Customizable Segmentation of Morphologically Derived Words in Chinese

Andi Wu[*]

Abstract

The output of Chinese word segmentation can vary according to different linguistic definitions of words and different engineering requirements, and no single standard can satisfy all linguists and all computer applications. Most of the disagreements in language processing come from the segmentation of morphologically derived words (MDWs). This paper presents a system that can be conveniently customized to meet various user-defined standards in the segmentation of MDWs. In this system, all MDWs contain word trees where the root nodes correspond to maximal words and leaf nodes to minimal words. Each non-terminal node in the tree is associated with a resolution parameter which determines whether its daughters are to be displayed as a single word or separate words. Different outputs of segmentation can then be obtained from the different cuts of the tree, which are specified by the user through the different value combinations of those resolution parameters. We thus have a single system that can be customized to meet different segmentation specifications.

Keywords: segmentation standards, morphologically derive words, customizable systems, word-internal structures

1. Introduction

A written sentence in Chinese consists of a string of evenly spaced characters with no delimiters between the words[1]. In any word-based Chinese language processing[2], therefore, segmenting each sentence into words is a prerequisite. However, due to some special linguistic properties of Chinese words, there is not a generally accepted standard that can be

---

[*] Microsoft Research
   Address: 21062 NE 81st Street, Redmond, WA, USA
   E-mail: andiwu@microsoft.com
   Phone: (O) 1-425-706-0985   (H) 1-425-868-8075

[1] See Sproat [2000] for a theoretical account of this orthographic convention.

[2] Character-based processing is also possible and has performed well in certain applications.

used to unambiguously determine "wordhood" in every case.[3] While native speakers of Chinese are often able to agree on how to segment a string of characters into words, there are a substantial number of cases where no agreement can be reached [Sproat *et al.* 1996]. Besides, different natural language processing (NLP) applications may have different requirements that call for different definitions of words and different granularities of word segmentation. This presents a challenging problem for the development of annotated Chinese corpora that are expected to be useful for training multiple types of NLP systems. It is also a challenge to any Chinese word segmentation system that claims to be capable of supporting multiple user applications. In what follows, we will discuss this problem mainly from the viewpoint of NLP and propose a solution that we have implemented and evaluated in an existing Chinese NLP system[4].

In Section 2, we will look at the problem areas where disagreements among different standards are most likely to arise. We will identify the alternatives in each case, discuss the computational motivation behind each segmentation option, and suggest possible solutions. This section can be skipped by readers who are already familiar with Chinese morphology and the associated segmentation problems. Section 3 presents a customizable system where most of the solutions suggested in Section 2 are implemented. The implementation will be described in detail and evaluation results will be presented. We also offer a proposal for the development of linguistic resources that can be customized for different purposes. In Section 4, we conclude that, with the preservation of word-internal structures and a set of resolution parameters, we can have a Chinese system or a single annotated corpus that can be conveniently customized to meet different word segmentation requirements.

2. Target Areas for Customization

How to identify words in Chinese has been a long-standing research topic in Chinese linguistics and Chinese language processing. Many different criteria have been proposed and any serious discussion of this issue will take no less than a book such as [Packard 2000]. Among the reasons that make this a hard and intriguing problem are:

·   Chinese orthography has no indication of word boundaries except punctuation marks.

·   The criteria for wordhood can vary depending on whether we are talking about the phonological word, lexical word, morphological word, syntactic word, semantic word, or psychological word [Packard 2000, Di Sciullo and Williams 1987, Dai 1992, Dai 1997, Duanmu 1997, Anderson 1992, Sadock 1991, Selkirk 1982, *etc.*].

---

[3] For a comprehensive review of this problem, see Packard [2000].

[4] This system is developed at Microsoft Research in the general framework of Jensen *et al* [1993] and Heidorn [2000]. Details of the Chinese system can be found in Wu *et al* [2000, 1998].

·   Unlike Japanese, Chinese has very little inflectional morphology that can provide clues to word boundaries.

·   Many bound morphemes in Chinese used to be free morphemes and they are still used as free morphemes occasionally. Therefore the distinction between bound morphemes and words can be fuzzy.

·   The character sequence of many Chinese words can be made discontinuous through morphological processes.

·   Word-internal structures look similar to syntactic structures. As a result, there is often confusion between words and phrases [Dai 1992].

·   Structural information is not always sufficient for identifying a sequence of characters as a word. Frequency of the sequence, mutual information between the component syllables, and the number of syllables in that sequence also play a role (Summarized in [Sproat 2002]).

As a result, native speakers of Chinese often disagree on whether a given character string is a word. As reported in [Sproat *et al*, 1996], the rate of agreement among human judges was only 76%. It is not hard to understand, then, why Chinese linguists have had such a hard time defining words.

However, we do not have to wait for linguists to reach a consensus before we do segmentation in NLP. In computer applications, we are more concerned with "segmentation units" than "words". While words are supposed to be well-defined, unambiguous and static linguistic entities, segmentation units are not. In fact, segmentation units are expected to vary from application to application. In information retrieval, for example, the segmentation units are search terms, whose sizes may vary according to specific needs. A system aimed at precision will require "larger" units while a system aimed at recall will require "smaller" ones. A good Chinese IR system should be flexible with the output of word segmentation so that search terms of different sizes can be generated. In machine translation, the segmentation units are strings that can be mapped onto the words of another language. An MT system should not be committed to a single segmentation, since the granularity of that segmentation may be good for some mappings but not for others. We can do better if a variety of segmentation units are generated so that all possible words are made available as candidates for alignment. In an N-gram language model, the segmentation units are the "grams" and their sizes may need to be adjusted against the perplexity of the model or the sparseness of data. In text-to-speech systems, the segmentation units can be prosodic units and the units that are good for IR may not be good for TTS. In short, a segmentation system can be much more useful if it can provide alternative segmentation units. Alternative units provide linguistic information at different levels and each alternative can serve a specific purpose. We

will see some concrete examples in the remainder of this section. To facilitate the use of terminology, we will use "words" to mean "segmentation units" in the rest of this paper.

Now where does the variability in segmentation units come from? If we compare the outputs of various word segmentation systems, we will find that they actually have far more similarities than differences. This is mainly due to the fact that the word lists used by different segmenters have a lot in common. The actual differences we observe usually involve words that are not typically listed in the dictionary. These words are more dynamic in nature and are usually formed through productive morphological processes. It is those morphologically derived words (MDWs hereafter) that are most controversial and most likely to be treated differently in different standards and different systems. This is the main focus of this paper.

The morphological processes we will be looking at have all been discussed extensively in the literature and a brief summary of them can be found in [Sproat 2002]. We will not attempt to review the literature here. Instead, we will concentrate on cases where differences in segmentation are likely to arise. Here are the main categories of morphological processes we will go through:

- · Reduplication
- · Affixation
- · Directional and resultative compounding
- · Merging and splitting
- · Named entities and factoids

During the discussion, we will make frequent reference to the following four existing segmentation standards:

(1) The segmentation guidelines for the Penn Chinese Treebank [Xia 2000] ("CHTB" hereafter).

