Word Segmentation Rules for Mandarin Chinese Corpus Transcription

August 3, 2019

ZhiYi Wu

1. **Introduction**

Word segmentation is the process of grouping together syllables from utterances to form meaningful words or phrases. This process is necessary for both human and machines to parse the meanings underneath a consecutive list of speech sounds or written characters.

Unlike English and many other languages, Mandarin Chinese (hereafter “Chinese”) is an isolating language whose words are either one syllable long or are compound words, meaning that every written character, or spoken syllable, has its own meaning. Also, there is no space between words in written text, as speakers segment words according to context and semantic meanings. Moreover, different ways of grouping for some characters generate different meanings and different parts of speech. This results in the difficulty of retrieving the correct parts of speech as well as the corresponding meanings for different words while automatically computing the morphosyntactic structure of transcripts for online corpus.

Because of the variability of ways to segment words in Chinese, the current ways of segmentation in different corpus are different, which causes conflicting morphosyntactic computation and grammatical analysis. Unlike Talkbank, many of the existing ways to segment Chinese words are used for accurate machine translation rather than for research on language acquisition. This article aims to address this problem and unify the ways to segment words in research-directed Chinese corpus.

In the following sections, segmentation rules for compound words, phrases as well as some special cases are explained with examples provided. The last section contains some specific segmentation rules on the sentence level. A glossary can be found after the rules.

1. **How to Segment Compound Words**
2. Compound Nouns:
   1. Derived Nouns are words are formed by adding word-forming affixes to a single character, such as [noun+子(zi)], [noun+头(tou)], [noun+儿(er)], [noun+性(xing4)], and [noun+们(men)]. They should not be segmented. Examples of these nouns include石头(shi2tou: stone), 房子(fang2zi: house), 花儿(hua1er: flower), 弹性(tan2xing4: elasticity), 我们(wo3men: we), etc.
   2. [Verb+Noun] compounds as Nouns are formed by combining a noun character after a verb. They should always be treated as one single word. There are two different conditions:
      * 1. When the **last noun character is a bound word** and the **compound means [the+Noun+that+can+Verb**]: 打手(da3shou3: a thug that is hired to beat someone up/a person that can beat), 打印机(da3yin4ji1: printer/a machine that can print)，消耗品(xiao1hao4pin3: consumable/a thing that can be consumed)
        2. When the compounds include a Verb and a Noun and if treated as two words, they are a verb phrase, meaning [verb+ing+noun], but if treated as one whole word, they are nouns, e.g., 烤肉(kao3rou4: grilled meat)，炒菜(chao3cai4: stir-fried dish)
   3. [ADJective+Noun] compounds should be treated as one word ONLY when they are NOT SEPARABLE, which can be determined by adding 的 in the middle to see if the meaning of the phrase changes. E.g., 红花(hong2hua1: safflower)，黑带(hei1dai4: black belt in Taekwondo)，苦瓜(ku3gua1: bitter melon), 老姑娘(lao3gu1niang: an old woman who has never married)
   4. [Noun+Noun] compounds should be treated as one single word. E.g., 垃圾袋(la1ji1dai4: garbage bag), 手包 (shou3bao1: handbag)
   5. [Nouns of Locality + Noun] should be segmented, e.g.,
      1. 前(qian2: front)\_院(yuan4: yard), 里(li3: inside)\_屋(wu1: room), 左(zuo3: left)\_手(shou3: hand);
      2. **Exceptions** include those phrases that indicate time—前年(qian2nian2: the year before last year), 后天(hou4tian1: the day after tomorrow), etc.
   6. [Proper Nouns] such as names of roads, restaurant should not be segmented, e.g., 东方明珠(Oriental Pearl TV Tower), 迪士尼乐园(Disneyland)
   7. Titles of songs, poems, and games as well as lines of classical Chinese poems should be considered as one complete phrase while the content of the games and songs should be treated as natural output. E.g.:
      1. 两只老虎(liang3zhi1lao3hu3: two tigers), 我是一个粉刷匠(wo3shi4yi1ge4fen3shua1jiang4: I am a wall painter)
      2. Lyrics: 我(wo3:I)\_是(shi4:be)\_一(yi1:one)\_个(ge4:cl)\_粉刷匠(fen3shu1jiang4:painter)
   8. People’a names and nicknames should be treated as one, although the titles should be segmented from the names. Different conditions include:
      1. **Prefixes** **used to show intimacy** should be combined with the names after them: 小 (xiao3: young), 老(lao3: old): e.g., 小张(xiao3zhang1), 老王(lao3wang2).
      2. **Titles** such as老师(Teacher), 博士(bo2shi4: Docter) should be separated: e.g., 张(Zhang: Surname)\_老师(lao3shi1: Teacher).
      3. **Titles that are abbreviated** should not be segmented from the names, e.g., 王导(Wang2dao3: Director Wang)
3. Compound Verbs:
   1. [Verb+Verb] such as 敲打(qiao1da3:hit) cannot be segmented. They do not make sense when the aspect marker得(de) or the negation marker不(bu4) is inserted in between the two/three verbs. More examples: 堆放(dui1fang4: stack), 抖动(dou3dong4: shake), 护送(hu4song4: escort).
   2. RESULTATIVE VERB PHRASES should be segmented. These phrases include a free root word and an adjective, a directional verb, or a resultative verb and can be separated by 得(de)/不(bu4). In these cases, when 得(de)/不(bu4) is inserted between the free root word and the following word, 得(de)/不(bu4) should be segmented as a single word:
      1. Root word + Adjectives (stative). E.g.:
         1. 放(fang4: put)\_大(da4: big)—(zoom in)🡺 放 (fang4: put)\_得(de: be able to)\_大(da4: big)—(be able to zoom in); 放(fang4: put)\_不(bu4: not able to)\_大(da4: big)—(cannot zoom in)
         2. 吃(chi1: eat)\_饱(bao3: full)—(become full);
         3. 坐(zuo4: sit)\_满(man3: full)—(the seats are all taken).

