October 2000

The Segmentation Guidelines for the Penn Chinese Treebank (3.0)

Fei Xia
University of Pennsylvania

University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-00-06.

The Segmentation Guidelines for the Penn Chinese Treebank (3.0)

Abstract

This document describes the segmentation guidelines for the Penn Chinese Treebank Project. The goal of the project is the creation of a 100-thousand-word corpus of Mandarin Chinese text with syntactic bracketing. The Chinese Treebank has been released via the Linguistic Data Consortium (LDC) and is available to the public.

The segmentation guidelines have been revised several times during the two-year period of the project. The previous two versions were completed in December 1998 and March 1999, respectively. This document is the third and final version. We have added an introduction chapter in order to explain some rationale behind certain decisions in the guidelines. We also include the English gloss to the Chinese words in the guidelines.

In this document, we first discuss the notion of word and tests for wordhood that have been proposed in the literature. Then we give the specification for word segmentation. The specification is organized according to the potential Part-of-Speech tag of an expression and the internal structure of the expression. Next, we specify the treatment for some common collocations. Finally, we compare our guidelines with two segmentation standards: the first (Liu et al., 1993) is used in Mainland China and the second (CKIP, 1996) is used in Academia Sinica in Taiwan.

Comments

University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-00-06.

The Segmentation Guidelines for the Penn Chinese Treebank (3.0)

Fei Xia

October 17, 2000

Contents

Chapter 1

Introduction

This document is designed for the Penn Chinese Treebank Project [XPX+00]. The goal of the project is the creation of a 100-thousand word corpus of Mandarin Chinese text with syntactic bracketing. The annotation consists of two stages: the first phrase is word segmentation and part-of-speech (POS) tagging and the second phrase is syntactic bracketing. Each stage includes at least two passes, that is, the data are annotated by one annotator, then the resulting files are checked by another annotator.

The segmentation guidelines, like POS guidelines and bracketing guidelines, have been revised several times during the project. So far, we have released all three versions on our web site: the first draft was completed in December 1998, after the first pass of word segmentation and POS tagging; the second draft in March 1999, after the second pass of word segmentation and POS tagging. This document, which is the third draft, is revised after the second pass of bracketing. The major changes in the third draft, compared with the previous two drafts, are (1) we add an introduction chapter in order to explain some rationale behind the guideline, (2) we add the gloss to the Chinese words in the guidelines,[1] and (3) we also turn the guidelines into a technical report, which is published by the Institute for Research in Cognitive Science (IRCS) of the University of Pennsylvania.

1.1 Notion of *word*

The difficulty in defining the notion of *word* is not unique to Chinese,[2] but the problem is certainly more severe for Chinese for a number of reasons. First, Chinese is not written with word delimiters so segmenting a sentence into "words" is not a natural task even for a native speaker. Second, Chinese has little inflectional morphology to ease word identification. Third, there is little consensus in the community on difficult constructions that could affect word segmentation. For instance, the segmentation of verb resultative compounds depends on the syntactic analysis of the construction. One view on how a verb resultative compound is formed says that a simple sentence with a compound is actually bi-clausal and the compound is formed by movement, therefore, the

[1] We'd like to thank Sylvia Lin for adding the gloss. The gloss for words is enclosed in square brackets ([and]) and the gloss for a phrase is enclosed in angle brackets (⟨ and ⟩).

[2] Even for languages that use delimiters between words, such as English, the distinction between a *word* and a non-word is not always clear-cut. For example, *pro-* (which means "supporting/favoring") normally can not stand alone, therefore, it is like a prefix. However, it can appear in a coordinated structure, such as *pro- and anti-abortion*, and under the assumption that only words and phrases can be coordinated, it is like a word. For more discussions of different notions of words (e.g., morphological object, syntactic atom, phonological word and listeme), please refer to [SW87].

compound should be treated as two words. Another view believes that the compound is formed in the lexicon, and therefore should be one word. The segmentation of the verb resultative compounds depends on which view we adopt for this construction. Fourth, many monosyllabic morphemes that used to be able to stand alone in non-Modern Chinese become bound in Modern Chinese. The influence of non-Modern Chinese makes it difficult to draw the line between bound morphemes and free morphemes, the notions which could otherwise have been very useful for deciding word boundaries.

Our approach is based on both linguistic and engineering consideration. The notion *word* in our Treebank is roughly a *syntactic atom* as defined in [SW87], that is, anything that can be inserted into an X^0 position in syntax. This includes both compounds and simple words.

1.2 Tests of wordhood

What tests can be used to decide whether a string of *hanzi*[Chinese character] is a word or not? Without loss of generalization, we assume the string that we are trying to segment is X-Y, which has two morphemes X and Y. The following tests for establishing word boundaries have been proposed by various authors:

- Bound morpheme: a bound morpheme should be attached to its neighboring morpheme to form a word when possible.
- Productivity: if a rule that combines the expression X-Y does not apply generally (i.e., it is not productive), then X-Y is likely to be a word.
- Frequency of co-occurrence: if the expression X-Y occurs very often, it is likely to be a word.
- Complex internal structure: strings with complex internal structures should be segmented when possible.
- Compositionality: if the meaning of X-Y is not compositional, it is likely to be a word.
- Insertion: if another morpheme can be inserted between X and Y, then X-Y is unlikely to be a word.
- XP-substitution: if a morpheme can not be replaced by a phrase of the same type, then it is likely to be part of a word.
- The number of syllables: several guidelines [LTS93, Chi96] have used syllable numbers on certain cases. For example, in [LTS93], a verb resultative compound is treated as one word if the resultative part is monosyllabic, and it is treated as two words if the resultative part has more than one syllable.

All of these tests are very useful. However, none of them is sufficient by itself for covering the entire range of difficult cases. Either the test is applicable only to limited cases (e.g., the XP-substitution test) or there is no objective way to perform the test as the test refers to vaguely defined properties (e.g., in the productive test, it is not clear where to draw the line between a *productive* rule and a *non-productive* rule). For more discussion on this topic from the linguistics point of view, please refer to [Pac98, SW87].

