

Segmentation Guideline(draft)

Fei Xia

December 18, 1998

1 Overview

In Chinese, it is hard to define what a word is. Our approach is based on both linguistic and engineering consideration. The notion “word” in our system is “syntactic atom”. In this paper, “word” will include both compounds and simple words.

For the sake of compatibility with other systems, we give internal structures of some words. Ideally, each leaf unit in the internal structure will be a simple word. In practice, there are some exceptions because there are

- (Un1-Un2) Unclear cases:

The boundaries among phrases, compounds, simple words, and morphemes are not clear sometimes. For the treatment of unclear cases, see the section ??.

- (Comp) Compounds without internal structures:

That happens only when the internal structure of the compounds can be easily constructed automatically AND many of the other systems don't segment the compounds either.

Examples: numbers, personal names, reduplications.

- (Simp) “Simple” word with internal structures:

That happens only when the “wordhoodness” of some morphemes is controversial and/or many of the other systems treat the morphemes alone as segments.

Examples: aspect markers, 得 in V-de.

Our goal is: in the final project, 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.

This paper lists mainly the decision for each case without going to detail elaborating other alternatives and the reasoning behind each decision. For more information, please refer to the supplement paper which includes dozens of tests for wordhoodness and the compoundness and the explanation for some decisions made in this paper.

1.1 Treatment for unclear cases

There are two types of unclear cases:

- (Un1) There are alternative treatments of a construction. Once a decision is made, it will apply to all the instances of the construction.

Ex: A-not-A, V-de construction, V-R, potential form, etc.

Our approach: we will choose one analysis, and annotate according to that analysis. Make sure the annotation is easy to convert to that of another analysis if the other analysis turns out to be correct.

- (Un2) Two constructions have the same forms and right now we don't have tests to tell them apart.

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 now.

- (Un2-pc): If the collocation is either a phrase or a compound, we segment the string.
- (Un2-cs): If the collocation is either a compound or a simple word, we don't segment the string. The POS of the string is largely determined by the word formation of the string except for a few exceptions.

That will guarantee that the current annotation is incomplete but not wrong in the sense that in the future we will only add information(e.g. word boundary) not remove information.

1.2 Post-processing

After the manual annotation, we will automatically add the word boundary around the following collocation:

- V+AS
- V+DER in V-de construction

We will also work on the tests for deciding whether certain N+N or JJ+N is a compound or not. We will also work on the V+P combination and check if the V+P has been reanalysed as a V.

2 Specification

Assume that the text has been segmented into chunks, the next step is to decide if each chunk is a word or a phrase. The section 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 subsubsection; or simply use the “word” formation information. For some common collocation, simply searching for the string might help.

2.1 Common nouns: NN

2.1.1 Names of relative

Treat it as one word.

Ex: 三叔/NN, 表叔/NN, 大姑父/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: 三叔/NN, 三连/NN, 一方/NN, 三者/NN, 一行/NN
two words: 一/CD 学生/NN

2.1.3 DT+N

Treat it as one word if both DT and N are monosyllabic and N is bound. Otherwise, treat it as two words.

Currently, a DT+N is treated as one word if and only if

- the DT+N is in the followig list: 本人/PN. or
- the DT is monosyllabic and the N is in the following list: 校, 国

One word: 本人/PN, 本校/NN, 全国/NN,
Two words: 本/DT 单位/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.

Currently, a PN+N is treated as one word if and only if

- the PN is monosyllabic and the N is in the following list: 校, 国

One word: 我校, 我国

Two words: 我/PN 单位/NN

2.1.5 JJ+N

The pattern is: X+N, where X modifies the N, X is either a JJ or part of the word.

Note: JJ+N can be a phrase. e.g. 全国性/JJ 网络/NN is extended into “全国性/JJ 观测/VV 苏梅克-列维/NR 9号/NN 彗星/NN 撞击/VV 木星/NN 的/DEC 网络/NN” in one of the files we annotated.

