AN ALGORITHMIC STUDY OF
THE MODERN CHINESE LANGUAGE

Nelson Chou

INTRODUCTION

During the long history of China, the written language has gone through many stages of changes in writing form, meanings, and syntactic structure. For example,

1. Change of writing form: The writing form of Chinese has changed from chia-ku-wen (甲骨文 bone and shell inscriptions) to chin-wen (金文 bronze inscriptions) to li-shu (隸書 official style) to k'ai-shu (楷書 regular writing);

2. Change of meanings: In classic Chinese (文言文), the word 女 means both ‘woman’ and ‘you’; but in modern Chinese (白話文) it is used exclusively to mean ‘woman’;

3. Change of syntactic structure: Before the 5th century, in classic Chinese, the adverbial phrase denoting place (e.g. 於 (yu) . . . (at . . .)) had always been placed after the verbal phrase it modifies, e.g. tzu-ching yu kou-tu 自經於溝濬 (commit suicide in a ditch). But, after the 5th century, the expression yu kuo-tu tzu-ching 於溝濬自經 (in a ditch, commit suicide) has been allowed. The change coincides with modern Chinese grammar, e.g. tsai chia li k’an tien-shih 在家裡 看電視 (at home, watch television) where the adverbial phrase tsai chia li 在家裡 (at home) is placed in front of the verbal phrase k’an tien-shih 看電視 (watch television).

For practical purposes, we limit our research to study only the modern Chinese language for this project.

The Chinese language is not only different from the alphabetical language in appearance, but also in its basic linguistic structure. From the written records, the visually distinguishable
units of the Chinese language are 'tzus' (字), or characters (ideographs). However, linguistically, the basic units of the Chinese language are 'tz’us' (詞), or 'words' which are mostly composed of two or more characters in modern Chinese. It is therefore imperative to begin the study by examining the structure of 'words', the relationships among member characters of a word and that among words of a 'word-group' (詞組), so as to identify the structural patterns of 'words' and 'word-groups'. These identified patterns will be the foundation upon which the much needed algorithm of the modern Chinese language can be built. In turn, this algorithm can be programmed for a computer to process the modern Chinese language for information needs such as information retrieval, machine translation, and even, speech understanding system (SUS).

This particular project is designed to study the semantic, syntactic and pragmatic properties of modern Chinese 'words' so as to lay a solid foundation for the establishment of a machine algorithm of the modern Chinese language. The research is outlined below:

1. The Definition:

   A natural start of the research is to define what a modern Chinese 'word' is, especially in comparison with 'character' and 'word-group'.

2. The Verification:

   For the ease of collecting data, and for the convenience of possible application in the future for SUS system, we decided to go through Hsien-tai Han-yu tz’u-tien 現代漢語詞典 (the Dictionary of Modern Chinese Words), which is arranged alphabetically by pronunciation, and record as many words and word-groups that fit our tentative definition as time allows. This produces a list of tentatively confirmed words and word-groups. (This list can be used as a thesaurus for manual processing.)
3. The Observation:

We, then, take this list as a check-list to observe and to record the behavioral patterns, or variation, of the listed words and word-groups in actual writings, with the following in mind:

a. internal grammatical structure of the listed items in relation to contexts;
b. their semantic variations in relation to contexts; and
c. their pragmatic changes in relation to contexts.

The result of the observation and recording is used to amend the preliminary list and to produce a revised thesaurus.

4. Tagging for Computer Processing:

With specially designed tagging process, all words are tagged for future computer processing.

In practical situations, we will want to assign priorities for different tasks. For each individual task, similar procedures are based upon the same rationale. This establishes a set of thesauri which can be integrated into a single thesaurus when necessary.