Since no single test is sufficient, we chose a set of tests for our segmentation guidelines which includes all of the ones mentioned except for the productivity test and the frequency test. Rather

than have the annotators try to memorize the entire set and make each decision from these principles, in the guidelines we spell out what the results of applying the tests would be for all of the relevant phenomena. For example, for the treatment of verb resultative compounds, we select the relevant tests (e.g., the number of syllables and the insertion test), and give several examples of the results of applying these tests to verb resultative compounds. This makes it straightforward, and thus efficient, for the annotators to follow the guidelines.

1.3 Compatibility with other guidelines

We have studied other groups' guidelines, such as the Segmentation Standard in China [LTS93] and the one in Taiwan [Chi96], and tried to accommodate them in our guidelines if possible.

Since the final result of the Treebank is a list of bracketed sentences, our guidelines have some flexibility with regards to the segmentation of certain constructions. For example, the string 走上来[walk up] is treated as two segments in [LTS93], but one segment in [Chi96]. In our Treebank, we will segment it into two parts, and then group them together as a compound — that is, (走[walk]/V 上来[up]/V)/V. We call 走上来 a word with internal structures. Our annotation, in this case, is compatible with both [LTS93] and [Chi96]. The comparisons of these three guidelines can be found in Appendix A.

Note: For the sake of annotation efficiency, the grouping of the words with internal structure is done at bracketing stage, rather than at the segmentation stage. In this document, we show the grouping format, but keep in mind that the format is the one AFTER the bracketing is completed. For example, we consider 走上来[walk up] as one word. It is segmented into "走[walk]/V 上来[up]/V" at the segmentation stage, and it will be grouped into (走[walk]/V 上来[up]/V)/V at the bracketing stage. In the paper, we just say 走上来[walk up] should be annotated as (走[walk]/V 上来[up]/V)/V.

Most disagreements among these three guidelines do not make much difference to parsing or sentence interpretation. For most patterns for which the guidelines give different treatments (e.g., numbers and reduplication strings), simple conversion programs can be written to convert the data from one format to another.

Our goal is: in the final output, the word boundary (the highest-level X^0 in the parse tree) should be as accurate as possible, while the internal structure serves as a bridge for the resource sharing with other systems.

1.4 Treatment for unclear cases

There are two types of unclear cases:

- A construction is easy to identify but there is no consensus on its treatment.

  Ex: A-not-A, V-de construction, V-R, potential form (i.e., V-de-R).

  Our approach: we will choose one analysis, and annotate the data according to that analysis. Make sure that the annotation is easy to convert to the structures for other analyses if necessary.

- Two constructions are difficult to tell apart by existing tests.

  Ex: some N+N are compounds, others are phrases.

  Our approach: for the sake of consistency and efficiency, we don't disambiguate the two constructions unless making the distinction is crucial for various reasons.

1.5 Organization of this guidelines

The guidelines are organized according to the internal structure of the corresponding expressions (e.g., a verb resultative compound is represented as V+V, while a verb-object expression is as V+N), so it is easy for the annotators to search the guidelines for reference. The Part-of-speech tags used in this paper are identical to the ones used in the POS tagging task except that the tags for verbs are merged into V and the ones for nouns are merged into N. For the descriptions of the complete POS tagset, please refer to our *Part-of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0)*. The list of POS tags can be found in Appendix B.

In this guidelines, we list mainly the decision for each case without going into detail elaborating other alternatives and the reasoning behind each decision.

Chapter 2

Specification

In this chapter, we assume that a sentence has been segmented into large chunks, and the next step is to decide whether each chunk should be further divided. The chapter is arranged by the potential POS of the chunk if the chunk is a word. To search through the section, first use the "POS" of the chunk to find the subsection, then use the "word" formation information to find the subsection; or simply use the "word" formation information.

2.1 Common noun: NN

2.1.1 Name of relative

Treat it as one word.

Ex: 三叔[uncle]/NN, 表叔[uncle]/NN, 大姑父[uncle]/NN.

2.1.2 CD+N

If a measure word can be inserted between CD and N without changing the meaning, tag it as CD+N; otherwise, tag it as one word (N).

One word: 三排[the third platoon]/NN, 一方[one side]/NN, 三者[three entities]/NN, 一行[a group traveling together]/NN, 21世纪[the 21th century]/NT.

Two words: 一[one]/CD 学生[student]/NN.

2.1.3 DT+N

Treat it as one word if both DT and N are monosyllabic and either DT or N is bound; otherwise, treat it as two words.

Sometimes, it is difficult to decide whether a morpheme is bound or not because of the influence of non-Modern Chinese. To be consistent, we maintain a list of nouns and a list of determiners. If a morpheme is in one of the lists, we consider it as *bound*:

- mono-syllabic bound nouns: 校[school], 球 (when it means *the earth*).
- mono-syllabic bound determiners: 当[this/that]

We also treat 本人[oneself]/NN as one word and tag it as NN.

One word: 本人[oneself]/NN, 本校[our school]/NN, 全球[whole world]/NN, 当地[the place mentioned]/NN, 当今[present time]/NT, 当代[the contemporary era]/NN.

Two words: 本[one's]/DT 单位[organization]/NN.

2.1.4 PN+N

Treat it as one word if both PN and N are monosyllabic and N is bound; otherwise, treat it as two words.

In this case, the current list of bound nouns is: 校[school].

One word: 我校[my school]/NN.
Two words: 我[my]/PN 单位[organization]/NN.

2.1.5 JJ+N

The pattern is: X+N, where X modifies the N, and X is either a JJ or a prefix.

Note: JJ+N can be a phrase. For example, in one of the files we annotated, 全国性[nationwide]/JJ 网络[network]/NN is extended into "全国性[nationwide]/JJ 观测[observe]/VV 苏梅克－列维/NR 9号[number 9]/NN 彗星[comet]/NN 撞击[hit]/VV 木星[Jupiter]/NN 的/DEC 网络[network]/NN".

Segment X+N according to the type of X:

- X is a prefix: treat X+N as one word.[1]

  A list of prefixes: 啊, 非[non-].

  Ex: 啊爸[father]/NN, 非商业化[non-commercial]/JJ 宗旨[purpose]/NN.

  A list of JJs: 原[former], 前[former]

  Ex: 原[former]/JJ 在[at]/P 华[China]/NR 老挝[Laos]/NR 难民[refugee]/NN;
  前[former]/JJ 民主德国[German Democratic Republic]/NR.

- X is a non-predicate adjective:[2] if both JJ and N are monosyllabic, tag it as one word; otherwise, treat it as JJ+N.