Xs in X+N have many types:

- X is prefix-like: treat X+N as JJ+N if N can be replaced by a NP.
one words: 啊爸/NN, 非商业化/JJ 宗旨/NN.
two words: 原/JJ: 原/JJ 在/P 华/NR 老挝/NR 难民/NN; 前/JJ: 前/JJ 民主德国/NR,
- X is a non-predicate adjective: if both JJ and N are monosyllabic, tag it as one word, otherwise, treat it as JJ+N.
one word: 女人/NN.
two words: 共同/JJ: 共同/JJ 利益/NN,
- X is an adjective: treat it as one word if the meaning is noncompositional. For unclear cases, if both JJ and N are monosyllabic, treat JJ+N as one word, e.g. 鲜花/NN, 强队/NN, 红茶/NN.
one word: 小媳妇/NN, 大洲/NN, 高山/NN.
two words: 厚/JJ 书/NN.

2.1.6 LC+N

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

Ex: 前院/NN, 前天/NT, 左肩/NN.

2.1.7 N+LC

When the N and LC are monosyllabic, the N is non-referential or bound, and the N is not modified by Det-M or other modifiers, treat N+LC as one word, otherwise, tag it as N+LC.

One word(some of them might be two words in other context): 室内(室内/NN 训练/NN); 境外(境外/NN 集团/NN); 天下/NN, 国内/NN, 午后/NT, 赛前/NT.

Two words: 中午/NT 以后/LC.

2.1.8 N+N: N1 modifies N2

If it is 1+1 or 2+1, treat it as one word. i.e. We treat all monosyllabic nouns as “接尾词”. If a noun 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/NN.

For other cases, the string is treated as two words.

one word: 北京市/NR, 研究室/NN, 发展史/NN, 始祖鸟/NN, 残疾人/NN, 警师会/NN, 清晰度/NN, 紧迫感/NN, 大奖赛/NN, 太阳系/NN.

two words: 北京/NR 大学/NN(later, they will be grouped as a N), 北京市/NR 副市长/NN(later, they will NOT be grouped as a N). 玩具/NN 工厂/NN, 合作/NN 领域/NN, 史学/NN 研究/NN,

2.1.9 PN+LC

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

One word: 此间/NT(or NN), 其中/NN(#167), 何时/NT

2.1.10 V+N

If V modifies N, treat V+N as one word(N).

V+N: 烤肉/NN, 炒菜/NN,
V/N+N: 证明信/NN, 讨论会/NN

2.2 Proper Nouns: 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.马克思/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. Note: 张/NR 李/NR 两/CD 位/M 教授/NN.

Ex: 张/NR 教授/NN

2.2.4 Name of Organization/Country/School/..

If it consists of several words, we don't group them as compounds now.

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/NR, 黄河/NR, 沙市/NR, 黑龙江省/NR.

complex names(we will group them later): 北京/NR 大学/NN, 北京/NR 第一/OD 服装厂/NN, 美国/NR 国会/NN

2.2.5 NR+NR: coordination without conjunction

Treat it as two words.

Ex: 中/NR 美/NR, 中/NR 美/NR relation/NN, 东/NR 新/NR 澳/NR.

2.3 Temporal nouns: NT

The names of years/months/day/.. etc. are words.

Ex: 1998年/NT 3月/NT 21日/NT, 5点钟/NT, 初一/NT, 去年/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/NT, 5点钟/NT, 90年代/NT,
two words: 3/CD 年/M, 3/CD 个/M 月/NN.

2.4 Localizer: LC

Localizers are separated from the noun it attaches to.

Localizers have two types:

- short-form localizers are monosyllabic: e.g. 内, 后
- long-form localizers are bisyllabic: e.g. 之间, 以来, 以后, 左右, 后面, 期间

2.5 Pronoun: PN

Treat it as one word.

personal pronoun: 他们/PN, 他自己/PN, 自己/PN,

others: 这里(the string can be a DT+M in other context): 这里/PN 有/VE
..., 这/DT {一/CD} 里/M 路/NN.

2.6 “Determiner”: DT

We separate DTs from the following words, e.g. 这/DT 三/CD 个/M 人/NN,
各/DT 国/NN,

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

There are some bisyllabic DTs: 个别, 全体, 其余, 一切, 这些, 那些, 所有

2.7 Cardinal number: CD

Treat it as one word without the internal structure. Note: the internal structure is very easy to recover if needed.