(2) The guidelines for the Beijing University Institute of Computational Linguistics Corpus [Yu 1999] ("BU" hereafter). These guidelines closely follow the GB standard [GB/T 13715-92, 1993] but have some additional specifications.

(3) The ROCLING standard developed at Academia Sinica in Taiwan. [Huang *et al*. 1997, ROCLING 1997] ("ROCLING" hereafter).

(4) The standard used in our own system.

Our segmentation system is developed as an integral part of a Chinese parser where initial word segmentation produces a weighted word lattice. The word lattice contains all the dictionary words plus the MDWs formed by morphological rules. Syntactic parsing takes this word lattice as its input and the final segmentation corresponds to the leaves of the best parse

tree [5]. Segmentation ambiguities are resolved in the parsing process and the correct segmentation is the one that enables a successful parse. In cases where parsing fails, we back off to partial parsing and use dynamic programming to assemble a tree that consists of the largest partial trees.

2.1 Reduplication

The main patterns of reduplication in Chinese are AA, ABAB, AABB, AXA, AXAY, XAYA, AAB and ABB. Examples of these patterns can be found in Appendix 1. Existing standards do not have much disagreement over the segmentation of AA, AABB, AXAY, XAYA, AAB and ABB. These are all considered single words for the simple reason that, except in the case of AA, breaking them up will result in segments that are not independent words. The problem cases are ABAB and AXA.

2.1.1 ABAB

A representative example of this is "讨论讨论" (taolun-taolun: discuss-discuss "*have a discussion*"). It is considered a single word in the CHTB and ROCLING standards, but two separate words in the BU standard. According to CHTB and ROCLING, ABAB is just a variation of AA, where the reduplicated word is made of two characters instead of one. Since the meaning of AA (such as "看看" (kan-kan: look-look "*take a look*")) or ABAB is not compositional,[6] they should be both considered single words. According to the BU standard, however, "讨论讨论" should be broken up because "讨论" can be looked up in the dictionary but "讨论讨论" can not.

Different NLP applications can also have different requirements. The one-word segmentation may simplify syntactic analysis but the two-word segmentation might be better for information retrieval or word-based statistical summarization. For pinyin-to-character conversion, adding the reduplicated form to the word list should improve accuracy but may not have the desired effect if the data is too sparse. In machine translation, it will be desirable to have both: the one-word analysis will make it easier for us to learn mappings between, say, "讨论讨论" and "have a discussion", whereas the two-word analysis will let us translate "讨论" into "discuss" in case no mapping is found for "讨论讨论" in the training data. In our system, we treat ABAB as a single word with internal structure, i.e. [讨论　讨论], so that we can have access to both kinds of information. The word also has a "lemma" attribute indicating that the "underlying form" is "讨论".

---

[5] The weights in the word lattice are considered in the selection of the best parse.

[6] The meaning of AA is not "A and A". The verb or adjective is duplicated here to represent certain grammatical aspects, such as short duration or attempted action.

2.1.2 AXA

This covers cases like the following:

试一试	shi-yi-shi: try-one-try	*"give it a try"*
试了试	shi-le-shi: try-LE-try[7]	*"gave it a try"*
试了一试	shi-le-yi-shi: try-LE-one-try	*"gave it a try"*

Both BU and ROCLING regard those expressions as separate words, while CHTB treats them as single words with internal structures. Our system also analyzes them as single words. To represent the fact that AXA is an instance of A with additional aspectual information, we store two additional attributes in this word: a "lemma" attribute that holds the "underlying form" of the MDW (e.g. "试" for "试了试") and an "aspect" attribute whose value(s) record the aspectual information carried by "一" and/or "了".

The lemma attribute is in fact assigned in each type of reduplication. This is especially important for AABB, AAB and ABB. In the case of AABB such as "清清楚楚" (qing-qing-chu-chu "*very clear*"), for instance, we will not get "清楚" (qingchu "clear") unless we segment it into "清 / 清楚 / 楚" which is not acceptable by any standard because of the dangling bound morphemes on the two sides. This problem disappears once we have "清楚" represented as the lemma of the whole reduplicated form.

2.2 Affixation

Affixation is a very productive morphological process in Chinese. Examples of various derivational processes can be found in Appendix II. As we can see, the morphological rules that combine stems with affixes are almost indistinguishable from the syntactic rules that attach a modifier to a head. The only difference is that the modifier (in the case of prefixation) or the head (in the case of suffixation) is supposed to be a bound morpheme. However, the line between free morphemes and bound morphemes is often hard to draw in Chinese.[8] There are some relatively clear cases, such as 非 (fei "*non-*") and 超 (chao "*super-*") as prefixes and 者 (zhe "*-er*") and 学 (xue *"-ology"*) as suffixes, but the distinction is fuzzy in many cases.

---

[7] Function words like 了 have no English translation and therefore will be glossed by the uppercase versions of their pronunciation.

[8] Here are a few borderline cases:

   总工程师   zong-gongchenshi *"chief engineer"*
   副主席   fu-zhuxi   *"vice-chairman"*
   足球场   zuqiu-chang   *"soccer field"*
   警察局   jingcha-ju   *"police station"*
   煤气炉   meiqi-lu   *"gas stove"*

  Are they words or phrases?

Even the agentive suffix 者 can act as a free morpheme in cases like "持枪闯入民宅者" (chi-qiang-chuang-ru-min-zhai-zhe: carry-gun-break-into-civilian-residence-er "*people who broke into houses with guns*") where 者 is the head of a noun phrase modified by a relative clause. To avoid this thorny issue, different segmentation standards resorted to different definitions of affixation. In the CHTB standard, the term "affixation" is not explicitly used. Instead, it describes prefixation as JJ+N where JJ is monosyllabic, and suffixation as N+N where the second N is monosyllabic. The ROCLING standard distinguishes between affixes, "word beginning" (接头词 jietouci) and "word endings" (接尾词 jieweici), but they are functionally equivalent in derivational rules. The BU standard tries to distinguish between affixation and modifier-head phrases by restricting affixation to words that end in a pre-specified list of affixes.

In terms of segmentation, all the standards agree that MDWs derived from affixation should be treated as single words. In actual NLP applications, however, we often wish to have access to both the derived word as a whole as well as its components as separate words. In machine translation, for instance, it might be desirable to have a choice of translating either the whole or the parts: translate the whole if a translation for the whole can be found and back off to the parts otherwise. Take 烘干机 (honggan-ji: dry-machine "*dryer*") as an example. Ideally the whole word should be translated into "dryer". However, if our translation knowledge base has no translation for 烘干机 but does have translations for 烘干 and 机, we should be able to translate it as "drying machine" given that the parts are also available. In information retrieval, we may also want to search for the parts if the query term as a whole is not found. For example, we may want to retrieve texts containing 警察 (jingcha "*police*") when the query term is 警察局 (jingcha-ju:police-bureau, "*police station*") .

In our system, we treat complex words derived from affixation as single words, just as the other standards do, but we also keep their internal structures. For example, the complex word 核物理学家 (he-wuli-xue-jia: nuclear-physics-science-expert "*nuclear- physicist*") is represented as [[[核 物理] 学] 家]. Each derived word contains such as a sub-tree. The sub-tree functions as a single leaf node in syntactic analysis but it can be made visible after parsing to become part of the parse tree if necessary.