5. The Machine Algorithm:

With a single comprehensive thesaurus specifying the structural patterns, the semantic variations, the syntactic relations to contexts, and the pragmatic changes in different environments of words and word-groups, we are able to construct a computerized document retrieval system. However, to construct a true information retrieval system where an ability to perform content analysis by machine is needed, we have to have a machine algorithm. We need to identify and incorporate all special "keys", e.g. non-substantive words which include particles, prepositions, and conjunctions, for this algorithm.

6. The Feedback:

The feedback process is an essential part of any research. For
both the thesaurus and the algorithm, experiments should be conducted beforehand to test feasibility in conjunction with actual texts, and periodically thereafter with new texts for amendments in order to ensure applicability. Because of constraints in time and manpower and lack of equipment for testing, our study was partially done up to the fourth step. Even with partial findings, we are confident that we have affirmed our hypotheses and clearly pointed out a direction of research in this particular field. The remaining steps are left for future research investigations.

CHARACTER, WORD, AND WORD-GROUP

A Tentative Definition

In modern Chinese, most words are combinations of two or more characters. The latter are visually distinguishable units of written Chinese. From earlier studies, our tentative definition of ‘word’ is as follows:
1) it is the smallest meaningful unit of the language;
2) it is functionally independent;
3) it has a unique meaning; and
4) for combined words, their structures are relatively stable.
The following elaborations may help us better understand our tentative definition:

I. Word and Character:

1. Words are essential elements of the language and are directly related to human thought, while characters are signals for recording the language and have no direct relation to human thought.
2. The formation of words sometimes changes as time goes by. A word in ancient usage may not be a word in modern usage. The differentiation of word from character should be based upon the style of the writing we are dealing with at the
moment, i.e. classic (文言) or modern (白話). For example,
a. The symbol p’eng 聰 in the sentence of classic Chinese
“Yu p’eng tzu yuan fang lai 有朋自遠方來 (a friend
comes from a far away place)” is a word, meaning
“friend”; but it is only a member character of the
modern word “p’eng-yu 朋友 (friend)”.
b. The symbol “p’ing 平 (equality)” in the classic Chinese
expression “Pu p’ing tse ming 不 平 則 鳴 (Unequal
treatment causes screaming)” is a word, meaning “equal-
ity” but it is a member character of the modern word
“p’ing-teng 平等 (equality)”.
3. The context in which a word or a character appears should
also be differentiated. For example, in the expressions
“mu-t’ou 木頭 (wood)” and “shih-t’ou 石頭 (stone)” the
symbol “t’ou 頭 ” is a member character of a word; while in
“wo-ti t’ou 我底頭 (my head)”, it is a word itself.

II. Word and Word-group:
1. Within different contexts, the same combination of char-
acters could mean different things. For example,
a. The combination “tso-yu 左右 (left and right)” in
“mou-jen ti tso-yu 某人的左右 (Mr. X’s aides)” means
“aides”, or in “tso-yu tsung-shih na-mo hui-shih 左右總是
么回事 (no matter what, it is so)” means “no matter
what”. In these cases, this combination should be
regarded as a “word”. However, in “tso yu feng yuan
左右逢源 (meet the source left and right)”, it means
two different directions, and should be considered as
a word-group, or two words tso 左 and yu 右.
b. The combination “pi-mo 筆墨 (pen and ink)” in “mo-
jen pi-mo pu ts’o 某人筆墨不錯 (Mr. X writes well)”
is a word, meaning “writing”. While in “mai lai pi mo
買來筆墨 (bought pen and ink)”, it becomes two words,
pi 筆 and mo 墨.
2. Changeability, or stability, of member characters may be used to differentiate a word from a word-group. For example, the expressions "ch’ih-tou-fu 吃豆腐 (to flirt)" and "ch’ih-hsien-fan 吃閒飯 (to lead an idle life)" have specific meanings; they are not changeable and obviously should be treated as words. However, expressions such as "ch’ih fan 吃飯 (eat, or eat rice)" and "ho ch’a 喝茶 (drink tea)" are word-groups, since either the verbs or their noun objects can be replaced by other verbs or nouns without changing the essential meaning of the original verb or noun. For example, in the expression "ch’ih t’ang 吃糖 (eat candy)", the object of the verb is different, but the verb retains its meaning "eat". It is also true that if we replace the verb ch’ih with chu 煮 (cook)—the new expression will be "chu fan 煮飯 (cook rice)"—we can see that the noun retains its meaning, "rice".