Pure numbers: 一亿三千万/CD, 30.1/CD, 123,456/CD, 35.6%/CD, 30万/CD,
30几/CD.

Estimation: e.g. 三四十/CD 岁/M

CD + X + CD(5.5.4): X is a morpheme such as 余, 分之, 点. e.g.
三十几亿/CD, 三分之一/CD, 三十点一/CD, 好几/CD 个/M

CD+X: X is a morpheme such as 余, 来: e.g. 四千一百余/CD 人/NN,
三十来/CD 个/M

2.8 Ordinal number: OD

Treat it as one word without internal structure.

Ex: 第一/OD

2.9 Measure word: M

Treat the measure word(incl. reduplication and compound measure word)
as one word. Treat the string such as 分钟 as one word.

Ex: 杯/M, 杯杯/M, 架次/M, 分钟/M

2.10 Verbs

Verbs(V) in this section includes VA, VC, VE, VV.

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

Treat it as one word with no internal structure.

- AA, A is a verb: AA/V
Ex: 看看/VV, 红红/VA
- ABAB: AB is a verb: ABAB/V
Ex: 研究研究/VV, 雪白雪白/VA
- AABB, AB is a verb: AABB/V
Ex: 来来往往/VV, 高高兴兴/VA
Note: most time, AA or BB is not a word.
- AAB(except for AA-看): AAB/V
Ex: 蒙蒙亮/VA
Note: most time, AA or B is not a word.
- ABB: ABB/V
Ex: 绿油油/VA, 红彤彤/VA,
Note: most time, A or BB is not a word.
- ABAC, etc.: ABAC/V
Ex: 马里马虎/VA, 有条有理/VA, 一清二楚/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.

Note: we annotate the internal structures for this type for two reasons: the compatibility with other guidelines and le5 here might be an inflectional morpheme.

- AA-看: [AA/V 看/V]/V:
Ex: [说说/VV 看/V]/V

Note: 看 might be an AS.

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

2.10.3 A-not-A

Treat it as one word with internal structure.

Ex: [高/VA-p1 不/AD 高兴/VA]/VA, [来/VV 没/AD 来/VV]/VV, [喜/VV-p1 不/AD 喜欢/VV]/VV

2.10.4 AD+V

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

- no free word can intervene between AD and V,
- the V can not 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, 胡打, 敬献, 尚余: 尚余/VV 一千零七十五/CD 名/M 老挝/NR 难民/NN, 历任, 并列, 不畏/VV

Two words: 已经/AD 采取/VV, e.g. 不/AD 应该/VV, 没/AD 完成/VV

2.10.5 MSP+V

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

One word: 以期/VV 在与美国、瑞典、挪威这些世界强队交锋中 ...

2.10.6 N+V

Some subject-predicate strings can be either a phrase or 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. e.g.

one word: This homework 让/VV 我/PN 很/AD 头疼/VA. 名列/VV 榜首/NN.

two words: He 头/NN {很/AD} 疼/VA.

2.10.7 V+N

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

If the V and the N are adjacent,

- If V-N is semantically transitive and its object can occur after N only when VN are adjacent (so V is not ditransitive), then treat V+N as one word. e.g. 投资/VV, 出席/VV, 关心/VV, 为期/VV.
- If V and VN have similar meaning and both are semantically intransitive, then treat VN as a word, e.g. 睡觉/VV.
- If N is “bound”, treat VN as one word. e.g. 游泳/VV.
- If the V or the N is short form of another word AND V-N is 1+1, treat VN as one word. e.g. 盈利/VV, 无法/VV, 辞职/VV
- For unclear cases, if V-N is 1+1 AND the meaning is “noncompositional”, treat V-N as one word, e.g. 念书/VV, 流血/VV.

treat V-N as two words: e.g. 曾/AD 七/CD 次/M 访/VV 华/NR.

