October 2000

The Part-Of-Speech Tagging 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-00-07.

The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0)

Abstract
This document describes the Part-of-Speech (POS) tagging 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 POS tagging 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 criteria for POS tagging and other factors that we considered when designing our POS tagset. Second, we describe each of the thirty-three POS tags in detail. Third, we provide tests to distinguish certain POS tag pairs and specify the treatment for some common collocations. Fourth, we list a number of words with each POS tag. Finally, we compare our tagset with three tagsets: the tagset for the Academia Sinica Balanced Corpus in Taiwan (CKIP, 1995), the tagset for the Grammatical Knowledge Base developed by Peking University in China (Yu et al., 1998), and the tagset for the English Penn Treebank (Santorini, 1990).

Comments
University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-00-07.

The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0)

Fei Xia

October 17, 2000

Contents

Chapter 1

Introduction

Chinese has little, if any, inflectional morphology. For example, words are not inflected with respect to tense, case, person, and number. As a result, it is often difficult to decide how a word should be tagged in a particular context.

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 POS guidelines, like the segmentation 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 this guidelines into a technical report, which is published by the Institute for Research in Cognitive Science (IRCS) of the University of Pennsylvania.

1.1 Tagging criteria

The central issue in Part-of-Speech (POS) tagging is whether the POS tagging should be based on meaning or on syntactic distribution. This issue has been debated since the 1950s [Gon97] and there are still two different viewpoints. For example, a word such as 毁灭 in Chinese can be translated into *destroy/destroys/destroyed/destroying/destruction* in English and it is used roughly the same way as its counterparts in English. According to the first view, POS tags should be based solely on meaning. Because the meaning of the word remains roughly the same across all of these usages, it should always be tagged as a verb. The second view says that POS tags should be determined by the syntactic distribution of the word. When 毁灭 is the head of a noun phrase, it should be tagged as a noun in that context; when it is the head of a verb phrase, it should be tagged as a verb.

[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 ⟩).

We have chosen syntactic distribution as the main criterion for our POS tagging because it complies with the principles adopted in contemporary linguistics theories, such as the notion of head projections in the X-bar theory and the GB theory.

One argument that is often used against the syntactic distribution approach is that, because many verbs in Chinese can also occur in noun positions, thus requiring two POS tags, using the syntactic distribution approach would increase the size of the lexicon. We believe that this argument is not convincing for two reasons. First, the two POS tags allow us to distinguish between verbs that can occur in noun positions and verbs that cannot (such as monosyllabic verbs and verbs in reduplication forms such as AABB, A-not-A). If there are generalizations about which verbs can occur in noun positions and which cannot, these generalizations can be represented as morphological rules, which allow the lexicon to be expanded automatically. On the other hand, if no such generalizations exist and the nominalization process is largely idiosyncratic, it supports the view that this is a lexical phenomenon and verbs that can be nominalized should have two POS tags in the lexicon. Second, the phenomenon that many verbs can occur in noun positions is not unique to Chinese, and the standard treatment in other languages is to give them both tags.

1.2 POS tagset

Our POS tagset has 33 tags:[2]

Verb, adjective (4): VA, VC, VE, VV.

Noun (3): NR, NT, NN.

Localizer (1) : LC.

Pronoun (1) : PN.

Determiner and number (3): DT, CD, OD.

Measure word (1): M.

Adverb (1): AD.

Preposition (1): P.

Conjunction (2): CC, CS.

Particle (8): DEC, DEG, DER, DEV, SP, AS, ETC, SP, MSP.

Others (8): IJ, ON, PU, JJ, FW, LB, SB, BA.[3]

[2]We want our tagset to be complete; therefore, we include tags such as ON and IJ, which do not occur in our corpus.

[3]Whether bei4(被) and ba3(把) are prepositions or verbs is highly controversial. In the tagging stage, we don't make any commitment to that. That's the reason why we tag them as LB, SB, BA, respectively, rather than tagging them as P or VV.