2.3 Directional and Resultative Compounding

There are many kinds of compounding in Chinese. In terms of word segmentation, the most problematic ones are directional compounding and resultative compounding. In directional compounding, a verb is followed by a directional complement, such as 上 (shang, "*up*"), 下 (xia "*down*"), 进去 (jinqu "*into*"), 出来 (chulai "*out*"), which indicates the direction of the action expressed by the verb. In resultative compounding, a verb is followed by a resultative complement which is a verb or adjective that indicates what results from the action of the first

verb. In both cases, the verb and the complement can be separated by 得 (de) or 不 (bu) to express the possibility of the verb-complement relationship. Here are some examples:

Directional compounding:

走进   zou-jin: walk-enter   "*walk into*"

走进去   zou-jinqu: walk-enter   "*walk in*"

走得进去   zou-de-jinqu: walk-DE-enter   "*can walk in*"

Resultative compounding:

带走   dai-zou : take-go   "*take away*"

带得走   dai-de-zou: take-DE-go   "*can take away*"

带不走   dai-bu-zou: take-not-go   "*cannot take away*"

看清楚   kan-qingchu: see-clear   "*see clearly*"

看得清楚   kan-de-qingchu: see-DE-clear   "*can see clearly*"

看不清楚   kan-bu-qingchu: see-not-clear   "*cannot see clearly*"

The segmentation of those compounds depends on many factors:

(1) Type of compounding. Directional compounds are more likely to be treated as single words than resultative compounds. Both CHTB and ROCLING follow this principle.

(2) Word length. Those compounds are more likely to be treated as separate units if their total length is more than 2. CHTB provides internal structures when the compound is longer than 2 characters. ROCLING treats "看清" (kan-qing: see-clear "*see clearly*") as one word but "看清楚" (kan-qingcu: see-clear "*see clearly*") as two words.

(3) Frequency. Compounds that are more frequent, either synchronically or diachronically, tend to be treated as one word. Compare 打破 (da-po: hit-break "*hit and make it break*") and 打痛 (da-tong: hit-hurt "*hit and make someone hurt*"). These two compounds have exactly the same internal structure and the same word length, but former is more likely to be regarded as a single word than the latter, simply because 打破 is more frequent. The BU standard assumes that all the frequent compounds are already in its lexicon. Therefore non-lexicalized compounds are to be broken up into independent words.

(4) Mutual information [Sproat and Shih 1990]. Compounds whose components have strong mutual information between them are usually taken as single words. For example, 撕裂 (si-lie: tear-split "*tear open*") is not as frequent as 撕坏 (si-huai: tear-bad "*tear and break*"), but 撕裂 is lexicalized in the BU dictionary while 撕坏 is not.

(5) Some resultative verbs are more independent and therefore more likely to stand on their own. Typical examples are 完 (wan "*finish*") and "给" (gei "*give*") which have some special grammatical functions[9] in addition to being resultative complements.

(6) "V + 得/不 + complement" structures are segmented into separate words in BU and ROCLING but kept as single items with internal structures in CHTB[10]. The main reason for keeping them together is that the verb and the complement can usually form a single word.

NLP applications have considerations that are not always compatible with human judgment. In machine translation, it often makes more sense to break up directional compounds into independent words and keep resultative compounds as single words, contrary to the tendencies we observed above. Directional compounds often correspond to verb-preposition sequences in other languages. The compound "走进", for example, corresponds to "walk into" in English. If "走进" is segmented into two words, we will be able to align "走" with "walk" and "进" with "into". After seeing other instances of Verb+进, such as "跑进" (pao-jin: run-enter "*run into*") and "跳进" (tiao-jin: jump-enter "jump into"), we can come to the generalization that Verb+进 is to be translated as Verb+into in English. If those compounds are reduced to single words, we can still learn the correspondence between "走进" and "walk into", but the generalization is not so easy to reach. Resultative compounds, on the other hand, are much more likely to correspond to single words in languages that are unrelated to Chinese. "打破", for example, will most likely align with "break" in English rather than "hit and break" or "break by hitting".

In the case of "V + 得/不 + complement" structures, it is important to know the relationship between the verb and the complement. We need a representation where 吃得下 (chi-de-xia: eat-DE-down "can eat up"), for instance, can be interpreted as having more or less the same meaning of "能 吃下" (neng-chi-xia: can-eat-down "can eat up"). This is crucial not only for semantic analysis, but for such seemingly simple computer applications as various types of Chinese input methods where a language model is used to select the best sequence of characters. Most existing IME systems are error-prone when the input contains the "V + 得/不 + complement" structure. They are unable to relate the verb and the complement even though the verb-complement bigram is in the language model.

To meet the needs of as many standards and applications as possible, our system treats all directional and resultative compounds as single words while preserving their internal

---

[9] 完 can be viewed as an aspectual marker indicating the completion of an action while 给 may have a role similar to the English "to" in dative constructions.

[10] Except in cases like 吃不了 (chi-bu-liao:eat-not-done "*unable to eat anymore*") where V+complement" is not a legitimate compound.

structures. In cases of "V + 得/不 + complement", we also represent the "lemma" which is equivalent to "V + complement". The result is a word tree, where the root node contains the lemma of the compound.

2.4 Merging and Splitting

Both merging and splitting result in word fragments, which often creates a dilemma as to whether to keep those strings as single units or not. We will look at them one by one.

2.4.1 Merging

This morphological process, also known as "telescopic compounding" [Huang *et al.* 1997], can be considered a sub-case of abbreviation, but unlike other kinds of abbreviation, it has a fixed pattern and a predictable semantic interpretation. It applies to cases where two adjacent and semantically related words have some characters in common. The common characters may be at the beginning or end of the words. Here are some examples.

> Common beginnings (AB+AC => ABC)
>
> 国内+国外 => 国内外   guo-nei-wai: country-inside-outside
>
> "domestic + foreign" => "domestic and foreign"
>
> Common endings (AC+BC => ABC)
>
> 进口+出口 => 进出口   jin-chu-kou: enter-exit-port
>
> "import + export" => "import and export"
>
> Ending = Beginning (AB+BC => ABC)
>
> 上海市+市长 => 上海市长 shanghai-shi-zhang: Shanghai-city-head
>
> "Shanghai City + city mayor" => "mayor of Shanghai"

All existing standards agree that we have a single word in the AB+AC and AC+BC cases[11] and two words in the AB+BC case. The problem in the first two cases is that, unless we store ABC in the dictionary as a whole, we will not be able to assign good semantic interpretations to them. However, not all words of this kind can be stored in the dictionary, since merging is a productive morphological process. To interpret a newly merged word, such as 存贷款 (cun-dai-kuan: deposit-borrow-fund "*deposits and loans*"), which is unlikely to be in the dictionary, we seem to need a level of representation where ABC shows up in its underlying form, i.e. AB AC or AC BC. 存贷款 should then be represented as 存款 贷款, not at the surface segmentation, but as the "lemmas" of 存贷款. This is what we do in our system where every merged word contains a tree where the lemmas are conjoined.