2.10.8 V+R

The tests for V-R verb: the potential forms(V-de-R, V-not-R) exist. So our definition of V-R includes resultative/directional verb compounds etc. e.g. 看见 is a V-R. but it does NOT include words like 改善, 鼓动.

We treat it as one word. For the sake of compatibility, we give the internal structure for some words.

For the internal structure, we use the syllable-count test:

If V-R is 1+1 AND R is not in the following list, don't give the internal structure of V-R, otherwise, give the internal structure.

The list of Rs where V-R is tagged as [V R]/V: 完,

words without internal structure: 吃掉/VV, 看见/VV, 擦净/VV

words with internal structures: [做/VV 完/VV]/VV, [擦/VV 干净/VV]/VV, [认识/VV 到/VV]/VV.

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.

Words with internal structure: [擦/VV 不/AD 净/VA]/VV, [擦/VV 得/DER 净/VA]/VV.

Words without internal structure: 吃不了/VV, 买不起/VV

Note: V-de-V can be ambiguous between potential form and V-de structure, e.g. “this table 擦得干净吗?” (can this table be wiped clean? or Has the table been wiped clean?). Normally, they can be differentiated by meaning, the position of the object and whether the second V can be modified by adverbs.

2.10.10 V+DIR

See the section for V-R.

Words with internal structure: [走/VV 出去/VV]/VV, [走/VV 不/AFF 出去/VV]/VV

Words without internal structure: 走出/VV, 想出/VV.

2.10.11 V+AS

It is one word with internal structure. We'll do the grouping later.

Ex: 走/VV 了/AS(in the final product, it will be [走/VV 了/AS]/V).

2.10.12 V+DER

The pattern is V-de in V-de construction. V-de is one word with internal structure. We'll do the grouping later.

Ex: 走/VV 得/DER 很/AD 快/VA (in the final product, it will be [走/VV 得/DER]/V).

2.10.13 Verb coordination without conjunctive words

Treat it as multiple words.

Ex: 宣传/VV 鼓动/VV

2.10.14 V+P/V

The pattern is V+X, where X is either a P or a V and X is monosyllabic and V+X is a chunk.

We first decide if it is a word, then decide whether to give the internal structure or not. For the second question, we use syllable-count unless stated otherwise: i.e. if V is monosyllabic, don't give the internal structure; otherwise, give the internal structure.

The decision on the first question is made according to each X:

Type 1: V+X is a word(or a V-V compound):

- gei3(给): if the pattern is V + gei3 + NP1 + NP2(or NP1 is preposed), where NP1 and NP2 are the objects of V-gei3, treat V-gei3 as one word.

Ex: 送给/VV, 交给/VV, [赠送/VV 给/VV]/VV

- wei2 etc.(为, 成, 作, 到, 出): it is a word.

Ex: without internal structure: 当作/VV, 起到/VV, 决出/VV

with internal structure: [翻译/VV 成/VV]/VV, [认识/VV 到/VV]/VV, 找到/VV, [感觉/VV 到/VV]/VV.

- zi4 etc(自, 向, 入, 以): it is a word.

Ex: 来自/VV, 面向/VV, 流入/VV, 迈向/VV, 报以/VV

Type 2: V+X is one word, or two words(V+P). We might later decide to group those two words together as V-P compound.

- yu2(于): If V alone can not be a verb or the meaning of V+yu2 is different from V+P(P is 在/由于/对/被), tag V+yu2 as one word, otherwise, tag it as V+P (we might later decide to group them as one compound).

Ex: one word: 等于/VV, 缘于/VV, 大于/VV, 小于/VV, 无助于/VV, 低于/VV, 利于/VV

two words: 生/VV 于/P, 建/VV 于/P, 有利/VV 于/P.

- zai4 etc(在, 似): as two words V+P for now.

Ex: 生/VV 在/P, 坐/VV 在/P, 留/VV 在/P, 深/VA 似/P 海/NN

2.10.15 Others

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

Ex: 各具/VV 特色/NN

2.11 Adverbs: AD

Adverbs are separated from the XP that it modifies.