1.3 Size of the POS tagset

Suppose we start with a small POS tagset that most people will agree on, which includes tags for nouns, verbs, adverbs, prepositions, and so on. The question is whether we should replace each tag T with a set of more specific tags, say $\{T_1, T_2, ..., T_i, ..., T_n\}$. There are several factors that needed to be considered:

- In general, for words with the same POS tag to have exactly the same syntactic distribution, a large tagset is needed. On the other hand, the larger the tagset is, the harder the annotation is. Because the annotators have to remember more tags, more tests and apply them consistently. Therefore, when we design a tagset, we need to make compromise.
- Let set(T) be the set of words with POS tag T. If there are no good tests to assign every word in set(T) to one of the $set(T_i)$, and $\{T_i\}$ do not provide significant more information than T, then we will not split T into $\{T_i\}$.
- If the set(T) is a closed set and for each pair (i, j), $set(T_i) \cap set(T_j)$ is empty, then $word/T$ can be replaced with $word/T_i$ automatically by a simple conversion program. Therefore, if we want to tag words with specific tags in $\{T_i\}$, this can be done automatically by a program, rather than manually by annotators.

1.4 Handling difficult cases

Sometimes it is not clear whether a word in a context should be tagged as X or Y. If we know for sure that the word is not in $set(X) \cap set(Y)$, then we can simple choose a tag, say X, for the word and later replace X with Y if necessary. For example, the word 许多[many] in 许多[many] 学生[student] is either a JJ or a DT or a CD, and it should have exactly one of the three tags. We simply choose one tag that seems to be more appropriate, and the tag can be replaced later automatically if needed.

On the other hand, if we know that in another context the word is tagged as one of the two tags, say X, then we should decide whether or not the word behaves the same in these two contexts. If we are positive that the behavior is the same, we should tag it as X; otherwise, tag it as Y. For example, 又 in "(1) 又[both] ... 又[and] ... (e.g., 又[both] 高兴[happy] 又[and] 难过[sad])" is either an AD or a CC, and 又 in "(2) 他[he] 又[again] 来[come] 了 [AS]⟨He came again⟩" is clearly an AD. Because we are not sure whether the behavior of 又 in these two contexts are the same, we tag the word in (1) as a CC. The rationale behind this decision is that, if later we want to tag the word in (1) as an AD, we can simply replace 又/CC with 又/AD. But if we tag the word in (1) as AD *now* and later we want to change it to CC, we have to distinguish the two contexts and make sure only the 又/AD in (1), not in (2), is changed to 又/CC.

1.5 Notation

Some notations used in this document:

- A word without context can have multiple tags "a word w is in set(T)" means T is one of the tags that w has.
- Tag N represents all noun tags (NT/NN/NR). Tag V represents all verb tags (VA/VV/VC/VE). "Det+M" is a short-hand for $DT + (OD|CD) + M$, where DT, OD, and CD may or may not be present.

- "A word can be negated" is a short form of saying "a word with a positive meaning can be negated". Similarly, "a word can occur in A-not-A" means "a word can occur in the question pattern A-not-A".
- For the gloss, we do not translate measure words, particles, and words with tags LB, SB, BA, or VC. Instead, we mark the words with their POS tags.

Chapter 2

The Treebank Part-of-Speech Tagset

2.1 Verb: VA, VC, VE, VV

Normally, a verb satisfies the following:

- Verbs (except auxiliary verbs etc.) serve as the predicate of a clause (main clause or embedded clause).
- Verbs can be negated by 不[not] or 没[not].
- Aspect markers can be attached to most (but not all) verbs.
- Most verbs can occur in A-not-A.

If a word w in set(V) is the head of an NP, it is tagged as N, not as V. If w in set(V) is a noun modifier (excluding the case where the V is the head of a relative clause), it is tagged as N or JJ(according to the tests for N and JJ), not as V.

2.1.1 Predicative adjective: VA

VA roughly corresponds to adjectives in English and stative verbs in the literature on Chinese grammar.[1] Our VAs include two types:

Type 1: predicates that have no object and can be modified by 很[very].

Type 2: predicates derived from type 1 either through reduplication(e.g., 红彤彤[bright red]) or through the pattern N + A meaning "as A as an N"(e.g., 雪白[snow white]). This type of VAs don't have objects, but some of them cannot be modified by 很[very] either, because the intensifying meaning is already built-in.

