guidelines and Ensuring Consistency for Chinese Text Annotation". In *Proceedings of the Second Chinese Language Processing Workshop* (in conjunction with ACL), HKUST, Hong Kong, pp. 46-51.

Xue, Nianwen. 1997. A Promotion Account of the Chinese Relative Construction. University of Delaware manuscript.

Zwicky, Arnold M. 1990. "Syntactic Words and Morphological Words, Simple and Composite". *Yearbook of Morphology* 3, 201-216.