Adverbs that modify numbers: 5/CD 分/M 多/AD 钟/NN, 近/AD 三十/CD,

The strings such as 越来越, 从不 are treated as adverbs because part of them alone can not be adverbs. e.g. He 从不 lies. *He 从 lies.

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

2.11.1 DT+M

The following are ADs when they modify VP/S: 这样/AD: 这样/AD 做/VV, 同时/AD(#112),

2.11.2 Reduplication

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

Ex: 好好/AD 干/VV, 常常/AD, 仅仅/AD.

2.11.3 P+PN

We tentatively treat the following as two words: 为/P 此/PN,

2.11.4 P+N

We treat the following as ADs: 迄今(#46), 沿途(#113), 即席(#122),

We tentatively treat the following as AD: 为何: 为何/AD 愈演愈烈/VA; 为什麼: 为什麼/AD 来/VV.

2.11.5 PN+LC

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

We tentatively treat the following as ADs: 此外/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 tentatively treat the following as ADs: 进一步(#82)

2.12 Prepositions: P

Separate it from NP/S that follows it.

Most prepositions are monosyllabic. The common bisyllabic prepositions are: 为了, 随着, 沿着, 本着, 鉴于, 除了, 经过, 作为, 截止

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

2.13 Conjunctions: CS and CC

Separate them from the XP that it connects.

Strings such as 只有 is ambiguous: 只有/CS ... 才/AD ..., 他只/AD 有/VE three dollars.

2.14 Particles: DEC, DEG, DEV, DER, SP, AS, MSP

Markers include localizers, 得, 的, 地, sentence-final ending, aspect marker, and misc. particles such as 所, 以.

Except AS and DER(see section for verbs), other markers are considered words and are separated from the XP it attaches to.

Most particles are monosyllabic. Some common non-monosyllabic particles are: 的话, 来说, 为止

2.15 IJ

Treat it as one word.

2.16 ON

Treat it as one word.

Ex: 哈哈/ON, 哗啦啦/ON

2.17 JJ

Separate it from the M or N that it modifies. Ex: 三/CD 大/JJ 杯/M 水/NN

When JJs modify nouns, the JJs can be adjectives, 区别词(非谓形容词), and “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.

Here, we list some examples of the “phrasal words”.

2.17.1 V+N

V+N: 随军/JJ 妓女/NN, 旅英/JJ 学者/NN, 成套/JJ 设备/NN, 发稿/JJ 时间/NN, 获奖/JJ 学者/NN, 驻华/JJ 使馆/NN, 给惠/JJ 国家/NN,

2.17.2 AD+VA

AD+VA: 最新/JJ 消息/NN, 超大/JJ 规模/NN 集成/NN 电路/NN,

The common “AD”: 最, 超

2.17.3 VA+N

VA+N/M: 高层/JJ 人士/NN, 高速/JJ 公路/NN, 大幅/JJ 标语/NN (#77).

2.17.4 CD+N

CD+N/M: 两国/JJ 关系/NN, 多国/JJ 部队/NN

2.17.5 P+N

P+N/LC: 对外/JJ 经济/NN

2.17.6 others

others: 关贸/JJ 总协定/NN, 年均/JJ 增长率/NN, 上述/JJ 三/CD 国/NN, 历届/JJ 世界/NN 体操/NN 大赛/NN 有关/JJ 方面/NN

2.18 PU

Treat it as one word, except when it is part of a word, e.g.

“,”: in a number, e.g. 123,456/CD

2.19 FW

Treat it as one word, except when it has become part of a word, e.g. 卡拉OK/NN.

2.20 X

Treat it as one word.

2.21 Others

2.21.1 Idioms

The frozen idioms(成语) are treated as words,
e.g. 各有所好/V, 一比高低/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/NN 贸易/NN, 国内外/NN 形势

long strings: 交响/JJ 乐团/NN, 北京/NR 市长/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:

e.g. 三好/JJ 学生/NN, 教科文/NN 组织/NN 七中/NN 全会/NN

3 Collocation with some morphemes

3.1 Strings with zhe5

Some prepositions end with zhe5.