Note: when a word in set(VA) modifies N without 的[DEC], it is tagged as JJ or a noun, rather than as VA. When a word in set(VA) has an object, it is tagged as VV, rather than VA. For example, 这[this] 项/M 活动[activity] 丰富[enrich]/VV 了/AS 他[he] 的/DEG 生活[life] ⟨This activity enriched his life⟩.

[1]The definition of stative verbs varies from system to system. Some include psych-verbs such as 喜欢[like], 了解[understand], and 怨恨[hate]. In our system, those psych-verbs are tagged as VV. One open question about adjectives is whether they form a subclass of verbs in Chinese. We will not get into the debate.

2.1.2 Copula: VC

The words 是[be] and 为[be] are tagged as VC. 非 is also tagged as VC if it means 不[not] 是[be] and there is no other verb in the sentence.

The word 是[be] has several usages:

- Link two NPs/Ss: 他[he] 是[be]/VC 学生[student] ⟨He is a student⟩,
- In cleft-sentences: 他[he] 是/VC 昨天[yesterday] 来[come] 的/SP ⟨It was yesterday that he came⟩.
- For emphasis: 他[he] 是[VC] 喜欢[enjoy] 看[see] 书[book]⟨He does enjoy reading books⟩.

Currently, in all these cases the word is tagged as VC.

2.1.3 you3 as the main verb: VE

Only 有[have], 没[not]{有[have]}, and 无[not have] are tagged as VE when they are the main verbs (including the possessive you3, existential you3, etc.).[2]

2.1.4 Other verb: VV

This includes the rest of the verbs, such as modals, raising predicates (e.g., 可能[maybe, probably]), control verbs (e.g., 要[want], 想[want to]), action verbs (e.g., 走[walk]), psych-verb (e.g., 喜欢[like]/了解[understand]/怨恨[hate]), and so on.

2.2 Noun: NR, NT, NN

A noun can be an argument of a predicate or a preposition. In general,

- Nouns cannot be modified by degree and negation adverbs such as 很[very] and 不[not].
- Many nouns can be modified by Det+M structure.
- Nouns can modify nouns directly (i.e., without 的/DEG).

If a word is the head of an NP, it is tagged as a noun.[3] Sometimes it is hard to tell whether or not a phrase is an NP. Some tests for NPs are:

- If the phrase XP is modified by a Det+M phrase, and the Det+M phrase in other context only modify NPs, then the XP is likely to be an NP.
- If the phrase XP is the argument of a verb or a preposition which in other context takes only NP arguments, then the XP is likely to be an NP.
- If the phrase XP is modified by "ZP 的/DEG-or-DEC", then the XP is likely to be an NP.

[2]The main reason why we assign those verbs with a new tag is that the treatment of existential sentences is controversial. Giving these verbs a different tag will make it easy to find the existential sentences in the corpus.

[3]In this Treebank, we assume the head of a NP is a noun, not a classifier or a determiner.

2.2.1 Proper Noun: NR

Proper Nouns (NRs) are a subclass of nouns.[4] An NR is a name of a particular person, politically or geographically defined location (cities, countries, rivers, mountains, etc.), or organization (corporate, governmental, or other organizational entity). A proper noun is usually unique and cannot be modified by a Det+M.

- The names of the following are NRs:
  region/country/county/city, mountain/river, newspaper/journal, organization/company, school/association/foundation, person/family.
- The names of the following are NOT NRs: nationality (e.g., 中国人[Chinese]), race (e.g., 白人[Caucasian]), title (e.g., 教授[professor]), disease, occupation, organ (e.g., 肺[lung]), instrument (e.g., 钢琴[piano]), game (e.g., 足球[soccer]), flower (e.g., 玫瑰[rose]), etc.

2.2.2 Temporal Noun: NT

Temporal Nouns can be the objects of prepositions such as 在[at], 从[since], 到[until], or 等到[until], They can be refered by 这个时候[at this moment], and questioned by 什麽时候[when]. They can also modify VP/S directly. Like other nouns, NTs can be arguments of some verbs.

Temporal Nouns are either the names of the time (e.g., 1990年[1990], 一月 [January], 汉朝[Han Dynasty]) or formed by PN+LC, N+LC, DT+N.