---

[11] Unless the sequence is interrupted by a punctuation mark, as in 国内、外 and 进、出口.

2.4.2 Splitting

Splitting is an active morphological process where a multiple-character word with an internal verb-object structure is split into two non-consecutive parts by the insertion of an aspect marker, a measure word or other functional elements.   Here are some examples:

Insertion of an aspect marker

洗了澡　　　xi-le-zao: wash-LE-bath   "*took a bath*"

Insertion of a measure word

洗个澡　　　xi-ge-zao: wash-one-bath   "*take a bath*"

Insertion of both an aspect marker and a measure word

洗了个澡　　　xi-le-ge-zao: wash-LE-one-bath   "*took a bath*"

Insertion of even more words

洗了个舒舒服服的澡 xi-le-ge-shushufufu-de-zao: wash-LE-one-comfortable-DE-bath

"*took a comfortable bath*"

Most segmentation standards require such expressions to be segmented into multiple words, such as 洗 / 了 / 澡. This can result in segments that are not independent words, as we see in the case of 澡 which is a bound morpheme. One may argue that in such cases the bound morpheme is acting as a free morpheme.   But it would still be desirable to have a representation which indicates that 洗 and 澡 actually form a single word and 洗了澡 has more or less the same meaning as 洗澡+了. In other words, the lemma of 洗了澡 should be 洗澡. Such a representation can be difficult in the case of 洗了个舒舒服服的澡, but even there 洗 and 澡 still form a single unit in some sense.

The lemma representation of a split word is obviously useful in the realm of information retrieval since it makes it possible to establish links between the split and non-split forms of the same verbs. As in the verb-complement case (2.3), it may also be beneficial to Chinese input methods that use an N-gram language model to select the correct character sequences. Most existing systems perform poorly when the input contains split words. While the non-split forms of those words (such as 洗澡) are usually in the N-gram model, the split forms are not. If future systems employ word segmentation where the split form is recognized as a single unit with its lemma represented, we will be able to relate 洗 and 澡 in 洗了澡 as long as we have the bigram "洗澡" in the model.

A special case of splitting is found in expressions like 跳起舞来 (tiao-qi-wu-lai "*start dancing*") where two words (跳舞 and 起来 in this case) cross each other. Here again we need a level of representation to encode the fact that 跳起舞来 actually means 跳舞+起来.

Our system regards a split word as a single unit with a single lemma and a subtree if the intervening characters are no more than 2. Syntactic analysis treats the unit as a single leaf

and has the option of exposing the subtree as part of the parse tree after parsing is done. For cases like 洗了个舒舒服服的澡, we parse them as separate words and, if 澡 is found to be the object of 洗 in the parse, we will concatenate the lemmas of the verb and the object (i.e. 洗+澡), look up 洗澡 in the dictionary, and make it the lemma of the subtree if it exists as a dictionary entry. This can also be done in the case of 洗了澡 but we choose to make it a single unit at the lexical level just to reduce the complexity of syntactic analysis. Once its subtree (which also has the verb-object structure in it) is merged into the main parse, we will have a unified representation for 洗了澡 and 洗了个舒舒服服的澡.

2.5 Named entities and factoids

This is an area with the greatest amount of variation among segmentation standards. This is also an area where linguistic theory has very little to say on the justification of a given standard. The differences are mostly computationally motivated and the main concern here is the granularity of segmentation. Different segmentation standards prefer different levels of granularity, but the differences are fairly systematic and can be easily specified in segmentation guidelines. Listed below are the most common types of named entities and factoids whose segmentation may vary across different standards.

2.5.1 Personal names

A personal name is usually composed of a first name and a last name. The BU standard segments a Chinese name into these two parts and treats a foreign name as a single unit if the first name and last name are connected by "·", as in "诺罗敦·西哈努克 nuoluodun-xihanuke "*Norodom Sihanouk*". Other standards treat both Chinese and foreign names as single words. In our system, a personal name is a single word with an internal structure which indicates not only the family name and the given name but the components of the given name as well.

2.5.2 Place names and organization names

There are many levels of granularity here. For instance, "江苏省盐城地区" (jiangsu-sheng-yancheng-diqu: Jiangsu-province-Yancheng-prefecture, "*Yancheng Prefecture, Jiangsu Province*") can be segmented as "江苏省盐城地区", "江苏省／盐城地区", "江苏省／盐城／地区" or "江苏／省／盐城／地区". Likewise, "世界贸易组织" (shijie-maoyi-zuzhi: world-trade-organization "World Trade Organization") can be segmented as "世界贸易组织" or "世界／贸易／组织". Existing standards usually break those names up as long as it does not result in single-character segments. So place names with single-character place-type suffixes (such as 江苏省) tend to be kept as one word while place names with multiple-character place-type suffixes (such as 盐城地区) will be separate words. The BU standard has additional annotation to represent the internal structure of place names. "世界贸易组织", for

example, is tagged as [世界/n 贸易/n 组织/n]nt.

Each level of granularity has its pros and cons. On the one hand, "世界贸易组织" has a better chance of being aligned with "WTO" in the automatic acquisition of translation knowledge if it is segmented as one word. On the other hand, "江苏省" can be more easily related to "江苏" in information retrieval or automatic summarization if it is segmented into two words. All of this points to the need of a hierarchical structure for all the place names and organization names that contain multiple words. This is what has been done in our system.

2.5.3 Factoids

Word trees are also needed for numbers and other factoids. The reasons are obvious and therefore we will simply list some common cases where internal structures exist and different kinds of segmentation are possible.

- Numbers

  四百五十 si-bai-wu-shi-liu: four-hundred-five-ten-six " *four hundred and fifty-six*"

  四百五十六; 四百 / 五十六; 四百 / 五十 / 六; 四 / 百 / 五 / 十 / 六

  三分之一 san-fen-zhi-yi: three-divide-ZHI-one "*one third*"

  三分之一; 三 / 分之 / 一; 三 / 分 / 之 / 一

  三十多 san-shi-duo: three-ten-more "*thirty or so*"

  三十多; 三十 / 多; 三 / 十 / 多;

  数千 shu-qian:several-thousand "*several thousand*"

  数千; 数 / 千

- Dates

  一九九七年三月五日 yijiujiuqi-nian-san-yue-wu-ri: 1997-year-3-month-5-date "March 5, 1997"

  一九九七年三月五日; 一九九七年 / 三月 / 五日; 一九九七 / 年 / 三 / 月 / 五 / 日;

- Time

  十点零五 shi-dian-ling-wu-fen: ten-clock-zero-five-minute "*five minutes past ten*"

  十点零五分; 十点 / 零 / 五分; 十 / 点 / 零 / 五 / 分;

- Money

  六块九毛三 liu-kuai-jiu-mao-san: six-dollar-nine-dime-three

  "Six dollars and ninety-three cents"

  六块九毛三; 六块 / 九毛 / 三; 六 / 块 / 九 / 毛 / 三

- Scores

三比一 san-bi-yi: three-match-one "*three to one*"

三比一; 三 / 比 / 一

- Range

三至五天 san-zhi-wu-tian: three-to-five-day "*three to five days*"

三至五 / 天; 三 / 至 / 五 / 天; 三 / 至 / 五天

These are just simple cases. The structure can be much more complicated when one kind of named entity is embedded in another. However, no matter how complicated they are, clear guidelines can be set up to make them segmented consistently as long as their internal structures are available.