THE PROCESS

I. Verification:

For the ease of collecting data and for the potential for application in future speech understanding systems, we decided to use a dictionary of Chinese words, which was available to us at the time, as the basic tool for our selection. The deciding factors are 1) it must be arranged by pronunciation alphabetically; 2) it should be composed of modern words only; and 3) it should contain words created in Mainland China after 1949. After careful consideration, we adopted the Dictionary of Modern Chinese (現代漢語詞典). Since the process of selecting words out off all the entries in the Dictionary (which includes a great deal of non-words—according to our tentative definitions) is very time-consuming, we spent a total of 18 months on the process and had to stop at the letter ‘H’ of the Dictionary. We collected a total of
II. Observation:

We know that, similar to other language, a great number of Chinese words have multiple meanings and functions differently in different contexts. Their parts of speech are likely determined by the context. For example, the words “pien-ch’ien 變難 (change)”, “ch’ung-ching 憧憬 (dream, wish)” can be used as verbs in “hsiao-yuan pien-ch’ien t’ai ta liao 校園變難太大了 (The campus has changed too much)” and “Wo ch’ung-ching i ko mei-li-ti wei-lai 我憧憬一個美麗的未來 (I wish (or dream) a beautiful future)”. But they can also be used as nouns in “Hsiao-yuan ti pien-ch’ien (shih) t’ai ta liao 校園的變難（是）太大了 (The change in campus (is) too much)” and “Wo-ti ch’ung-ching li yu i-ko mei-li ti wei luo 我的憧憬裡有一個美麗的未來 (There is a beautiful future in my wish (dream)).” Also, words “ts’un-hsin 存心 (intend)” and “ch’ou-ch’u 勇躇 (hesitate)” can be used as verbs in “T’a ts’un-hsin tso huai shih 他存心做壞事 (He intends to do bad things)” and “T’a ch’ou-ch’u pu ch’ien 他躇躇不前 (He hesitates to go forward)”. But they can also be used as nouns in “t’ai-ti ts’un-hsin (shih) pu liang 他的存心 (is) 不良 (His intention (is) not good)” and “Che ch’ou-ch’u huei liao shih 這躇躇壞了事 (The hesitation spoiled the matter)”. Further examples such as “huai 壞 (bad)” and “ts’ung-yung 從容 (leisurely)” can be used as adjective and adverb respectively in “huai jen 壞人 (bad person (or people))” and “ts’ung-yung chiu-i 從容就義 (go to the altar (as a martyr) leisurely)”. But “huai 壞” can also be used as a verb in “huai liao shih 壞了事 (spoiled the matter)” and as a noun in “shih liao ko huai 使了個壞 (played a dirty trick)”. In the meantime, “ts’ung-yung 從容” can be used as a noun in “T’ai-ti ts’ung-yung 他的從容 (His easy attitude)”. It is clear that the context of words determines their nature. We should first identify the words whose parts of speech are changeable, as well as those which are non-changeable. We should then separately study them.
based on their structural variations in relation to the context. Nevertheless, each word must have a primary role, e.g. noun, verb. In the preliminary stage, we categorize the words by their primary roles. Then, within each category, we arrange the entries in alphaytical order according to their pronunciations.