  One word: 女人[woman]/NN.

  Two words: 共同[mutual]/JJ 利益[interest]/NN.

[1] The difference between a JJ and a prefix is that the latter, not the former, is bound. As mentioned before, sometimes, it is difficult to tell whether a morpheme is bound or not, so we keep a list of morphemes that we regard as prefixes. In this case, if the N in X+N can be replaced with an NP, we treat X as a JJ, rather than a prefix.

[2] A word is a non-predicate adjective if it can not appear as a predicate after the subject without the help of 是 ... 的.

- X is an adjective: treat it as one word if X or N is bound or the meaning of X+N is non-compositional. For unclear cases, if both JJ and N are monosyllabic, treat JJ+N as one word (e.g., 鲜花[fresh flower]/NN, 强队[strong team]/NN, 红茶[black tea]/NN, 好评[favorable comment]/NN).

  One word: 小媳妇[daughter-in-law]/NN, 大洲[continent]/NN, 大海[sea]/NN.

  Two words: 厚[thick]/JJ 书[book]/NN.

2.1.6 LC+N

If both LC and N are monosyllabic, treat the string as one word, and tag it as NN or NT according to its meaning.

Ex: 前院[front yard]/NN, 前天[day before yesterday]/NT, 左肩[left shoulder]/NN.

2.1.7 N+LC

Treat N+LC as one word if:[3]

- the N and LC are monosyllabic; and
- in this context, the N is non-referential or bound; and
- in this context, the N can not be modified by Det-M or other modifiers.

Otherwise, treat it as two words.

One word (some of them might be two words in *other* context): 室内[indoor](室内[indoor]/NN 训练[training]/NN), 台下[off stage], 眼前 [at present],境外[foreign](境外[foreign]/NN 集团[group]/NN), 境内外[domestic and international] /NN, 海外[oversea](海外[oversea]/NN 市场[market]/NN), 背后[at the back]/NN, 天下[world]/NN, 国内[domestic]/NN, 午后[afternoon]/NT, 赛前[before the contest]/NT.

Two words: 中午[noon]/NT 以后[afterwards]/LC.

2.1.8 N+N: N1 modifies N2

If it is 1+1 or 2+1 (i.e., N1 has one or two *hanzi* and N2 has one hanzi), treat N1+N2 as one word (i.e., we treat all monosyllabic nouns as potential "接尾词"). If a noun with no more than 2 hanzi is followed by multiple "接尾词"(i.e., each monosyllabic noun attaches to the preceding "chunk"), the whole string is treated as one word (e.g., 物理学家[physicist]/NN).
For other cases, the string is treated as two words.

One word: 北京市[Beijing]/NR, 研究室[research lab]/NN, 发展史[developmental history]/NN, 始祖鸟[proto-bird]/NN, 残疾人[the physically challenged]/NN, 清晰度[visibility]/NN, 紧迫感[sense of urgency]/NN, 大奖赛[tournament]/NN, 太阳系[the solar system]/NN.

Two words: 北京[Beijing]/NR 大学[University]/NN, 玩具[toy]/NN 工厂[factory]/NN, 合作[collaboration]/NN 领域[area]/NN, 史学[history]/NN 研究[research]/NN.

[3] N+LC1+LC2, where LC1 and LC2 denote opposite directions, is treated similarly.

2.1.9 PN+LC

If both PN and LC are monosyllabic, treat PN+LC as one word and tag it as NT or NN.

One word: 此间[here]/NN, 此前[before this]/NN, 其中[among them]/NN, 何时[when]/NT.

Two word: 这[this]/PN 以后[after]/LC.

2.1.10 V+N

In this pattern, we assume V is VV (For VA+N, please refer to the section for JJ+N) If V modifies N, treat V+N as one word and tag it as a noun.

one word: 烤肉[barbecue]/NN, 炒菜[stir-fried dishes]/NN, 证明信[certificate]/NN, 讨论会[symposium]/NN.[4]

2.2 Proper Noun: NR

Currently, if the proper noun is composed of multiple words, we don't group them.

2.2.1 Personal name

Treat it as one word. Don't give the internal structure unless there is a space between two names (in foreign alphabet).

Ex: 张胜利/NR, 卡尔[Karl]·马克斯[Marx]/NR, John/NR Smith/NR.

2.2.2 Personal name with affixes

Treat it as one word.

Ex: 老张/NR, 张老/NR

2.2.3 Personal name + title

Treat it as two words.
Ex: 张/NR 教授[professor]/NN, 张/NR 李/NR 两[two]/CD 位/M 教授[professor]/NN.

2.2.4 Name of Organization/Country/School/..

If the pattern is N1+N2, where N2 is a common noun, then if N2 is monosyllabic, treat N1+N2 as one word, else treat N1+N2 as two words.

Simple names: 北京市[Beijing]/NR, 黄河[the Yellow River]/NR, 沙市[Sha City]/NR,黑龙江省[Heilongjiang Province]/NR.

[4]In either of the last two examples, the first morpheme is bisyllabic, and it could be tagged as nouns in some context. Because the second morpheme is mono-syllabic, the expression should be treated as one word regardless of the POS tag of the first morpheme.

Complex names: 北京[Beijing]/NR 大学[University]/NN, 北京[Beijing]/NR 第一[First]/OD 服装厂[Clothing Factory]/NN, 美国[the United States]/NR 国会[Congress]/NN.

2.2.5 NR+NR: coordination without conjunction

Treat it as two words.

Ex: 中[China]/NR 美[the United States]/NR, 中[China]/NR 美[the United States]/NR 关系[relation]/NN, 东[Eastern Asia]/NR 新[Singapore]/NR 澳[Macao]/NR.

2.3 Temporal noun: NT

The names of years/months/day/hour and so on are words.

Ex: 1998年[1998]/NT 3月[March]/NT 21日[21st]/NT, 5点钟[5 o'clock]/NT, 初一[the first day of a lunar month]/NT, 去年[last year]/NT.

2.3.1 CD+N

If CD+N is the name of a time, treat it as one word (NT). If it is the count of the time, treat it as two words (CD+M).

One word: 1998年[1998]/NT, 5点钟[5 o'clock]/NT, 90年代[the 90s]/NT,
Two words: 3/CD 年[year]/M, 3/CD 个/M 月[month]/NN.