3. A Customizable System

In this section, we give a detailed description of how our system has been designed to address the problems and requirements discussed in the previous section. We will see how the word-internal structures are built, how the system can be customized to produce different outputs, and what the initial evaluation results are. Suggestions will also be made as to how the design principle here can be applied to the development of annotated corpora.

3.1 Dynamic Words

There are two types of words in our system: static words and dynamic words. Generally speaking, static words are those words that are stored in the dictionary while dynamic words are constructed at run time. All the MDWs belong in the category of dynamic words. These words are not supposed to be stored as headwords in our lexicon. Instead, they are to be built dynamically during sentence analysis through the application of a set of word-formation rules.

There are about 50 word-formation rules in our system, covering all the cases listed in Section 2 and more[12]. They are augmented phrase structure rules that have the form of A(conditions)+B(conditions) => C{actions} and each rule has a unique name that describes the particular morphological process involved. The rules are executed like a small grammar in a morphological parser before sentence-level parsing begins. They interact with each other, with some rules feeding into others, but they do not interact with the grammar rules used in sentence analysis.[13] The derivational history from the rule application then forms a tree that represents the internal structure of a given word. Figure 1 is the word tree for a fictional

---

[12] Some of these rules assemble unknown words that are not discussed in Section 2.

[13] We do have the option to run these rules together with the grammar rules, but that has been found to affect the system negatively both in efficiency and accuracy.

organization name, where the labels of non-terminal nodes represent the rules that are applied in constructing the tree.

赵元任语言学基金会

zhao-yuanren-yuyan-xue-jijin-hui:Zhao-Yuanren-language-science-fund-committee

"*Yuan-Ren Chao Linguistics Foundation*"

```
                            OrgName
                   ┌───────────┴───────────┐
               FullName                 OrgName
              ┌───┴────┐            ┌──────┴──────┐
              │    GivenName    NounSfx        NounSfx
              │    ┌───┴───┐    ┌───┴───┐      ┌───┴───┐
            Char  Char   Char Noun  Suffix  Noun   Suffix

             赵    元     任   语言    学    基金     会
```
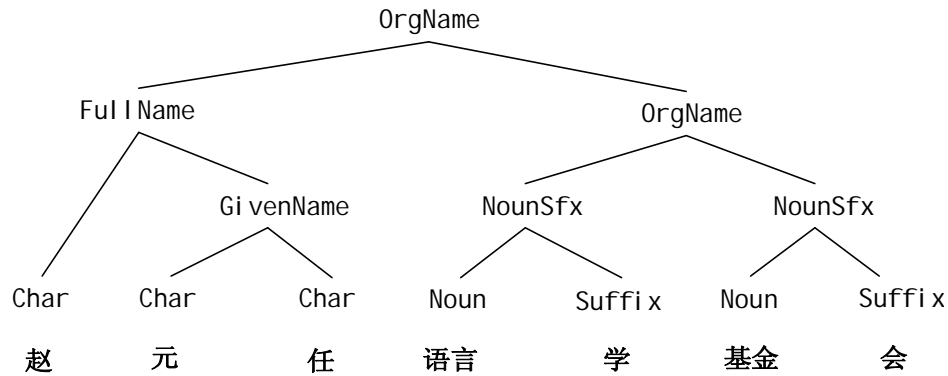
Figure 1

Trees of this kind are built for all types of MDWs, so that all of them can be treated as single words if necessary. These "maximal word trees" or "maximal words" are submitted to the sentence parser as single lexical units, which significantly reduces parsing complexity.

In all cases of merging, splitting and reduplication, the feature structure of the parent node also has an attribute that holds the lemma of the word, as we have already mentioned in Section 2. The value of the lemma is computed by piecing together the relevant characters to hypothesize a word and then checking this word against the dictionary. In the case of AABB reduplication, for instance, the hypothesized word will be AB, such as 清楚 in 清清楚楚. Since 清楚 is a word in the dictionary, it becomes the value of the lemma attribute of 清清楚楚. Similarly, the lemma of 洗了澡 is 洗澡.[14] In the case of AC+BC => ABC merging, both AC and BC will be hypothesized and put into the lemma attribute of ABC if verified in the dictionary. For example, the lemma of 进出口 is 进口+出口. These operations all take place in the "actions" part on the right-hand side of the rule.

An interesting question that arises naturally at this point is what words should be listed

---

[14] In addition to the lemma, we also have attributes that record the information associated with the inserted part. In 洗了澡, we store the tense/aspect information contributed by 了, so that 洗了澡 as a single verb will be equivalent to 洗了澡 as a verb phrase in terms of semantic content.

in the dictionary. According to our design, none of the MDWs should go into the dictionary. This way the word trees we get will have the maximal word at the top node, the minimal words at the leaves, and the intermediate words at the other nodes. We can thus accommodate the widest range of segmentation variations. In practice, however, there are some complications that need to be dealt with.

First of all, none of the existing dictionaries has been built strictly in line with this "minimal word" principle. They do have the minimal words, but they usually also contain words that are supposed to be dynamic in our system. It is not hard to imagine that a dictionary may contain words like 意大利式 (yidali-shi:Italy-style "*Italian-style*", 搬进 (juan-jin:move-enter "*move into*"), and 中小学 (zhong-xiao-xue:middle-small-school "*middle school and elementary school*"). Since our original dictionary was acquired rather than created in house, we do have this problem. We do not ***add*** any MDW to our dictionary, but we have to find a way to deal with those words that are already in the lexicon.

The easy way out is to leave the existing dictionary alone, with the assumption that words like 意大利式, 走进, and 进出口 are lexicalized in the dictionary because they have been lexicalized in a Chinese speaker's mind. We can also assume that they are all high-frequency words or words with strong mutual information between their components. Therefore they should stay unsegmented for probabilistic reasons. Yet another assumption is that the dictionary has listed all the exceptional MDWs that should never be segmented. If any of these assumptions turns out to be true, we should respect the dictionary entries, regarding every word in the dictionary as a minimal word, and build word trees only for words that are not in the dictionary.

These assumptions do not always hold, of course. We do find many dictionary words that can be further segmented. The solution we adopted is to keep those MDWs in the dictionary while assigning internal structures to them at run time. For all the lexicalized words that need internal structures, we mark them with two simple attributes: Type and Segs. [15] The value of Type is the name of the rule that would have been used to construct the word dynamically had this word not been lexicalized. Segs marks the potential internal word boundaries in the word. For 语言学 (yuyan-xue, language-study, "linguistics"), for example, we will have Type = "NounSfx" and Segs = "语言_学". With these two pieces of information, we are able to reconstruct the internal word tree at run time. In terms of structure, therefore, a lexicalized 语言学 will be identical to a dynamically constructed 语言学. This enables us to handle all MDWs in a unified way in later stages of processing, regardless of whether they are

---

[15] The addition of such information to the dictionary was done semi-automatically. We automatically extracted from the dictionary candidates for a given type of MDWs and then had a human evaluator remove the invalid ones.

from the lexicon or from the rules.