Ex: 一月[January], 汉朝[Han Dynasty], 当今[present time], 何时[when], 今后[from now on]

2.2.3 Other Noun: NN

NN includes all other nouns. NNs, except the ones for locations, normally cannot modify VPs with or without 地/DEV.

2.3 Localizer: LC

Many nouns alone cannot be the argument of prepositions such as 在[at] and 到[until] or modify VP/S directly. One function of localizers is to attach to the preceding NP/S so that the whole phrase can act as the argument of those prepositions or modify VP/S.

Some localizers can stand alone as the arguments of the prepositions/verbs. Some localizers can be modified by 最[the most]. Localizers cannot be modified by Det+M.

Localizers are of two types:

[4] Although there are some differences between proper nouns and common nouns with respect to syntactic distribution, we assign proper nouns a different tag from other nouns mainly for pragmatic reasons. First, they are useful for some NLP applications, such as information extraction task. Second, most unknown words in a corpus are proper nouns. The user of this corpus can take advantage of this tag in order to identify and deal with unknown words. In this sense, the user can regard an NR as an NN with some additional information, and this additional information may or may not be helpful for their applications. Because there is no perfect test to distinguish proper nouns from common nouns, it is not always easy to tell them apart, especially for the names of organizations. As a result, "this additional information" might not be consistently marked in the corpus. To be as consistent as possible, we made a list of things that we will tag as NRs. Fortunately, mis-tagging an NN as an NR or vice versa will not affect syntactic structures.

- fan1wei4ci2 (方位词): this type of localizer denotes direction, location and so on. They come from nouns. Some can stand alone as the arguments of the prepositions/verbs. Some can be modified by 最[the most]. They cannot be modified by Det+M.
  - mono-syllabic localizers: e.g., 前[before], 后[after], 里[in], 外[out], 内[in], 北[north], 东[east], 边[side], 侧 [side], 底[end/bottom], 间[between], 末[end], 旁[next to].
  - bisyllabic localizers: they are formed by
    * mono-syllabic localizers plus morphemes such as 以, 之 etc.
      Ex: 之间[between], 以北[to the north of].
    * two mono-syllabic localizers.
      Ex: 前后[around], 左右[around], 上下[or so], 东北[northeast].
- others: we tag the following as LCs.[5]
  . 为止[until]: 到[at] 目前[present] 为止[until] ⟨until now⟩.
  . 开始[starting from]: 从[from] 四月[April] 开始[starting from] ⟨starting from April⟩.
  . 来[ever since]: 5 年[year] 来[ever since] ⟨in the past five years⟩.
  . 以来[since]: 1998年[1998] 以来[since] ⟨since 1998⟩.
  . 起[since]: 一九九三年[1993] 起[since] ⟨since 1993⟩.
  . 在内[inside]: 包括[include] 他[he] 在内[inside] ⟨including him⟩.

2.4 Pronoun: PN

Pronouns function as substitutes for noun phrases and denote persons or things asked for, previously specified, or understood from the context. They are normally not modified by Det-M or adjectival expressions.[6]

PNs include personal pronouns (e.g., 我[I], 你[you]), demonstratives when used alone as NPs (e.g., 这[this], 此[this]), possessive pronouns (e.g., 其[his/her/its]), and reflexives (e.g., 我自己[myself], 自己[self]).

2.5 Determiners and numbers: DT, CD, OD

2.5.1 Determiner: DT

This includes demonstratives (e.g., 这[this], 那[that], 该[the]) and words such as 每[every], 各[each], 前[the preceding], 后[the following]. DTs does NOT include cardinal numbers and ordinal numbers.

See Section 4.11 for all the DTs.

[5] We could choose to mark some of them as verbs, but this will complicate the bracketing annotation.

[6] Because unlike English pronouns in Chinese are not inflected with case markers, it is not always easy to decide whether some words (such as 大家[everyone], 本人[oneself], 本身[self]) should be tagged as PN or NN. To be consistent, we compiled a list of PNs appeared in our 100K-word corpus.