2.4 Localizer: LC

Localizers are separated from the noun that it attaches to except for the case mentioned in Section 2.1.7 (i.e., N+LC).

A localizer is either one or two syllables:

- mono-syllabic localizers: e.g. 内[in], 后[after].
- bisyllabic localizers: e.g. 之间[between], 以来[since], 以后[afterwards], 左右[around].

2.5 Pronoun: PN

Treat it as one word.

Ex: 他们[they]/PN, 他自己[himself]/PN, 自己[self]/PN.

2.6 Determiner: DT

We separate DTs from the succeeding words.

Ex: 这[this]/DT 三[three]/CD 个/M 人[people]/NN, 各[each]/DT 国[nation]/NN.

Currently, we treat 这些[these] as one word, and tag it as DT.

Some examples of bisyllabic DTs: 全体[all], 其余[the rest], 一切[all], 这些[these], 那些[those], 所有[all].

2.7 Cardinal number: CD

Treat it as one word. Note: the internal structure of a CD is very easy to recover if needed.

Some examples:

- Pure numbers: 一亿三千万[one hundred and thirty million]/CD, 30.1/CD, 123,456/CD, 35.6%/CD, 30万[three hundred thousand]/CD, 30几[thirty odd]/CD.

- Estimation: 三四十[between thirty and forty-nine]/CD 岁[years old]/M.

- CD + X + CD(5.5.4): X is a morpheme such as 余[odd], 分之[fraction], 点[point]. 三十几亿[three billion odd]/CD, 三分之一[one third]/CD, 三点一[three point one]/CD, 好几[many]/CD 个/M.

- CD+X: X is a morpheme such as 余[odd], 来[over/odd]: 四千一百余[four thousand and one hundred odd]/CD 人[people]/NN, 三十来[about thirty]/CD 个/M.

2.8 Ordinal number: OD

Treat it as one word.

Ex: 第一[first]/OD, 第三十一[thirty-first]/OD.

2.9 Measure word: M

Treat the measure word, including a reduplicated or a compound measure word, as one word. Treat the string such as 分钟[minute] as one word.

Ex: 杯[cup]/M, 杯杯[cup-cup]/M, 架次[number of flights]/M, 分钟[minute]/M.

2.10 Verb: VA, VC, VE, and VV

2.10.1 Reduplication: AA, ABAB, AABB, AAB, ABB, ABAC

Treat it as one word.

- AA, A is a verb: AA/V
  Ex: 看看[see]/VV, 红红[vivid red]/VA.

- ABAB: AB is a verb: ABAB/V
  Ex: 研究研究[research]/VV, 雪白雪白[snow white]/VA.

- AABB, AB is a verb: AABB/V
  Ex: 来来往往[come and go]/VV, 高高兴兴[happy]/VA
  Note: most of the time, AA or BB is not a word.

- AAB (except for AA-看 in 2.10.2): AAB/V
  Ex: 蒙蒙亮[dim]/VA.
  Note: most of the time, AA or B is not a word.

- ABB: ABB/V
  Ex: 绿油油[bright green]/VA, 红彤彤[bright red]/VA.
  Note: most of the time, A or BB is not a word.

- ABAC, etc.: ABAC/V
  Ex: 马里马虎[careless]/VA, 有条有理[orderly]/VA, 一清二楚 [very clear]/VA.

2.10.2 "Reduplication": AA-kan, A-one-A, A-le-one-A, A-le-A

Treat it as one word with internal structure.

- AA-看: (AA/V 看/V)/V:[5]
  Ex: (说说[say]/VV 看/VV)/V.

- A-one-A: (A/V one/CD A/V)/V
  Ex: (想[think]/VV 一[one]/CD 想[think]/VV)/V.

- A-le-A: (A/V le/AS A/V)/V
  Ex: (想[think]/VV 了/AS 想[think]/VV)/V.

[5]The basic meaning of the word 看 is to "see", but in this context, it roughly means "try to do something".

- A-le-one-A: (A/V le/AS one/CD A/V)/V
  Ex: (想[think]/VV 了/AS 一[one]/CD 想[think]/VV)/V.

Note: V+CD+M is treated as three words, e.g. 看[look]/V 一[one]/CD 眼[eye]/M ⟨take a look⟩.

2.10.3 A-not-A

Treat it as one word with internal structure.

Ex: (来[come]/VV 没[not]/AD 来[come]/VV)/V, (高[happy]/VA 不[not]/AD 高兴[happy]/VA)/V, (喜[like]/VV 不[not]/AD 喜欢[like]/VV)/V.

2.10.4 AD+V

If one or more of the following hold, treat AD+V as one word (V):

- no free word can intervene between AD and V,
- the V cannot be a predicate without the AD,
- the subcategorization frame of AD+V is different from that of the V.

Otherwise, treat it as two words.

One word: 胡说[talk nonsense], 胡来[mess things up], 敬献[present with great respect], 尚余[remain] (尚余[still remain]/VV 七十五[75]/CD 名/M 难民[refugee]/NN), 历任[have served successively as], 并列[tied], 不畏[not afraid of].

Two words: 已经[already]/AD 采取[take]/VV, 不[not]/AD 应该[should]/VV, 没[not]/AD 完成[complete]/VV.

2.10.5 MSP+V

If the V can not be a predicate without the MSP, treat MSP+V as one word (V).

One word: 以期[in order to]/VV (以期[in order to]/VV 在[at] 与[with] 美国[the United States] 、瑞典[Sweden] 、挪威[Norway] 这些[these] 世界[world] 强队[strong teams] 交锋[competition] 中[during] ...).

2.10.6 N+V

Some subject-predicate strings can be either a phrase or a word depending on the context.

If a VP-modifier can be inserted between the subject and the predicate part and the "subject" is referential, then the string is a phrase, otherwise it is a word.

One word: 头疼[headache]/VA in "他[he]/PN 让[make]/VV 我[me]/PN 很[very]/AD 头疼[headache]/VA ⟨He gives me a headache⟩".

Two words: 头[head]/NN 疼[ache]/VA in "我[I]/PN 头[head]/NN {很[very]/AD} 疼[ache]/VA ⟨I have a headache⟩".

2.10.7 V+N

If the V and the N are separated (by the aspect markers, by the modifiers of the N, or because the V is reduplicated), treat V+N as two words.