Ex: 随着, 沿着, 本着

3.2 Strings with zhi1

zhi1+LC where LC is monosyllabic is treated as one word(LC): e.g. 之外/LC, 之中/LC

zhi1+CD is treated as DEG+CD. e.g. 方法/NN 之/DEG 一/CD, 方法/NN 之/DEG 三/CD.

zhi1+N is treated as DEG/DEG+N. e.g. 少年/NN 之/DEG 家/NN.

3.3 Strings with bu4

If X in X+不(or 不+X) must co-occur with bu4, then X+bu4 is part of a morphological word.

Words that include bu4(不): 不到 5 minutes, 不足 five pounds, 不便, 从不, 不久,

3.4 Strings with shi4

Words:

特别是/AD (#67),

3.5 Strings with xie1

The following are treated as one word: 这些/PN(or DT), 一些/CD

3.6 Strings with you3

V+有 is often a verb: 刻有/VV, 具有/VV, 富有/VV,

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

Many idioms include the word you3: e.g. 若有所思/VV

you3suo3(有所) is tagged as 有/V 所/MSP.

The following are two words: 仅/AD 有/V(#42), 有/V 可能/NN(#43),

The following are ambiguous without the context:

- you3-dian3(有点): V+M or AD

Ex: He 有点/AD 下不了台.

This book 有/V 点/M 意思.

This book 有/V 点/M 看头.

you3-dian3 is an AD when it can be replaced by other degree adverbs such as hen3 or when it is followed by a VP.

It is V+M when 点 can be dropped or replaced by 一点. you3-dian3 in you3 + dian3 + XP is an AD when:

- you3-de5(有的): V+DEC or DT

Ex: you3-de5/DT people already left.

He you3/VV de5/DEC book Mary too you3/VV.

- you3-xie1(有些): V + M or DT:

Ex: Mary 不 like you3-xie1/DT people.

Mary 只(only) you3/VV xie1/M old books.

- suo3-you3(所有): DT or MSP + V:

Ex: 所有/DT 这些/DT answers 都错了.

Mary 所/MSP {拥}有/V de5/DEC 一切 ...

- zhi3-you3(只有): it is a word(tagged as CS) or AD+V:

You zhi3-you3/CS study 才/AD 能 improve your work.

Mary zhi3/AD you3/VV three dollars.

3.7 Strings with zai4

One Word: 正在/AD(#44),

3.8 Strings with zi4ji3

When PN+zi4ji3 is in the object position, it is always one word. When it is in the subject position, it can be either one word(PN) or two words(PN+AD).

4 Common Collocations

4.1 As one word

AD: 迄今为止, 迄今(#46), 进一步(#82), 越来越, 同机(#112), 沿途(#113), 即席(#122)

DT: 这些,

JJ: 对外(#173, 对外/JJ 经济/NN), 各界/JJ

LC: 之间, 在内

NN: 其中(#167), 一行(#110),

P: 为了

V: 来自, 面向, 流入, 迈向, 报以, 为期

4.2 As two words

AD-like: 并/AD 未/AD(#160),

CC-like: 及/CC 其/PN(#195), 而/CC 又/AD,

DT-like: 各/DT 个/M

NN-like: 超大/JJ 规模/NN (#167), 我/PN 国/NN,

NT-like: 零点/NT 零一分/NT (#175),

VV-like: 有利/VV 于/P(#143)

4.3 Other cases

V-V: [迎上/VV 前去/VV]/VV(#102),

others: 总的/AD 来/MSP 说/VV,

5 Comparison with Other Guidelines

Table 1: Comparison with PRC and Rocling’s Guidelines

	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
unreduced 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-p1 not/AD A/V]/V	A-not-A	A-not-A	喜不喜欢
V-R(R is monosyl.)	IntStr depends	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/bu r	v de/bu r	打得破
(V-R exists)	[v/V bu4/AD r/v]/V			
V-de/bu-R	v-de(bu)-r/V	??	v-de/bu-r	来得及
(V-R doesn’t exist)				
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	v de5	v de5	走得
V-AS	[v/V asp/AS]/V	v asp	v asp	走了

Table 2: Comparison with PRC and Rocling’s Guidelines(Ctd)

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