2.5.2 Cardinal Number: CD

It includes cardinal numbers (optionally followed by 概数词[approximate number indicators] such as 来[over/odd], 多[odd], and 好几[over]) and words such as 好些[some], 若干[several], 半[half], 许多[many], 很多[many] (e.g., 很多[many] 学生[student]).

Ex: 1245, 一百[a hundred].

2.5.3 Ordinal Number: OD

Ordinal numbers(序列词) are tagged as ODs. We treat 第+CD as one word, and tag it as OD.

Ex: 第一百[the one hundredth].

2.6 Measure word: M

Measure words follow determiners/numbers to form Det+M structure to modify nouns or verbs. They include classifiers (e.g., 个), group measure words (e.g., 群[group]), and words such as 公里[kilometer] and 升[liter].

Some measure words can be modified by a limited set of adjectives (e.g., 一[one]/CD 小[small]/JJ 瓶[bottle]/M 水[water]/NN). 临时量词[temporary measure word] can be modified by nouns and adjectives (e.g., 一[one]/CD 铁[iron]/NN 箱子[box]/M 书[book]/NN).

2.7 Adverb: AD

The adverb is a big class. It includes manner adverbs, frequency adverbs, degree adverbs, conjunctive adverbs, and so on. The behaviors of adverbs differ a lot. The main function of most adverbs is to modify a VP or an S.

Ex: 仍然[still], 很[very], 最[most], 大大[greatly], 又[again], 约[approximately].

2.8 Preposition: P

A prepositions can take a noun phrase or a clause as its argument.

Note: words such as 把/BA and 被/LB-or-SB are not tagged as P. See Section 2.11 for detail.

Ex: 从[from], 对[to/for].

2.9 Conjunctions: CC, CS

Note: the words that are called 连词[connective words] in traditional Chinese grammar books are tagged as CC, CS, or AD according to their syntactic distribution. CC conjoins two equivalent constituents (noun phrases, clauses, etc.) of the same function, whereas CS precedes a subordinating clause. Conjunctive adverbs often appear in the main clause and pair with a subordinating conjunction (e.g., 如果[if]/CS ... 就[then]/AD).

2.9.1 Coordinating conjunction: CC

A coordinating conjunction (CC) conjoins two constituents, each of which has approximately the same function as the whole construction. The main pattern for CCs is: XP {,} CC XP.

Ex: 与[and], 和[and], 或[or], 或者[or], 还是[or].

2.9.2 Subordinating conjunction: CS

Words that join two clauses, one subordinating to the other, are tagged as subordinating conjunctions (CS). The patterns for CSs are:

- CS S1, S2.
- S2 CS S1.

Where S1 is the subordinating clause, S2 is the main clause. In the first pattern, the CS can also appear after the subject of S1; For example, 你[you] 要是[if] 不[not] 去[go], 我[I] 就[then] 去[go] 了/SP. ⟨If you don't want to go, then I will go.⟩

Ex: 如果[if]/CS ... 就[then]/AD ....

2.10 Particle: DEC, DEG, DER, DEV, AS, SP, ETC, MSP

2.10.1 de5 as a complementizer or a nominalizer: DEC

This only includes 的 and 之 when they function as a complementizer or a nominalizer(e.g., 吃[eat] 的/DEC). The pattern is: S/VP DEC {NP}.

Note: 的 also has other tags:

- DEG: 他[he] 的/DEG 车[car] ⟨his car⟩.
- SP: 他[he] 是/VC 一定[definitely] 要[must/should] 来[come] 的/SP ⟨He should definitely come⟩.
- AS: 他[he] 是/VC 在[at] 这里[here] 下[get off] 的/AS 车[car] ⟨It was here that he got off the bus⟩.

2.10.2 de5 as a genitive marker and an associative marker: DEG

This only includes 的 and 之 when they function as a genitive marker or an associative marker. The pattern is: NP/PP/JJ/DT DEG {NP}.

Note: 的 has other tags: DEC, SP, and AS.

2.10.3 Resultative de5: DER

de5(得) is tagged as DER in potential form V-得-R, and in V-de construction (他[he] 跑[run] 得/DER 很[very] 快[fast] ⟨He runs very fast⟩).