* + 1. Root word + directional/resultative verb. E.g.:
       1. 坐(zuo4: sit))\_下(xia4: down), 看(kan4: see)\_出(chu1: out), 走(zou3: walk)\_进(jin4:in).
       2. 问(wen4:ask)\_住(zhu4:stop), 碰(peng4:encounter)\_见(jian4:see), 看(kan4:see)\_到(dao4: arrive).
    2. **Exceptions**:
       1. When the phrase is [Root word + Directional/Resultative verb] with very high frequency, it should not be segmented, e.g., 进去, 起来, 上去. Even when these phrases follow another root word, these phrases should still not be segmented, e.g., 走\_出去, 看\_起来, 放\_下去. See Table 1 for the complete list of these words.

**Table** **1**: list of verb phrases with high frequency

进去 jin4qu4 go\_in

进来 jin4lai2 come\_in

出去 chu1qu4 go\_out

出来 chu1lai2 come\_out

过去 guo4qu4 go\_over

过来 guo4lai2 come\_over

回去 hui2qu4 go\_back

回来 hui2lai2 come\_back

起来 qi3lai2 get\_up

上去 shang4qu4 go\_up

上来 shang4lai2 come\_up

下去 xia4qu4 go\_down

下来 xia4lai2 come\_down

* + - 1. When the phrase is [Root word + Adjectives (stative)] which does not make much sense when segmented but makes sense when 得(de)/不(bu4) is inserted, it should not be segmented, e.g., 打开(da3kai1: open), 打得开(da3bu4kai1:can be opened) 舍得(she3de2: not grudge), 舍不得(she3bu4de2: grudge).
  1. GESTALT VERBS should not be segmented because the whole phrase would mean more than the two single characters simply added together. E.g., 解决(jie3jue2: solve), 过问(guo4wen4: ask), 研制(yan2zhi4: do research and develop), etc.
  2. FUNCTIONAL WORDS include **word-forming affixes**, such as -化(hua4), and **grammatical words**, such as 了, 着.
     1. Word-forming affixes are added after nouns or adjectives to change their parts of speech into verb. They should not be segmented from the core words, such as 商业化(shang1ye4hua4: commercialized), 美化(mei3hua4:beautify)
     2. Grammatical words follow verbs and signal the change of tense of the sentence/phrase. They should be segmented and treated as a single word, e.g., 听(ting1:listen to)\_着(zhe:aspect marker)\_歌(ge1:song)
  3. SEPARABLE VERBS (Verb+Noun phrases)