Based upon this notion, we find that among the collected 13,327 words and word-groups:

1) 8,007, or 60.08% of the total, are nominals (including proper nouns and borrowed words—transliterations of foreign words) which are subdivided as shown in Table I and presented in Graph I:

2) 3,195, or 23.97% of the total, are verbals which are subdivided as shown in Table II and presented in Graph II;

3) 787, or 5.90% of the total, are adjectivals which are subdivided as shown in Table III and presented in Graph III;

4) 317, or 2.38% of the total, are adverbials which are subdivided as shown in Table IV and presented in Graph IV;

5) 1,021, or 7.66% of the total, are miscellaneous words, (i.e. idomatic usage, colloquial expressions, and non-substantive words) which are subdivided as shown in Table V.

Of all the 8,007 nominals, there are 5,849 words (excluding word-groups) in 17 classes. The first three classes which comprise 82.15% of the total nominal words are represented in the following formats:

<table>
<thead>
<tr>
<th>FORMAT</th>
<th>NO. OF WORDS</th>
<th>% OF THE TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. (Noun + Noun)</td>
<td>2,962</td>
<td>50.64</td>
</tr>
<tr>
<td>2. (Adj. + Noun)</td>
<td>1,226</td>
<td>20.96</td>
</tr>
<tr>
<td>3. (Verb + Noun)</td>
<td>638</td>
<td>10.91</td>
</tr>
</tbody>
</table>

Clearly, Formats 2. and 3. are composed of a modifier and a substantive component in the structure

M (modifier) ------- S (substantive component)
However, Format 1. needs further categorization. Accordingly, subdivisions are made for Format 1. as follows:

<table>
<thead>
<tr>
<th>STRUCTURE</th>
<th>NO. OF WORDS</th>
<th>% OF THE GROUP</th>
<th>EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. (M ---- S)</td>
<td>2,155</td>
<td>72.75</td>
<td>茶盤，友愛</td>
</tr>
<tr>
<td>b. (S ---- M)</td>
<td>114</td>
<td>3.85</td>
<td>茶磚，冰棒</td>
</tr>
<tr>
<td>(modifier follows the substantive component)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c. (S ---- S)</td>
<td>170</td>
<td>5.74</td>
<td>本源，朋友</td>
</tr>
<tr>
<td>(Components redundant in meaning)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d. (I ---- I)</td>
<td>523</td>
<td>17.66</td>
<td>柴火，潮汐</td>
</tr>
<tr>
<td>(independent components, different in meaning but somewhat relevant to the newly formed words)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If we combine all the nominal words in the structure M ---- S (i.e. 1.a. +2. +3.), we have a total of 4,019 words, or 68.71% of all nominal words, which can be described by the structure M ---- S. We also find among most of the nominal words, that the relative positions of their component are fixed. The few that allows a change of the relative positions of their components have very distinctive patterns. The orginal meaning of the words which are composed of a modifier and substantive component, i.e. M ---- S, S ---- M, or of substantive components only, i.e. S ---- S will change very slightly after a change of the relative positions of their components, e.g. 茶磚 and 砖茶；兜肚 and 腹兜；本源 and 源本。But the original meaning of the words of the structure I ---- I will change completely after a change of positions of their components, e.g. 柴火 and 火柴。

Of the 3,195 verbals, there are 2,989 words (excluding word-groups) in 20 classes. The first two classes which comprise 76.21% of the total verbal words are represented in the following formats:

<table>
<thead>
<tr>
<th>FORMAT</th>
<th>NO. OF WORDS</th>
<th>% OF TOTAL WORDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. (verb + verb)</td>
<td>1,742</td>
<td>58.28</td>
</tr>
<tr>
<td>2. (verb + noun)</td>
<td>536</td>
<td>17.93</td>
</tr>
</tbody>
</table>
It is easy to see that all the 536 words in Format 2 are of the structure \textit{V} (verb) - \textit{O} (object). In other words, they can be represented also by the structure \textit{M} - \textit{S}. Words of Format 1 can be subdivided into the following structure:

<table>
<thead>
<tr>
<th>STRUCTURE</th>
<th>NO. OF WORDS</th>
<th>% OF THE GROUP</th>
<th>EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. (V --- V)</td>
<td>787</td>
<td>45.18</td>
<td>包裝・查問</td>
</tr>
<tr>
<td>(components are relevant but not exactly redundant in meaning)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. (V --- V)</td>
<td>504</td>
<td>28.93</td>
<td>測量・篡改</td>
</tr>
<tr>
<td>(components are redundant in meaning)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c. (V --- V)</td>
<td>271</td>
<td>15.56</td>
<td>拆殺・拆毁</td>
</tr>
<tr>
<td>(first component leads to the result described by the second)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d. (V --- V)</td>
<td>155</td>
<td>8.90</td>
<td>垂釣・裝運</td>
</tr>
<tr>
<td>(M --- S)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e. (V --- V)</td>
<td>25</td>
<td>1.43</td>
<td>唱和・查究</td>
</tr>
<tr>
<td>(S --- M)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

After a close examination, we find that the words of the structure 1.a., 1.c., and 1.d also can be represented by the structure M --- S since their first component usually serves as the modifier of the second, although they can be separately grouped by their semantic differences. Adding up 1.a., 1.c., 1.d. and 2., we have a total of 1,549 verbal words (53.33\% of all verbal words) that can be represented by M --- S structure.

We also find that the relative positions of the components of verbal words are usually not exchangeable without changing the original meaning of the words, except some of those in 1.b. (verb - verb, components are redundant in meaning), e.g. 抒發 and 發抒, 放逐 and 逐放.

There are 760 adjectival words (excluding the word-groups) in 17 classes. Of these 760 words, or 69.47\% of the total adjectival words, are in Format (adjective + adjective) and of the structure S --- S. With very few exceptions, the relative positions of the components of the adjectival words are fixed. It is not possible to change the order of the components of an adjectival word without changing its original meaning. The only exception
that we have come across so far is the word 道地 in the structure of (noun + noun). The word is composed of two components both of which are read in the fourth tone. This particular phenomenon may suggest that the total quality of an adjectival word may have something to do with the changeability of the relative positions of its components.

We know that within certain contexts, a great deal of nominal words can be used as adjectives. One of the well-known markers, the particle 的 which is usually attached to the end of a word, will change a nominal word into an adjective. The inverse is also true that a great deal of adjectival words can also be used as nouns. However, the markers, e.g. 底, 這, 那, unlike the particle 的, usually appear before the word to be changed.

The 288 adverbial words (excluding word-groups), which constitute the smallest group of all, are classified into 24 classes. There are 151 words, or 52.52% of the total, clustered in the following Formats:

<table>
<thead>
<tr>
<th>FORMAT</th>
<th>NO. OF WORDS</th>
<th>% OF THE TOTAL</th>
<th>EXAMPLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Repetition</td>
<td>80</td>
<td>27.87</td>
<td>紛紛, 剛剛</td>
</tr>
<tr>
<td>2. (adv. + adv.)</td>
<td>28</td>
<td>9.72</td>
<td>剛才, 非但</td>
</tr>
<tr>
<td>3. (verb + noun)</td>
<td>22</td>
<td>7.64</td>
<td>頓時, 到家</td>
</tr>
<tr>
<td>4. (adv. + verb)</td>
<td>21</td>
<td>7.29</td>
<td>方始, 何妨</td>
</tr>
</tbody>
</table>

We notice that about 72% of all adverbial words have fixed positions for their components. The other 28% are all in the Format 1. Since the words of this Format are composed of repetitious components, a change of order of the components will not make any difference. Also, we notice that by adding the particle 地 at the end of a verb or an adjective we can change the word into an adverbial word-group.