If the V and the N are adjacent,[6]

- If V-N is semantically transitive and its object can occur after N only when VN are adjacent (therefore the V is not a ditransitive verb), treat V+N as one word (e.g., 投资[invest]/VV, 出席[be present]/VV, 关心[care]/VV, 为期[scheduled for a specific duration of time]/VV).
- If V and VN have similar meaning and both are semantically intransitive, treat VN as one word (e.g., 睡觉[sleep]/VV).
- If N is "bound", treat VN as one word (e.g., 游泳[swim]/VV,无望[hopeless]/VV, 无效[invalid]/VV, 无法[unable to]/VV, 辞职[resign]/VV).
- If V-N is 1+1 AND the meaning is noncompositional, treat V-N as one word (e.g., 念书[study]/VV, 流血[bleed]/VV).

Examples of V-N as two words: 访[visit]/VV 华[China]/NR in the sentence 他[he]/PN 曾[previously]/AD 七[seven]/CD 次[time]/M 访[visit]/VV 华[China]/NR ⟨He has visited China seven times⟩.

2.10.8 V+R

The tests for verb resultative compounds (V-Rs): both V and R are verbs and the potential forms (V-de-R, V-not-R) exist. So our definition of V-R includes resultative and directional verb compounds (e.g., 看见[see] and 走上来[walk up]), but it does NOT include words such as 改善[improve] and 鼓动[agitate].

We treat it as one word. For the sake of compatibility with other guidelines, we give the internal structure for the words if they have more than 2 syllables or if the R is the following: 完[finish]/VV.

Words without internal structure: 吃掉[eat up]/VV, 看见[see]/VV, 擦净[wipe clean]/VV.

Words with internal structures: (做[do]/VV 完[finish]/VV)/V, (擦[wipe]/VV 干净[clean]/VV)/V, (认识[realize]/VV 到[reach]/VV)/V.

2.10.9 Potential form: V-de/bu-R

We treat it as one word.

If V-R exists, give the internal structure of V-de/bu-R, otherwise, don't give one.

Ex: words with internal structure: (擦[wipe]/VV 不[not]/AD 净[clean]/VA)/V, (擦[wipe]/VV 得/DER 净[clean]/VA)/V.

[6]The V+N combination is among the hardest cases for the word definition. The tests proposed here are not perfect. They tend to treat idiomatic phrases (similar to "kick the bucket" in English) as words. However, Those errors can be easily corrected if later a dictionary becomes available.

words without internal structure: 吃不了[unable to eat anymore]/VV, 买不起[cannot afford]/VV.

Note: the string "V *de* R" can be ambiguous between potential form and V-de construction. For example, "这[this] 张[M] 桌子[table] 擦[wipe] 得[DER] 干净[clean] 吗[SP]?" can either be a potential form (which means *Can this table be wiped clean?*), or it could be a V-de construction (which means *Has the table been wiped clean?*). The two constructions have different syntactic structures. Normally, we can tell them apart by meaning, by the position of the object or by checking whether adverbs can be inserted between the *de* and the R.

2.10.10 V+DIR

See Section 2.10.8 (i.e., the section for V+R).

Words with internal structure: (走[walk]/VV 出去[out]/VV)/V, (走[walk]/VV 不[not]/AD 出去[out]/VV)/V.

Words without internal structure: 走出[walk out of]/VV, 想出[think of]/VV.

2.10.11 V+AS

Treat it as two words.[7]

Ex: 走[walk]/VV 了/AS.

2.10.12 V+DER

The pattern is V-de in V-de construction. We treat V-de as two words.[8]

Ex: 走[walk]/VV 得/DER (走[walk]/VV 得/DER 很[very]/AD 快[fast]/VA).

2.10.13 Verb coordination without conjunctive words

If the pattern is 1+1, treat it as a word; otherwise, treat it as multiple words.

One word: 修建[build]/VV.

Two words: 宣传[propagate]/VV 鼓动[agitate]/VV.

2.10.14 V+coverb

The pattern is V+X, where X is monosyllabic and it is either a P or a V.[9]

We first decide whether V+X is a word. If it is, we use its syllable count to decide whether to show its internal structure. That is, if V is monosyllabic, don't give the internal structure;

[7] It has been argued that aspect markers are affixes (e.g., [ID92]). Right now, we do not group the V and the AS together.

[8] The function of *de* in the V-de construction is controversial. It ranges from an affix, a particle, to a verb. We will not get into details here.

[9] Many of Xs in this pattern are "coverbs" and it is highly debated which tag, V or P, X should have in this pattern and whether V+X forms a word by the process such as reanalysis.

otherwise, give the internal structure.

- treat V+X as one word if X is in the following list: 给[give]; 为[become], 成[become], 作[treat as], 到[arrive], 出[out]; 自[from], 向[toward], 入[in], 以[with].

  Ex:
  - 给[give]: 送给[give/send to]/VV, 交给[hand in]/VV, (赠送[give as a gift to]/VV 给[give]/VV)/V.

  - 为[to], 成[become/into], 作[do/as], 到[arrive], 出[out]: (翻译[translate]/VV 成[become/into]/VV)/V, 当作[treat as]/VV, 起到[take effect]/VV, 找到[find]/VV, (认识[realize]/VV 到[reach]/VV)/V, 决出[decide victors]/VV.

  - 自[from], 向[toward], 入[in], 以[with]: 来自[come from]/VV, 面向[face toward]/VV, 流入[flow into]/VV, 迈向[step toward]/VV, 报以[respond with]/VV, 加以[supplement with]/VV.

- treat V+X as two words if X is in the following list: 在[at], 似[like].

  Ex: 生[to be born]/VV 在[at]/P, 坐[sit]/VV 在[at]/P, 留[stay]/VV 在[at]/P, 深[deep]/VA 似[like]/P 海[sea]/NN.

- treat V+X as one word or two words (V+P) according to the meaning of the X, if X is in the following list: 于[at].

  If 于 in V + 于 can be replaced by 在[at], tag V+于 as two words (V+P). Otherwise, tag it as one word.

  One word: 等于[equal to]/VV, 缘于[due to]/VV, 大于[bigger than]/VV, 小于[smaller than]/VV, 无助于[of no help to]/VV 低于[lower than]/VV, 利于[be beneficial for]/VV, 有利于[be beneficial for]/VV.

  Two words: 生[to be born]/VV 于[at]/P, 建[build]/VV 于[at]/P.