3.2 Multi-resolution parameters

Once every MDW is assigned a word tree representing its internal structure, how to segment those words becomes merely a display problem, since different segmentations of the same word can now be obtained by taking different cuts of the word tree. Borrowing a term from the graphical world, we can say that we just have to decide on the degree of "resolution" in displaying the internal structure or the granularity of output.

To control the resolution, we let every non-terminal node in the tree be associated with a multi-resolution parameter. Since every non-terminal node corresponds to a word formation rule with which the node was built, the parameter is in effect associated with a given rule or a particular type of morphological process. In the current system, those parameters are binary-valued: 0 if the daughters of a node are to be displayed as a single word and 1 if they are to be displayed as separate words. To illustrate this, we go back to the MDW in Figure 1: 赵元任语言学基金会. We find four different types of node labels in its word tree – OrgName, NounSfx, FullName and GivenName – which are the names of the rules that are used to construct this MDW. Each of them has a multi-resolution parameter: P(OrgName), P(NounSfx), P(FullName) and P(GivenName). Different settings of those parameters then result in different granularities of segmentation:

·   P(OrgName) = 0:

    赵元任语言学基金会

·   P(OrgName) = 1; P(NounSfx) = 0; P(FullName) = 0:

    赵元任 / 语言学 / 基金会

·   P(OrgName) = 1; P(NounSfx) = 1; P(FullName) = 0:

    赵元任 / 语言 / 学 / 基金 / 会

·   P(OrgName) = 1; P(NounSfx) = 0; P(FullName) = 1; P(GivenName) = 0:

    赵 / 元任 / 语言学 / 基金会

·   P(OrgName) = 1; P(NounSfx) = 0; P(FullName) = 1; P(GivenName) = 1:

    赵 / 元 / 任 / 语言学 / 基金会

·   P(OrgName) = 1; P(NounSfx) = 1; P(FullName) = 1; P(GivenName) = 0:

    赵 / 元任 / 语言 / 学 / 基金 / 会

·   P(OrgName) = 1; P(NounSfx) = 1; P(FullName) = 1; P(GivenName) = 1:

    赵 / 元 / 任 / 语言 / 学 / 基金 / 会

We notice that the values of these parameters are not independent in a given structure.

When the parameter of a node is set to 0, the parameter values of all the nodes dominated by that node must be 0 as well. It is impossible to keep a MDW as a single word while separating some of its sub-words at the same time. The value of a parameter can be 1 only if the parameter of its parent node is set to 1. Therefore, although we have about 50 rules and consequently about 50 parameters, there do not exist 2^{50} different ways of segmenting sentences even theoretically. But we do provide enough options to adapt the segmentation to any reasonable standard. A user of our system can set those parameters according to any specification to produce the desired segmentation without making any modification in the system itself. The system is thus easily customizable.

Our current system also provides a parameter whose value determines whether word length is to be taken into consideration. As we have seen in Sections 2.3, words formed through directional and resultative compounding are sensitive to word length when it comes to segmentation. These MDWs are more likely to be treated as single words if it has fewer than three characters. The additional parameter covers this case. When it is set to 1, all MDWs built through derivational and resultative compounding will be segmented into separate words if it contains more than two characters, regardless of the values of other parameters. Suppose the name of the directional compounding rule is "DirCmpd". When the length parameter is set to 0, 走进 and 走进来 will both be kept as single words if P(DirCmpd) is set to 0. They will be segmented into two words if P(DirCmpd) is set to 1. When the length parameter is set to 1, however, 走进 will be kept as one word but 走进来 will be cut into two words even if P(DirCmpd) is set to 0.

We also added a parameter whose value determines whether the lemma or the surface string of a MDW is to be displayed. When this parameter is set to 1, the lemma will be displayed and 跳起舞来 will be displayed as 跳舞 起来. This is of course more like stemming than word segmentation, but this is a functionality that some applications may require. In fact, this might be one of the steps we have to take to go from the "truthful" level of segmentation to the "graceful" level [Huang *et al*. 1997].

3.3 Evaluation

To find out the degree of customization that can be achieved by the parameterization described above, we evaluated our system against two annotated corpora that were made publicly available for SIGHAN's First International Chinese Word Segmentation Bakeoff: the training data of the Penn Chinese Tree Bank and the Beijing University Institute of Computational Linguistics Corpus. These two annotated corpora follow very different guidelines and it should be interesting to see how well our system can adapt to them. The evaluation metric we used to measure our performance was the scoring tool written by Richard Sproat for the First International Chinese Word Segmentation Bakeoff. This scoring

tool measures word recall, word precision, the F-measure, the OOV rate, and the OOV recall rate, among other things.   Given a reference (the gold standard) and a hypothesis (the segmentation hypothesized by the word segmenter), word recall is the percentage of words in the reference that are also in the hypothesis, and word precision is the percentage of words in the hypothesis that are also in the reference. The F-measure is a simple average of precision and recall. The OOV rate is the percentage of words in the reference that are not found in the dictionary, and the OOV recall rate is the percentage of OOV words that are found in the hypothesis. The OOV scores are of interest in this paper because many of the OOV words are MDWs according to our dictionary and the OOV recall rate tells us how many OOV words are covered by the word-formation rules.   The wordlist used in running the scoring tool consists of all the 89,845 entries in our dictionary.

In the evaluation, we first segmented the text using our default setting where every parameter was set to 0. This gave us the maximal word in each case. We then did a quick resetting of the parameters following the relevant guidelines. Results of both the default segmentation and the adjusted segmentation were evaluated against the CHTB and BU gold standards respectively. The differences between the default setting scores and the scores after parameter value adjustment thus reflect the amount of customization that has been achieved:

When evaluated against the CHTB gold standard, our system received the following scores when the default setting was used:

Word Recall:	83.4 %
Word Precision:	90.1%
F-measure:	86.6%
OOV Rate:	8.4%
OOV Recall Rate:	58.8%

After a quick adjustment, during which 19 parameters were reset from 0 to 1, the scores became:

Word Recall:	96.5%
Word Precision:	96.3%
F-measure:	96.4%
OOV Rate:	8.4%
OOV Recall Rate:	86.5%

When evaluated against the BU gold standard, our system received the following scores when the default setting was used:

    Word Recall: 84.4%

    Word Precision: 90.4%

    F-measure: 87.3%

    OOV Rate: 7.5%

    OOV Recall Rate: 49.2%

After the resetting of 22 parameters from 0 to 1, the scores became

    Word Recall: 96.8%

    Word Precision: 95.9%

    F-measure: 96.3%

    OOV Rate: 7.5%

    OOV Recall Rate: 81.1%

We see that the scores improved dramatically across the board in both the CHTB and BU data after the parameter values were adjusted to the relevant standards. In particular, there is a high correlation between the rise of OOV Recall Rate and the F-measure, which indicates that the improvements indeed came from the area of MDWs.