The 1,021 miscellaneous words and word-groups are roughly grouped into three broad categories: colloquial, idiomatic usage, and non-substantive words (see Table V).
The colloquials are actually word-groups that have to be used as inseparable language units. They are volatile; new colloquials are created constantly and some of the existing colloquials may die of in frequent usage at time goes by. Since we encounter them daily, to collect them is not a difficult task. But, to ensure the completeness of the collection and the timely weeding of the dead ones is extremely difficult. Grammatically, the colloquials also have parts of speech and can be tagged accordingly. Caution should be taken during the tagging process concerning where in a sentence certain colloquials appear (e.g. the colloquial word-group 不管三七二十一 usually appears at the beginning of a sentence, etc.).

The idioms have etymological roots. Some of them cannot be taken literally by the individual meaning of their components. Like the colloquials, they are all inseparable word-groups with identifiable parts of speech. Unlike the colloquials, idioms tend to live on. New idioms are continually created. But rarely do any of them die of infrequent usage. The tagging process should be done in a manner similar to that of colloquials, i.e. identification of parts of speech, then indication of probable place(s) of their appearance in a sentence.

Non-substantive words and word-groups do not have distinctive, complete meanings by themselves. Most of them are meaningful only when used in conjunction with other substantive words. The subdivisions of the non-substantives are: 1) conjunction, e.g. 而且, 但是; 2) preposition, e.g. 關於, 對於; 3) suffix, e.g. 而已, 的話; 4) auxiliary, e.g. 非得...不; and 5) exclamation, e.g. 哎呀, 哎喲. They all have to be used as an inseparable unit.

All miscellaneous words and word-groups have parts of speech and have defined position(s) in a syntax. For example,

1) the colloquial word-group 不管三七二十一 is an adverbial phrase and usually placed at the beginning of a sentence or in front of the predicate as in 不管三七二十一, 我總是要來的 or 我不管三七二十一總是要來的 (No matter what, I will come).
while 花花公子 is a nominal and can be used as the subject as well as the predicate of a sentence in 花花公子才開跑車 (only playboy drives a sports car.), and 他是一個花花公子 (He is a playboy.).

2) the verbal idiomatic word-group 蓋棺論定 can be placed in front of a sentence such as 蓋棺論定，他是好人 (In final judgment, he is a good man); or in the middle of the sentence such as 我們在蓋棺論定之後，才發覺他是好人 (After having gone through the final judgment, (we) find that he is a good man.). (Note: in the former case, the verbal word-group is changed into a prepositional phrase, while in the latter case, the verbal word-group is changed into a statement of perfect tense when a prepositional phrase 在...之後 is added.)

Furthermore, the verbal idiomatic word-group 畫餅充饑 can be used as the subject of a sentence in 畫餅充饑不是辦法 (Drawing (or to draw) a pie to satisfy the hunger is not a solution.); or used together with the preposition 在 to indicate the progressive tense in 他在畫餅充饑 (He is drawing a pie to satisfy (his) hunger, or he is doing something in vain.).

From these examples, we know that usages of idioms are quite flexible, depending heavily upon contextual environments.

3) Non-substantives can be subdivided into at least 5 groups: a) conjunctions; b) prepositions; exclamations; and d) auxiliaries; and e) suffixes.

Conjunctions, prepositions and exclamations do not have substantive meaning and usually used similarly as their counterparts in English.

Conjunctions connect equal and/or dependent statements, be they words of the same parts of speech, phrases, or sentences. They are syntactically important and can serve as marks for syntactic analysis.

Prepositions, similar to their English counterparts, are semantically dependent upon the other component(s) of the prepositional phrase. For example, the preposition 在 in 他在說話 (He is speaking) denotes a progressive tense; while in 他在台北 (He is in Taipei), it denotes a place. But, in both cases, it coincides
with the English word “in” (as we can translate the former into a
deep structure as “He is in the process of speaking.”) In other
words, in semantical analysis, the preposition should be regarded
as an integral part of the whole phrase. Nevertheless, they
always appear at the beginning of the phrase.

Exclamations such as 哎呀 , 哎喲 express a mental state of
surprise and usually appear at the beginning of a sentence, e.g.
哎呀 , 我的天 哪 !(Oh, My God!), 哎呀 , 這是怎麼回事 ?! (Oh,
what is this?!), or they stand by themselves.