2.10.15 Others

Generally, in X+V(or V+X) where X modifies V, if X cannot modify other verbs, or V cannot be a predicate without the X, treat X+V as one word.

Ex: 以期[in order to]/VV

2.11 Adverb: AD

Adverbs are separated from the XP that it modifies.

Adverbs that modify numbers: 近[almost]/AD 三十[thirty]/CD, 5[five]/CD 分[minute]/M 多[odd]/AD 钟[minute]/NN.[10]

The string such as 极大[extremely big] is an adverb when it modifies VPs, not AD+VA, because the VA(大[big]) cannot modify VPs without the AD(极[extremely]).

2.11.1 Reduplication

When VA(or AD) reduplicates, the resulting word can be an AD.

Ex: 好好[well]/AD 干[do]/VV, 常常[always]/AD, 仅仅[only]/AD.

2.11.2 DT+M/N

The following are tagged as ADs when they modify VP/S: 这样[this way]/AD (这样[this way]/AD 做[do]/VV), 同机[on the same airplane]/AD (同机[on the same airplane]/AD 到达[arrive]/VV).

2.11.3 P+PN

We treat the following as two words: 为[for]/P 此[this]/PN.

2.11.4 P+N

The following can be seen as frozen PPs. Since they have the same function as the ADs, we treat them as words, and tag them as ADs: 迄今[until now], 沿途[on the way], 即席[impromptu], 为何[why](为何[why]/AD 愈演愈烈[get worse and worse]/VA), 为什麽[why](为什麽[why]/AD 来[come]/VV).

2.11.5 PN+LC

If a PN+LC totally loses the function of an NP and the string acts like an adverb, treat it as an adverb.

We treat the following as ADs: 此外[in addition]/AD.

2.11.6 Others

If in that context a string totally loses the function of the XP(where X is the head of the string) and the string behaves like an adverb, tag it as AD.

We treat the following as ADs: 进一步[a step further]/AD.

2.12 Preposition: P

Separate it from NP/S that follows it.

Most prepositions are monosyllabic. Some common bisyllabic prepositions are: 为了[in order to], 随着[along with], 沿着[along], 本着[in conformity with], 鉴于[due to], 除了[except], 经过[through],

[10]Note: 50多分钟 is segmented as 50多[50-odd]/CD 分钟/M.

作为[being/regard as], 截止[until].

When a coverb follows a verb, we have to decide whether the word is part of a verb compound. A list of such coverbs are: 于, 给, 为, ... See Section 2.10.14 for details.

2.13 Subordinating Conjunction: CS

Separate it from the XP that follows it.

Strings such as 只有[only] is ambiguous:

- CS: 只有[only if]/CS ... 才[then]/AD ....
- AD+VE: 他[he] 只[only]/AD 有[have]/VE 三[three]/CD 块/M 钱[money]/NN ⟨He only has three dollars⟩.

2.14 Conjunction: CC

Separate it from the XPs that it conjoins.

Ex: 和[and]/CC, 与[and]/CC.

2.15 Particle: DEC, DEG, DEV, DER, AS, SP, ETC, and MSP

Separate it from the XP that it attaches to.[11]

Most particles are monosyllabic. One of bisyllabic particles is 的话[if so]/SP.

2.16 Interjection: IJ

Treat it as one word.

Ex: 哈[expressing satisfaction and so on]/IJ.

2.17 Onomatopoeia: ON

Treat it as one word.

Ex: 哈哈[sound of laughter]/ON, 哗啦啦[sound of water/rain]/ON

[11] In the literature (e.g., [lD92]), it has been argued that some of these particles such as 得, 了 are affixes. For the sake of compatibility with other guidelines and also because it is very easy to automatically group these particles with preceding words, we separate the particles from the preceding words.

2.18 Other noun-modifier: JJ

Separate it from the measure word (M) or the noun (N) that it modifies. Ex: 三[three]/CD 大[big]/JJ 杯[glass]/M 水[water]/NN

When JJs modify nouns, the JJs can be adjectives, 区别词(非谓形容词), or "phrasal words". Most of the "phrasal words" have two parts: X+Y, both X and Y are monosyllabic, and X or Y is the short-form of the corresponding words. Some examples of the "phrasal words" are as follows:

2.18.1 V+N

V+N: 随军[being with the army]/JJ 妓女[prostitute]/NN, 旅英[having studied in England]/JJ 学者[scholar]/NN, 成套[forming a complete set]/JJ 设备[equipment]/NN, 发稿[sending manuscripts to press]/JJ 时间[time]/NN, 获奖[receiving award]/JJ 学者[scholar]/NN, 驻华[being stationed in China]/JJ 使馆[embassy]/NN, 给惠[giving benefit]/JJ 国家[nation]/NN,

2.18.2 AD+VA

AD+VA: 最新[the newest]/JJ 消息[news]/NN, 超大[extra-large]/JJ 规模[scale]/NN 集成[intergrate]/NN 电路[circuit]/NN, 较大[relatively big]/JJ 增长[growth]/NN.

The common "AD": 最[the most], 超[extra-], 较[relatively].

2.18.3 VA+N

VA+N/M: 高层[high-ranking]/JJ 人士[official]/NN, 高速[high speed]/JJ 公路[highway]/NN, 大幅[big size]/JJ 标语[slogan]/NN.

2.18.4 CD+N

CD+N/M: 两国[two-nation]/JJ 关系[relation]/NN, 多国[multi-nation]/JJ 部队[troop]/NN

2.18.5 P+N

P+N/LC: 对外[foreign]/JJ 政策[policy]/NN

2.18.6 Others

others: 关贸[tariff and trade]/JJ 总协定[treaty]/NN, 年均[annual average]/JJ 增长率[growth rate]/NN, 上述[aforementioned]/JJ 三[three]/CD 国[nation]/NN, 历届[all previous sessions]/JJ 世界[world]/NN 体操[gymnastics]/NN 大赛[championship]/NN, 有关[related]/JJ 方面[parties]/NN.

2.19 Punctuation: PU

Treat it as one word, except when it is part of another word; for example, "," in a number (e.g., 123,456/CD) or "·" in proper names, (e.g., 卡尔[Karl]·马克斯[Marx]/NR).

2.20 Foreign word: FW

Treat it as one word, except when it is part of another word (e.g., 卡拉OK[Karaoke]/NN).