Note: Some collocations ending with 得 are not V-de constructions. They are verbs (e.g., 记得[remember], 获得[gain]).

2.10.4 Manner de5: DEV

This only includes 地 when it occurs in "XP 地 VP", where XP modifies the VP. In some old literature, 的 is used in this pattern too. In that case, we will tag that 的 as DEV.

Ex: 高兴[happy]/VA 地/DEV 说[speak]/VV ⟨speak happily⟩.

2.10.5 Aspect Particle: AS

Verbal particles that indicate aspect are tagged as aspect particles (AS). This category includes ONLY 了, 着, 过, and 的.

2.10.6 Sentence-final particle: SP

SP often appears at the end of a sentence. For example, 他[he] 好[good] 吧[SP] ⟨Is he OK⟩ ?

Some of them can also be used for a pause. For example, 他[he] 吧[SP], 人[people] 很[very] 好[good] ⟨Speaking of him, he is a very nice guy⟩.

Ex: 了,呢,吧,啊,呀,吗.

2.10.7 ETC

The tag is used for the word 等 and 等等. Two patterns are:

- XP 等 NP: 科技[science and technology] 、 文教[culture and education] 等/ETC 领域[area].
- XP 等/等等: 科技[science and technology] 、 文教[culture and education] 等等/ETC.

2.10.8 Other particle: MSP

This includes particles, such as 所, 以, 来, and 而, when they appear before a VP.[7]

The MSPs are:
. 所: 他[he] 所[MSP] 需要[need] 的/DEC ⟨what he needs⟩.
. 以 or 来: 用[use] ... 以/MSP (or 来) 维持[maintain] ⟨use... to maintain⟩.
. 而: 为[for]... 而[MSP] 奋斗[fight] ⟨to fight for...⟩.

See Section 4.19 for details.

2.11 Others: IJ, ON, LB, SB, BA, JJ, FW, PU

2.11.1 Interjection: IJ

Interjections appear in the sentence-initial position (i.e., "IJ, S").

Ex: 啊[Ah].

[7] suo3(所) differs from the other three words in that it appears between the subject and the VP in a relative clause whereas the other three words appear between a PP/VP and another VP. However, because the set(MSP) is small, we don't give suo3 a different tag.

2.11.2 Onomatopoeia: ON

The term 象声词[onomatopoeia], a word that imitates sounds, has been mentioned in several Chinese grammar books and POS tagsets. However, it is not clear to us whether those words form a unique syntactic category. We have not found any occurrence of 象声词[onomatopoeia] in our 100K-word Treebank; nevertheless, we reserve the tag ON for this type of word. The following are some patterns in which an ON can occur:

- modify VPs in the pattern "ON 地 V": 雨[rain] 哗哗[ON] 地[DEV] 下[fall down] 了[AS] 一[one] 夜[night] ⟨The rain has been pouring down for the whole night⟩.
- modify NPs in the pattern "ON 的 N": 砰[ON] 的/DEG 一声[a sound] ⟨Bang⟩!
- form a sentence by itself: 砰砰[ON]! 屋里[in the house] 传出[spread] 两[two] 声/M 枪响[gunfire] ⟨Bang! Bang! Two sounds of gunfire spread out from the house⟩.
- ONs normally cannot be modified by adverbs, etc.

Ex: 哗啦啦, 咯吱

2.11.3 bei4 in long bei-construction: LB

This only includes 被, 叫, 给(in spoken language), and wei2(为) when they occur in the long bei-construction (i.e., NP0 + LB + NP1 + VP). For example, 他[he] 被/LB 我[I] 训[scold] 了/AS 一[one] 顿/M ⟨He was scolded by me⟩.

Note: 叫 is tagged as VV when it is used as a telescopic verb; for example, 他[he] 叫[order]/VV 你[you] 去[go] ⟨He ordered you to go.⟩.

2.11.4 bei4 in short bei-construction: SB

This only includes 被 and 给 (in spoken language) when they occur in the short bei-construction (i.e., NP0 + SB + VP); for example, 他[he] 被/LB 训[scold] 了/AS 一[one] 顿/M ⟨He was scolded⟩.