1. **How to Segment Phrases**
   1. ADJECTIVE Phrases
      1. [Adjective] should not be segmented when there are more than one character, e.g., 好看(hao3kan4: beautiful), 勤奋(qin2fen4: diligent).
         1. When adjectives, especially one-character adjectives, are put before nouns as simple modifiers, they should be separated from the nouns, e.g., 红(hong2:red)\_旗(qi2:flag), 小(xiao3:small)\_狗(gou3:dog).
         2. However, there are some **Special Cases with 的(de)**:
            1. when [Adjective]s are followed by 的 whose presence does not nominalize them, 的should be combined with the adjectives, e.g. 一(yi1:one)\_条(tiao2:classifier)\_漂亮的(piao4liang4-de: beautiful)\_裙子(qun2zi:dress)
            2. when的(de) is added to nominalize the adjective, 的(de) should be segmented, e.g., 我(wo3:I)\_要(yao4:want)\_这(zhe4:this)\_个(ge4:classifier)\_圆圆(yuan2yuan2:round)\_的(de:nominalizer)
   2. [ADVERB] and [ADVERB+地] should not be segmented. When 地(de) is added after the adverbs, it should not be segmented, e.g., 快速地(kuai4su4de:fastly), 非常地(very).
   3. [Numbers and Quantifiers] should be segmented, e.g., 一(yi1:one)\_张(zhang1:classifier)\_桌子(zhuo1zi: table), 一(yi1: one)\_样(yang4:classifier)\_东西(dong1xi:thing). However, when [一(yi1) + QUANT] phrases are used to indicate a period of time, these phrases should not be segmented, e.g., 一下(yi1xia4: a while), 一会儿(yi1hui4er: a while).
   4. DETERMINERS, including 这(zhe4:this) and 那(na4:that), should be segmented from words that follow them, e.g., 这(zhe4:this)\_个(ge4:classfier)\_人(ren2:person).
      1. **Exceptions**: there are situations when the determiner and the following character does not make sense when they are segmented. Under these circumstances, these words should be combined, e.g., 这里(zhe4li3: here), 那样(na4yang4: like that), 这些(zhe4xie1: these) etc.
2. How to segment some SPECIAL CASES

There are some words that can be segmented in multiple ways, while there are others that must be segmented in irregular ways.

* 1. 只有, 只要, and 还有 should be segmented if they mean [adverb+verb], e.g., 只(zhi3:only)\_有(you3:have), 还(hai2)\_有(you3:have). However, if they act as conjunctions in the sentence, they should not be segmented, e.g., 只要(zhi3yao4: as long as), 只有(zhi3you3: only if)
  2. 没有should be seen as one word when it is followed by a verb, but two words when it is followed by nouns:
     1. e.g., 我(wo3:I)\_没有(mei2you3: not)\_吃(chi1:eat)\_水果(shui3guo3:fruit),
     2. e.g., 我(wo3:I)\_没(mei2:not)\_有(you3:have)\_水果(shui3guo3:fruit).
  3. 照着 should be one word when it means “according to”, while two words when it means “lighting up”:
     1. E.g., 照着(zhao4zhe: according to)\_书(shu1:book)\_念(nian4:read);
     2. E.g., 用(yong4:use)\_火把(huo3ba3:torch)\_照(zhao4:light up)\_着(zhe:aspect marker)\_洞(dong4:cave)
  4. 真的should be one word when it is an adverb meaning “really”, but two words when it is a nominalized adjective meaning “real thing”:
     1. 糖果(tang2guo3:candy)\_真的(zhen1de:really)\_很(hen3:very)\_好吃(hao3chi1:delicious)
     2. 我(wo3:I)\_要(yao4:want)\_这(zhe4:this)\_个(ge4:classifier)\_真(zhen1:real)\_的(de:nominalizer)
  5. 一样 should be one word when it means “the same”, while it should be segmented when 样is a CLASSIFIER.
     1. E.g., 我们(wo3men:we)\_一样(yi1yang4:the same)\_高(gao1:tall);
     2. E.g., 你(ni3: you)\_一(yi1:one)\_样(yang4:classifier)\_一(yi1:one)\_样(yang4: classifier)\_做(zuo4:do).
  6. 的时候(deshi2hou4: when) and 的话(dehua4: if) should be one word.
  7. 好了(hao3le) should be one word when it acts as a communicator and does not have a tangible meaning, but should be separated as 好(hao3:good)\_了(le:aspect marker) when it has specific meanings.
  8. When 来着 is an aspect marker without specific meaning, these two characters should not be segmented, e.g., 我(wo3:I)\_记得(ji4de2:remember)\_你(ni3:you)\_要(yao4:is going to)\_走(zou3: go)\_来着(lai2zhe:aspect marker).