2.21 Others

2.21.1 Idioms

The frozen idioms(成语) are treated as words when they function as an NP or a VP.

Ex: 各有所好[each has his likes and dislikes]/V, 一比高低[compete]/V.

2.21.2 Telescopic strings

Telescopic strings are treated as one word if they are not too long (less than four characters). If it is too long, segment them according to pauses.

Short strings: 进出口[imports and exports]/NN 贸易[trade]/NN, 国内外[foreign and domestic]/NN 形势[situation]/NN.

Long strings: 交响[symphony]/JJ 乐团[orchestra]/NN, 北京[Beijing]/NR 市长[mayor]/NN.

2.21.3 Short form

Shortened part is treated as one word. If the shortened part is longer than 3 syllables, segment them according to phonologic evidence (e.g., pauses). The structure of the short form might be different from that of the full form.

Ex: 三好[three-merit]/JJ 学生[student]/NN, 教科文[education, science,and culture]/NN 组织[organization]/NN (UNESCO), 七中[the seventh central government]/NN 全会[convention]/NN.

Chapter 3

Collocation with Some Morphemes

3.1 Strings with zhe5

Some prepositions end with 着.

Ex: 随着[along with]/P.

3.2 Strings with zhi1

zhi+LC, where LC is monosyllabic, is treated as one word (LC).

Ex: 之外[aside from]/LC, 之中[among]/LC.

zhi1+CD is treated as DEG+CD (e.g., 方法[method]/NN 之/DEG 一[one]/CD, 方法[method]/NN 之/DEG 三[three]/CD).

For simplicity, 之一 in a sentence such as 中国 是 发展中 国家 之一 is treated as one word and tagged as an NN.

zhi1+N is treated as DEG+N (e.g., 少年[Children]/NN 之/DEG 家[Club/Center]/NN).

3.3 Strings with bu4

If X in X+不[not] (or 不[not]+X) must co-occur with bu4 or the meaning of X+不[not] is not compositional, we treat X+bu4 as one word.

Words that include bu4(不[not]): 不到[less than] (不到[less than] 5 分钟[minute]), 不足[less than] (不足[less than] 5 公斤[kilogram]), 不便[inconvenient], 不久[not before long].

3.4 Strings with shi4

For simplicity, we treat 特别是[particularly]/AD as one word.

3.5 Strings with xie1

The following are treated as one word: 这些[these]/PN(or DT), 一些[some]/CD.

3.6 Strings with you3

V+有[have] is often a verb; for example, 刻有[engraved with]/VV, 具有[possess]/VV, 富有[rich]/VV.

mei2you3(没有) is always treated as one word(VV or VE or SP).

Many idioms include the word 有[have]; for example, 若有所思[as if lost in thought]/VV.

The following are two words: 有[have]/V 所/MSP, 仅[only]/AD 有[have]/V, 有[have]/V 可能[possibility]/NN.

The following are ambiguous without the context:

- you3-dian3(有点): V[have]+M or AD[a little bit]

  It is V+M when 点 can be dropped or replaced by 一点[a little bit].

  you3-dian3 is an AD when it can be replaced by other degree adverbs such as 很[very] or when it is followed by a VP.

  . 他[he]/PN 有点[a little bit]/AD 下不了[unable to get off]/VV 台[stage]/NN ⟨He felt embarrassed⟩.

  . 这[this]/DT 本/M 书[book]/NN 有[have]/V 点/M 意思[meaning]/NN ⟨This book is interesting⟩.

  . 这[this]/DT 本/M 书[book]/NN 有[have]/V 点/M 看头[worth reading]/NN ⟨This book is worth reading⟩.

- you3-de5(有的): V[have]+DEC or DT[some]

  . 他[he] 有[have]/V 的/DEC 书[book] 我[I] 也[also] 有 [have] ⟨The books that he has, I have, too⟩.

  . 有的[some]/DT 人[people] 已经[already] 走[leave] 了[AS] ⟨Some people have already left⟩.

- you3-xie1(有些): V[have]+M or DT[some]:

  . 我[I] 只[only] 有[have]/VV 些[some]/M 旧书[old books] ⟨I only have some old books.⟩

  . 他[he] 不[not] 像[like] 有些[certain]/DT 人[people] 专门[especially] 爱[like] 抬杠[argue] ⟨He is not like certain people who especially like to argue⟩.

- zhi3-you3(只有): AD[only]+V[have] or CS[only if]:

  . 你[you] 只有[only]/CS 学习[learn] 才[then]/AD 能[able to] 改进[improve] 工作[work] ⟨You can only improve your work by learning⟩.

  . 他[he] 只[only]/AD 有[have]/VV 10 块[M] 钱[dollars] ⟨He only has ten dollars⟩.

3.7 Strings with zai4

One word: 正在[in the process of]/AD.

3.8 Strings with zi4ji3

Always treat PN+zi4ji3 (自己[self]) as one word.

Ex: 他自己/PN.

Chapter 4

Common Collocations

4.1 As one word

- AD: 迄今为止[until today], 迄今[until now], 进一步[one step further], 越来越[more and more], 同机[on the same airplane], 沿途[on the way], 即席[impromptu].
- DT: 这些[these].
- JJ: 对外[foreign] (e.g., 对外[foreign]/JJ 政策[policy]/NN), 各界[all circles]/JJ.
- LC: 之间[between], 在内[inside].
- NN: 其中[among them], 一行[group traveling together].
- P: 为了[in order to].
- V: 来自[come from], 面向[face toward], 流入[flow in], 迈向[step toward], 报以[respond with], 为期[scheduled for a specific duration of time], 有利于[be beneficial for].

4.2 As two words

- AD-like: 并[yet]/AD 未[not]/AD.
- CC-like: 及[and]/CC 其[his/its/her]/PN, 而[and]/CC 又[in addition]/AD.
- DT-like: 各[each]/DT 个/M.
- NN-like: 超大[extra-large]/JJ 规模[scale]/NN, 我[our]/PN 国[nation]/NN.
- NT-like: 零点[midnight]/NT 零一分[one]/NT ⟨one minute past midnight⟩.

4.3 Other cases

V-V: (迎上[step forward]/VV 前去[go forward]/VV)/V.