We also tried the setting where every parameter was set to 1, which resulted in the display of minimal words. Here are the scores:

    CHTB: Word Recall: 86.4 %

              Word Precision: 78.6 %

              F-measure: 82.3 %

              OOV Rate: 8.4 %

              OOV Recall Rate: 12.7 %

BU: Word Recall: 91.8 %

           Word Precision: 86.1%

           F-measure: 88.9 %

           OOV Rate: 7.5 %

           OOV Recall Rate: 21.9 %

This is the result we would get if we depended only on our dictionary and no MDWs rules were applied. The scores dropped sharply in both the CHTB and BU cases. Of particular interest is the drop in the OOV recall rates. If all the OOV words were constructed by MDW rules, the OOV recall rate would be 0 when we display the minimal words, which are all in the dictionary. However, there are other processes in our system that assemble dictionary words into bigger units and these units are invariably displayed as single words. For example, "1978" always appears as a single word in spite of the fact that it is assembled from "1", "9", "7" and "8" at run time. Another example is English words in Chinese texts, such as "IBM" which is not in our dictionary. MDWs thus account for 85.8% of the OOV recall rate in CHTB and 73.1% of the OOV recall rate in BU.

The evaluation results show clearly that (1) the variation among different standards does come largely from the area of MDWs and (2) our system can adapt to different standards successfully by parameterizing the display of MDWs.

3.4 Customizable resources

So far we have focused on the customization of a single segmentation system to produce different outputs. We can also envision an approach where segmenters for different standards are built by training them on texts that have been segmented according to those standards. This leads to the question of whether we can develop language resources that can be customized to serve different purposes. The annotated corpora that are currently being developed in the Chinese NLP community mostly follow a single standard and they are usually not designed for the training of segmenters that do not follow the same standard. However, we cannot afford to build a different tagged corpus for each different standard. It will be highly desirable, therefore, to develop resources that are customizable. The requirement for segmented texts, then, is that it should be capable of being converted to segmentations of varying granularity. To achieve this goal, we have to tag our texts in such a way that (1) the internal structures of words (at least the MDWs) are represented and (2) word boundaries of different types can be selectively kept or removed with ease.

Certain word-internal structures are already preserved in some annotated corpora. In

CHTB, for example, verbs and their directional/resultative complements are grouped into single units with internal word boundaries. 走进去 is thus tagged as "(走 进去)" and 走不进去 as "(走 不 进去)". The bracketing of named entities in the BU corpora is another step in this direction. The ROCLING standard has set even higher goals. It classifies segmentation into three increasingly demanding levels: faithful (信 xin), truthful (达 da) and graceful (雅 ya) [Huang *et al*. 1997].[16] The segmentation units at the faithful level basically correspond to the minimal words in our system. Those at the truthful level are usually MDWs. Segmentation units at the graceful level are not as well defined, but some of them correspond to the maximal words in our system, such as company names. Units at these levels are to be tagged with different SGML tags: faithful-level words tagged as <w0>, truthful-level words tagged as <w1>, and graceful-level words tagged as <w2>. "赵元任语言学基金会" will probably be tagged as the following in this scheme, assuming 赵, 元 and 任 are in the dictionary but 赵元任 and 元任 are not:

    <w2>

    <w1> <w0>赵</w0> <w0>元</w0> <w0>任</w0> </w1>

    <w1> <w0>语言</w0> <w0>学</w0> </w1>

    <w1> <w0>基金</w0> <w0>会</w0> </w1>

    </w2>

This tagging scheme makes the tagged data customizable, since all the potential word boundaries are preserved. But it does not distinguish between different types of MDWs and therefore the choices for customization are more limited. To preserve the type information of MDWs, we will need the following representation:

<OrgName>
  <FullName>
    <Char>赵</ Char >
    <GivenName> < Char >元</ Char > < Char >任</ Char > < /GivenName >
  </FullName >
  <OrgName>
    <NounSfx> <Noun>语言</ Noun > <Suffix>学</ Suffix > </ NounSfx >
    < NounSfx > < Noun >基金</ Noun > < Suffix >会</ Suffix > </ NounSfx >
  </OrgName>
</ OrgName >

This representation is equivalent to the word tree in Figure 1. It is somewhat clumsy,

---

[16] (a) Faithful (信 xin): All segmentation units listed in the reference lexicon should be successfully segmented; (b)Truthful (达 da): In addition to (a), all segmentation units derivable by morphological rules should be successfully segmented; Graceful (雅 ya): Segmentation units are ideal linguistic words for fully automated language understanding.

however, and may not be optimal when it comes to large-scale tagging. A simpler representation might be:

赵<3>元<4>任<1>语言<2>学<1>基金<2>会

where each number corresponds to a label, namely 1 = OrgName, 2 = NounSfx, 3 = Fullname, and 4 = GivenName. Since each label represents the morphological rule that assembles the pieces into a single unit, we replace each word-internal boundary with the relevant number that corresponds to the rule that puts the pieces together. We can then obtain different segmentations by specifying the types of boundaries to be kept or removed. During customization, the boundaries to be kept will be replaced by spaces and the ones to be removed will disappear. In the above example, if we want to treat personal names and words derived from suffixation as single words while keeping components of an organization name apart, we can remove <2>, <3> and <4> and turn the other numbers into spaces. The result will be "赵元任 语言学 基金会". We will get "赵 元任 语言 学 基金 会" if the number to be removed is just 4. It should be noted that, just like the case of parameter setting in our system, not all the number combinations are possible in the replacement/removal. For example, we cannot remove <1> and replace all the other numbers with spaces, since we cannot keep the whole organization name as a single piece if we break up its components. Therefore, there need to be a partial order of those numbers where the removal of a given number implies the removal of some other numbers. The original motivation of this representation was to avoid the need to process the same text *N* times to get *N* different segmentations. We were able to process the corpus just once and use the same output for multiple purposes. It seems that this can be an option in the future development of Chinese language resources.

In principle, all the information represented in the word trees of our system can be represented in a tagged corpus. In practice, however, textual representation of certain information (e.g. the lemma attribute) can be cumbersome and it can be labor-intensive for the annotators. Besides, the tagging is not easy to change once it is done. The main advantage of a customizable system over a customizable corpus is that the former can adapt to new specifications of representation very quickly, with large-scale systematic changes made within a very short time. This is especially so in cases of "bracketing paradoxes" where incompatible representations might have to be generated for different purposes. Of course, the output of an automatic system may be inferior in accuracy to a hand-tagged corpus, but we can maintain a set of surface sentences which are known to have the correct output from the system. Every time the "spec" changes, we can modify the system and process those sentences again to produce the updated output instead of modifying the whole tagged corpus.

3.5 Future refinement

In our current implementation of the multi-resolution parameters, the parameter values are not probabilistic in nature. They are either 0 or 1 and therefore it is not able to make the finer distinctions that we sometimes need when we try to determine wordhood on the basis of statistical information. As we have seen in Section 2, the segmentation of certain MDWs can depend on the frequency of those MDWs and the mutual information between their components. To make our customization more fine-tuned, we need to take such probabilistic information into account. One way to do it is to gather statistical information for every MDW and normalize it into a value between 0 and 1. This value can then be combined with the parameter values that we set by hand to produce a probability that represents the likelihood of a MDW being broken into individual words. We can then set a threshold to determine the "resolution" of the segmentation.