1. How to Segment Sentences
2. Special markers are needed when interactional markers and other communicators are used before or after a complete sentence. Explicit use of prefixed (F2+v to enter ‡) and suffixed (F2+t to enter ,,) interactional markers:
   1. Name calling for attention should be separated by ‡: e.g., 妈妈(ma1ma1:mom)\_‡\_我(wo3:I)\_要(yao4:want)\_吃(chi1:eat)\_饭(fan4:meal)\_.
   2. One-syllable verbs used to call for attention such as 看(kan4:look), 听(ting1:listen). E.g., 看(kan4:look) ‡ 插(cha1:stick)\_上(shang4:on)\_ .
   3. Short answer responding to the former Yes/No questions: e.g., 是(shi4:yes)\_的(de)\_‡\_有(you3:have)\_两(liang3:two)\_个(ge4:classifier)\_杯子(bei1zi:cup)\_.
   4. Communicators (but not filler words):
      1. E.g., 对了(dui4le:btw)\_‡\_你(ni3:you)\_在(zai4:is)\_做(zuo4:do)\_什么(shen2me)\_?
      2. E.g., 来(lai2:come on)\_‡\_看看(kan4kan4: look)\_这(zhe4: this)\_个(ge4:classifier)\_.
   5. Tag questions: e.g., 你(ni3:you)\_吃(chi1:eat)\_饭(fan4:meal)\_了(le:aspect marker)\_,,\_对不对(dui4bu4dui4:right)？
   6. Common catchphrases in the end: e.g., 你(ni3:you)\_要(yao4:want)\_这样(zhe4yang4:like this)\_做(zuo4:do)\_,,\_知道(zhi1dao4:know)\_吗(ma1)?
3. A不A, A不AB, AB不AB should be considered as one whole piece: 对不对(dui4bu4dui4: correct or not), 做不做(zu4bu4zuo4: do or not), 牢不牢固(lao2bu4lao2gu4:is it steady),喜欢不喜欢(xi3huan1bu4xi3huan1: like it or not) (Affirmative-negative questions)
4. When there are no conjunction words or phrases between two grammatically independent sentences but the semantic connection between two sentences is too strong to annoy, they should be treated as one sentence.

Take this sentence as an example: 你(ni3:you)\_不(bu4:not)\_说(shuo1:say). 我(wo3:I)\_怎么(zen3me:how)\_知道(zhi1dao4:know)\_你(ni3:you)\_想(xiang3:want)\_要(yao4:want)\_什么(shen2me:what)? (You don’t say. How would I know what you want?)

Even though there is no conjunction word “if”, every native speaker of Chinese can tell there is a conditional relationship between these two sentences, which means that “if you don’t say anything, how would I know what you want”.

Glossary

* Segmentation -- A mechanism to phonologically group syllables to create meaning.
* Although all languages require the listener to group phonemes, written Chinese is different because it lacks spaces between words.
* Bound word -- A character that cannot be used by itself
* Root word -- A character that cannot be used by itself
* Aspect marker -- (example could show "had run" vs. "is running")
* Negation marker -- (not running vs running)