The auxiliary 非得 , where the character 得 is read in the third
tone, indicates a necessity. (Sometimes, it is followed by one of
negative word-groups such as 不可 , 不能 , 不行 , or 不成 to em-
phasize the necessity.) For example, 他非得去上了 or 他非得去
上了不可 (He has to go to class). In the second example, the
necessity of going is exaggerated and the word-group 非得 . . . 不
can be divided into two parts which usually surround a verbal
statement. When the character 得 is read in the second tone, it
becomes a verb meaning “to obtain” and the word 非得 is usually
used together with another auxiliary 到 , e.g. 他非得到獎學金不可
(He is determined to obtain a scholarship.). Here the word-group
非得到 . . . 不可 shows a determination and surrounds nominal
statement.

From these examples, we begin to see the influence of
phonic quality upon the semantic quality. This phenomenon
foves the serious suggestion that, for a complete algorithm of
modern Chinese, we probably should also consider phonic rules.

Suffixes such as 而已 , 的話 should be attached at the end of
a verbal or descriptive statement. For example, 他只是看看而已
(He only wants to look); 如果下雨的話 , 我不上課 (If it rains, I
will not go to class.) (Note: The former is usually used together
with the limitative word 只是 and emphasizes the limitation,
while the latter is usually used together with the subjunctive word
如果 and emphasizes the subjunctive.
III. Tagging

All words of any natural language can be classified by their parts of speech. However, different words of the same parts of speech are not necessarily used the same way. Linguists have studied language phenomena from different points of views, e.g. phonological, grammatical, etc. But from the point of view of machine adaptability, the theory of transformational grammar is probably the most suitable means. Since the early 1950's, through the efforts of many scholars, this theory has yet well developed. Noam Chomsky's method of constituent analysis makes parsing the natural language possible. This method prompts the writer to think that if we reverse Chomsky's process by first establishing a lexicon in which each entry is "tagged adequately" for the purposes of possible reconstruction of the natural language, or possible recognition of a printed text by a computer, we will be able to accomplish what we initially proposed to do. Based upon this notion, the tagging process is designed for this project. Since the collection of words for the project was by no means exhaustive, the tentative classification of words is naturally incomplete. However, the sample is large enough to draw a relatively reliable inference on whether this is a correct approach. Tagging becomes all the more difficult if a complete information retrieval system is designed.

1. According to its categorial feature, we tag the word by its parts of speech, e.g. N(oun), V(erb), etc.

2. Based upon syntactic or semantic feature redundancy rule, we will further specify the word:
   a. Nominal words can be first divided as
      +COM and -COM
      where +COM = common noun
      -COM = uncommon noun, e.g. proper noun
      For common nouns, further divisions are
      +CON and -CON
      where +CON = concrete noun
      -CON = abstract noun
For a concrete noun, further divisions are
+ANI and -ANI
where +ANI = animate noun
-ANI = non-animate noun
For animate nouns, further divisions are
+HUM = human
-HUM = non-human
The last divisions of a noun are
+MAS, OMAS and -MAS
where +MAS = masculine
OMAS = neutral gender
-MAS = female
Similarly, the -COM nouns can be further specified as +ANI,
+HUM and then, +MAS.
This process may be further illustrated by the following diagram:

Based upon this scheme, we assign a noun a minimal necessary number of tags which is enough to describ the word in hand uniquely. For example, for the word 基隆, we assign N -COM -ANI; while, for the word 張飛 tags N -COM +HUM
+MAS and for the word 皇帝 , N + COM + MAS should suffice. This general principle of minimization is applied to all words.
b. Verbal words:

Chinese verbal words present a much more difficult problem. They cannot be as neatly classified as nominals. However, based upon grammatical functions, the coconcurrency relations, and selection relations, the verbals can be tentatively classified as follows:

A. Based upon the concurrency relations and selection relations among words, we classify them 1) Actional verb = +ACT and 2) Stative verb = -ACT

Under +ACT, we have
- Causative verb = CAU 叫來，使去
- Durative verb = DUR 成長
- Locomotion verb = LOM 來，去，走
- Meteorological verb = MET 下雨，刮風，打雷
- Momentary verb = MOM 看見
- Performative verb = PER 判斷，宣佈
- Quotative verb = QUO 說，問，告訴
- Resultative verb = RES (+ACT +INC) 拉開，想出

Under -ACT, we have
- Classfactory verb = CLA 是，像，姓
- Descriptive verb = DES 開著，關著，盛放，茂盛
- Factive verb = +FAC 知道，發現，忘記，後悔
- Monfactive verb = -FAC 以為，認為，希望，相信
- Inchaotive verb = INC 出，開，通，發生
- Location verb = LOC 有，坐，站，躺

B. We then categorize them by their grammatical functions:

1) Transitive verb = TR
2) Intransitive verb = IN
3) Ditransitive verb = DI (which has double objects.)

The tagging sequence can be further illustrated in the following:
c. Adjectival words:
Adjectives can be classified tentatively as
Qualified = QUL and
Quantifier = QUN (Numyr + Measure word)
For QUL, further divisions can be made as
Determinator = DET 這，那
Positive = + POS 美，好
Negative = -POS 醜，惡
The selection restrictions between the nominal words and the adjectives are very strict. An adjective, therefore, should be tagged clearly to show not only its own category but also the category(ies) of the nominal words being modified in order to avoid contradictions in meaning, such as 美麗的大便
(beautiful shit).

d. Adverbial words:
Besides those transformed adverbs (e.g. by adding the character at the end of an adjective), the rest of the adverbs can be tentatively classified as follows:
Sentential = + SEN and Non-sentential = -SEN
Both of them can be further divided into
Level = LEV 最，極，稍
Modal = MOD
Negative = NEG
Place = PLA
Repetition = REP
Scope = SCO
Time = TIM

能，得，必
不，別，莫，勿
內，外，上，下，前，後
又，再，也，還
都，光，只，總
今，昔，先，後

The diagram of the tagging process of adjectives is as follows:

ADV

+SEN

LEV MOD NEG PLA REG SCO TIM -SEN

For words having multiple “interpretive components” (or meanings), each component should be listed and tagged so as to eliminate the lexical ambiguity. For example,

長大：1. V DUR TR (grow)
2. ADJ +POS (long and big)
   mod: N +CON -ANI

Detailed tagging process for miscellaneous words are not included in this study. The project should be regarded as a part of the trial-and-error process in establishing a true machine grammar of the modern Chinese language. Although the results are not yet conclusive, they should serve adequately as an indicator to further research in this field.
Footnotes:

3. But in classic Chinese, most words are single characters.
6. The definition of ‘word’ has not yet firmly established. Up to the present time, a great number of entries in the published dictionaries of words do not fit out tentative definition. Hence, differentiation and screening are needed.
7. There is one additional class found in the second phase of the Project in the format of ADV + N.
8. Three more classes are found in the second phase of the Project.
9. Six more classes are found in the second phase of the Project.
10. Ten more classes are found in the second phase of the Project.
12. ‘Tagged adequately’ means to assign minimal necessary number of tags which adequately describes the word in hand for a computer to generate sentences, or to recognize the meaning of a printed text, with the help of the mechanized lexicon.
13. This design is deeply influenced by T’ang T’ing-ch’ih’s (湯廷池) Kuo Yu Pien Hsing Yu Fa Yen Chiu (國語變形語法研究). (Taipei: Student Book Co., 1979).
14. For words in N + COM + CON, a table can be constructed
with words of measurement to show the coordination between nouns and their modifying measurement words. For example, the noun 馬 should only be described by the word 匹.