4. Conclusion

The standards for Chinese word segmentation can vary according to different definitions of words and the different requirements of NLP applications. It is therefore important that the segmentation systems we develop or the tagged corpora we construct be capable of being customized to meet different needs. In this paper, we have concentrated on the segmentation of morphologically derived words (MDWs). We have demonstrated that a segmentation system can be customized to produce different outputs for different standards if the word-internal structures of MDWs are preserved in a tree structure and different types of nodes in the tree are associated with different resolution parameters. Different settings of those parameters then result in segmentations of different granularities. Evaluation shows that the effect of customization is significant and MDWs are indeed the main area where customization is most needed. A similar approach can also be used in the development of linguistic resources where a single annotated corpus can be customized to provide training and testing data for different applications.

References

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

Dai, J. X.-L, "Syntactic, morphological and phonological words in Chinese", in Packard (1997), pp. 103-134.

Dai, J. X.-L *Chinese Morphology and its Interface with the Syntax*, Ph.D. thesis, The Ohio State Univeristy, Columbus, OH, 1992.

Di Sciullo, A. M. and E. Williams, *On the Definition of Word*. MIT Press, Cambridge, MA, 1987.

Duanmu, S., "Wordhood in Chinese". In Packard (1997) pp.135-196.

GB/T 13715-92. Contemporary Chinese language word-segmentation for information processing. Technical report, Beijing, 1993.

Heidorn, G. E., "Intelligent writing assistance", in *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*, Dale R., Moisl H., and Somers H. eds., Marcel Dekker, New York, 2000, pp. 181-207.

Huang, C., K. Chen, F. Chen and L. Chang, Segmentation standard for Chinese natural language processing. *International Journal of Computational Linguistics and Chinese Language Processing*, 2(2), 1997, pp. 47-62.

Jensen, K., G. Heidorn and S. Richardson. *Natural Language Processing: the PLNLP Approach"*. Kluwer Academic Publishers, Boston, 1993.

Packard, J. (ed.), *New Approaches to Chinese Word Formation: Morphology, phonology and the lexicon in modern and ancient Chinese. Trends in Linguistics Studies and Monographs 105*. Mouton de Gruyter, Berlin and New York, 1997.

Packard, J., *The Morphology of Chinese: A Linguistic and Cognitive Approach*. Cambridge University Press, Cambridge, 2000.

ROCLING Segmentation Principle for Chinese Language Processing, 1997, http://godel.iis.sinica.edu.tw/ROCLING/juhuashu1.htm

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

Selkirk, E., *The Syntax of Words*. The MIT Press, Cambridge, MA, 1982.

Sproat, R., Corpus-Based Methods in Chinese Morphology. Tutorial at the 19th International Conference on Computational Linguistics, 2002.

Sproat, R., *A Computational Theory of Writing Systems*. Cambridge University Press, Stanford, CA, 2000.

Sproat, R., C. Shih, W. Gale and N. Chang, "A stochastic finite-state word-segmentation algorithm for Chinese". *Computational Linguistics*, 22(3), 1996, pp. 377-404.

Sproat, R. and C. Shih, "A statistical method for finding word boundaries in Chinese text", *Computer Processing of Chinese and Oriental Languages*, Vol. 4, 1990, pp. 336-351.

Wu, A. and Z. Jiang, "Statistically-Enhanced New Word Identification in a Rule-based Chinese System". In *Proceedings of the Second ACL Chinese Processing Workshop*, HKUST, Hong Kong, 2000, pp. 46-51.

Wu A. and Z. Jiang, "Word Segmentation in Sentence Analysis". In *Proceedings of the 1998 International Conference on Chinese Information Processing*, Beijing, China, 1998, pp. 169-180.

Xia, F., The Segmentation Guidelines for the Penn Chinese Treebank (3.0). Technical report, University of Pennsylvania, 2000, http://www.cis.upenn.edu/~chinese/.

Yu, S., Guidelines for the Annotation of Contemporary Chinese Texts: word segmentation and POS-tagging, Institute of Computational Linguistics, Beijing University, Beijing, 1999

Appendix

I. Examples of reduplication

- AA

  看看　kan-kan: look-look　　　"*take a look*"

  红红　hong-hong: red-red　　　"*very red / kind of red*"

  慢慢　man-man: slow-slow　　　"*slowly*"

  年年　nian-nian: year-year　　　"*every year*"

- ABAB

  研究研究 yanjiu-yanjiu: research-research　　　"*do some research*"

  舒服舒服 shufu-shufu: comfortable-comfortable　　"*have a comfortable time*"

- AABB

  方方面面 fang-fang-mian-mian　　　　　　　　"*every aspect*"

  清清楚楚 qing-qing-chu-chu　　　　　　　　"*very clear*"

  痛痛快快 tong-tong-kuai-kuai　　　　　　　　"*thoroughly*"

  年年月月 nian-nian-yue-yue: year-year-month-month "*year after year, month after month*"

- AXA

  试一试　shi-yi-shi: try-one-try　　　　"*give it a try*"

  试了试　shi-le-shi: try-LE-try　　　　"*gave it a try*"

  试了一试 shi-le-yi-shi: try-LE-one-try　　　"*gave it a try*"

- AXAY

  跑来跑去 pao-lai-pao-qu: run-come-run-go　　　　"*run around*"

  送医送药 song-yi-song-yao:send-doctor-send-medicine　"*deliver medical aid*"

  一砖一瓦 yi-zhuan-yi-wa: one-brick-one-tile　　　"*every brick / brick by brick*"

  所言所行 suo-yan-suo-xing: SUO-speak-SUO-do　　　"*every word and deed*"

- XAYA

  东看西看 dong-kan-xi-kan: east-look-west-look　"*look here and there*"

  左挑右挑 zuo-tiao-you-tiao: pick-left-pick-right　"*pick and choose*"

- AA 看

  试试看　shi-shi-kan: try-try-look　　　"*give it a try*"

- AAB

  充充电　chong-chong-dian:　　　"*charge the battery a bit*"

  溜溜光　liu-liu-guang　　　　　"*very smooth*"

- ABB

  亮堂堂　liang-tang-tang　　　　　"*very bright* "

II. Examples of Derivational Affixation

1. Prefixation

- Prefix + Noun => Noun

  微电子　wei-dianzi　　　　　　"*micro-electronics*"

- Prefix + Noun => Adj

  防病毒（软件）　fang-bingdu　　　"*anti-virus*"

- Prefix + Verb => Adj

  可再生 （能源）　ke-zaisheng　　　"*re-usable*"

- Prefix + Number => Number
  第一　　　di-yi　　　　　　　　　　　　　"*first*"

2. Suffixation

- Noun + Suffix => Noun
  科学家　kexue-jia　　　　　　　　"*scientist*"
- Noun + Suffix => Adj
  意大利式　yidali-shi　　　　　　"*Italian-style*"
- Verb + Suffix => Noun
  邮递员　youdi-yuan　　　　　　"*mail-man*"
- Verb + Suffix => Adj
  渐进式　jianjin-shi　　　　　　"*gradual-mode*"
- Adj + Suffix => Noun
  积极性　jiji-xing　　　　　　　"*proactive-ness*"