Appendix A

Comparison with Other Guidelines

In this appendix, we compare our guidelines with the guidelines from PRC [LTS93] and from Rocling [Chi96]. The grouping of words in our system is done in bracketing stage.

	Ours	PRC	Rocling	Example
Verb				
AA	AA	AA	AA	看看
ABAB	ABAB	AB AB	ABAB	研究研究
AABB	AABB	AABB	AABB	高高兴兴
ABB	ABB	ABB	ABB	绿油油
AAB(excl AA-看)	AAB	AAB	AAB	蒙蒙亮
ABAC etc.	ABAC	ABAC	ABAC	有条有理
AA-看	(AA/V kan/V)/V	AA kan	AA kan	说说看
A-yi-A	(A/V yi/CD A/V)/V	A yi A	A yi A	走一走
A-le-A	(A/V le/AS A/V)/V	A le A	A le A	走了走
A-le-yi-A	(A/V le/AS yi/CD A/V)/V	A le yi A	A le yi A	走了一走
nonreduced A-not-A	(A/V not/AD A/V)/V	A not A	A not A	喜欢不喜欢
reduced A-not-A	(A/V not/AD A/V)/V	A-not-A	A-not-A	喜不喜欢
V-R(R is monosyl.)	v-r except v/V 完/V	v-r	v-r	打破
V-R(R is bisyl.)	(v/V r/V)/V	v r	v r	扫 干净
V-de/bu-R	(v/V de/DER r/v)/V	v de r	v de r	打得破
(V-R exists)	(v/V bu4/AD r/v)/V	v bu r	v bu r	打不破
V-de/bu-R	v-de-r/V	??	v-de-r	来得及
(V-R doesn't exist)	v-bu-r/V	??	v-bu-r	来不及
V-DIR	(v/V dir/V)/V	v dir	v-dir	走上来
V-x-O	v/V x/X o/N	v x n	v x n	吃了饭
VO	depends	depends	depends	关心,吃饭
V-de	v/V de/DER	v de5	v de5	走得
V-AS	v/V as/AS	v as	v as	走了

Table A.1: Comparison with PRC's and Rocling's Guidelines

	Ours	PRC	Rocling	Example
Nouns				
Proper Names(NR)				
LstNm+FstNm	one seg	two segs	one seg	王鸣
lstNm+title	name/NR title/NN	name title	name title	王市长
NR + 接尾词	nr-nn/NR	depends	nr-nn	北京市
NR + common noun	nr/NR nn/NN	nr nn	nr nn	北京大学
complex names	several segs	depends	several segs	北京第一服装厂
Common nouns				
N+men5	one seg	one seg	two segs	学生们
VA+N	depends	depends	depends	小媳妇
N+N	depends	depends	depends	牛肉
Temporal nouns				
name of time	cd-year/NT	cd year	cd-year	1998年
count of time	cd/CD year/NN	cd year	cd year	3年
DP-related				
CD	one seg	??	one seg	一万三千
CD+X+CD	one seg	several	one seg	三分之一
AD + CD	ad/AD + cd/CD	ad cd	ad cd	约三百
CD + X	cd-X/CD	cd X	cd-X	三百多
di4-CD	di4-cd/OD	di4 cd	di4-cd	第一
CD+M	cd/CD m/M	cd m	cd m	这个
M + M	m-m/M	m-m	m-m	片片
yi1+M+M	yi1/CD m-m/M	yi1 m-m	yi1-mm	一片片
yi1-M-yi1-M	yi1/CD m/M yi1/CD m/M	??	yi1 m yi1 m	一个一个
Markers				
V-AS	v/V as/AS	v AS	v AS	打了
V-de	v/V de/DER	v de5	v de5	走得
SP	one seg	one seg	one seg	吗
de5(的, 地)	one seg	one seg	one seg	我的, 高兴地
zhi1(之)+CD/N	two segs	two segs	two segs	方法 之 三
zhi1(之)+LOC	one seg	??	one seg	之外
Others				
成语(no insertion)	one seg	one seg	one seg	鼠目寸光
ACROM	one seg	one seg	one seg	北大

Table A.2: Comparison with PRC and Rocling's Guidelines(Ctd)

Appendix B

Treebank Part-of-Speech Tagset

The following is the Part-of-Speech Tagset used in our Penn Chinese Treebank.

AD	adverb	还
AS	aspect marker	着
BA	把 in ba-construction	把, 将
CC	coordinating conjunction	和
CD	cardinal number	一百
CS	subordinating conjunction	虽然
DEC	的 in a relative-clause	的
DEG	associative 的	的
DER	得 in V-de const. and V-de-R	得
DEV	地 before VP	地
DT	determiner	这
ETC	for words 等, 等等	等, 等等
FW	foreign words	I S O
IJ	interjection	啊
JJ	other noun-modifier	男, 共同
LB	被 in long bei-const	被, 给
LC	localizer	里
M	measure word	个
MSP	other particle	所
NN	common noun	书
NR	proper noun	美国
NT	temporal noun	今天
OD	ordinal number	第一
ON	onomatopoeia	哈哈, 哗哗
P	preposition excl. 被 and 把	从
PN	pronoun	他
PU	punctuation	、？。
SB	被 in short bei-const	被, 给
SP	sentence-final particle	吗
VA	predicative adjective	红
VC	是	是
VE	有 as the main verb	有
VV	other verb	走

Table B.1: Our POS tagset in alphabetical order

Bibliography

[Chi96] Chinese Knowledge Information Processing Group. Shouwen Jiezi - A study of Chinese Word Boundaries and Segmentation Standard for Information Processing (in Chinese). Technical report, Taipei: Academia Sinica, 1996.

[lD92] John Xiang ling Dai. The Head in Wo Pao De Kuai. *Journal of Chinese Linguistics*, 1992.

[LTS93] Y. Liu, Q. Tan, and X. Shen. Segmentation Standard for Modern Chinese Information Processing and Automatic Segmentation Methodology, 1993.

[Pac98] Jerome L. Packard, editor. *New Approaches to Chinese Word Formation*. Mouton de Gruyter, 1998.

[SW87] Anna Maria Di Sciullo and Edwin Williams. *On the Definition of Word*. The MIT Press, 1987.

[XPX+00] Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Shizhe Huang, Tony Kroch, and Mitch Marcus. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. In *Proc. of the 2nd International Conference on Language Resources and Evaluation (LREC-2000)*, Athens, Greece, 2000.