Note: 给 has other tags: LB, VV, and P (e.g., 你[you] 给[to]/P 他[he] 写[write] 封/M 信[letter] ⟨You should write a letter to him⟩).

2.11.5 ba3 in ba-construction: BA

This only includes 把 and 将 when they occur in the ba-construction (i.e., NP0 + BA + NP1 + VP). For example, 他[he] 把/BA 你[you] 骗[cheat] 了/AS ⟨He cheated you⟩.

Note: 将 has other tags: AD and VV (e.g., 他[he] 将[check]/VV 了[AS] 我[I] 的[DEG] 军[king] ⟨ (In chess) My king is in check by him⟩).

2.11.6 other noun-modifier: JJ

JJs include the following three types:

Type 1 "区别词"(非谓形容词): They modify nouns in the pattern JJ+的+{N} or JJ+N, but they cannot be the predicate of a sentence without the help of 的. They cannot be modified by

degree adverbs.

The patterns: JJ + 的/DEG + N, JJ+N.

Ex: 共同[mutual]/JJ {的/DEG} 目标[goal]/NN, 她[she] 是[VC] 女[female]/JJ 的/DEG ⟨She is a woman.⟩

Type 2 "hyphenated-compound": Those words can be seen as shortened forms of relative clauses or preposition phrases. The words normally have two syllables. One (or both) is a shortened form of a longer word. The common POS combinations for this type of JJ are V+N, P+N/LC, AD+VA, and so on.

The pattern: JJ+N.

Ex: 留美[having studied in the US]/JJ scholar/NN.

Type 3 adjectives: 新[new]/JJ 消息[news]/NN.

The pattern: JJ+N.

Ex: 新[new]/JJ 消息[news]/NN.

Note: when 的/DEC is inserted between the adjective and the noun, the adjective is tagged as VA.[8]

2.11.7 Foreign Word: FW

FW is used to tag foreign words. FW excludes the translations of foreign words. It also excludes the words that have mingled with Chinese words (e.g., 卡拉OK[karaoke]/NN, A型[type A]/NN). It also excludes words whose meaning and POS is clear from the context. We should avoid the tag as much as possible. It is used only when the POS tag is not clear from the context.

2.11.8 Punctuation: PU

Punctuation marks are tagged as PU. If they are part of other words, they are not tagged.

Ex: 张三/NR ,/PU 李四/NR 和[and]/CC 王五/NR, 123,456/CD.

[8] We tag an adjective X in X+N as JJ because:

- Sometimes the distinction between adjectives and non-predicative adjectives in X+N is not clear.
- Unlike in the predicative position, the adjectives in the noun modifier position cannot be modified by adverbs (e.g., 好[good] 学生[student], *很好[very good] 学生[student]).
- Many adjectives cannot occur in this position; that is, without 的/DEC they cannot modify nouns (e.g., 这[this] 个[M] 学生[student] 很[very] 失望[disappointed], * 失望[disappointed] 学生[student]).

Chapter 3

Problematic Cases

3.1 Confusing parts of speech

Two tags X and Y are confusing when

- they have similar functions (e.g., both JJ and CD can modify nouns), or
- some words have both tags. For example, 政治[politics] is a noun most of the time, but it is an AD in 只[only] 能[can] 政治[politically]/AD 解决[resolve]/VV 这[this] 个/M 问题[problem] ⟨This problem can only be resolved through political means⟩.

3.1.1 AD or AS

zai4(在) before a verb is treated as AD, not as AS.

Note: Other adverbs can intervene between zai4 and a verb, but they cannot intervene between a verb and aspect markers.

3.1.2 AD or CC

In the pattern: X+S/VP, X is either an AD or a CC.

A CC links two equivalent XPs whereas an AD does not. A conjunctive adverb often pairs with a CS, and its function is to refer back to the subordinating clause.

The words with both tags are: 又[and/CC, again/AD], 还是[or/CC, still/AD].

ADs: 否则[otherwise], 但是[but], 但[but].

See Section 4.4 for a list of CCs.

3.1.3 AD or CS

In the pattern: X+S/VP, X is either an AD or a CS.

A CS leads a subordinating clause, and it normally can occur before the subject of the clause.

There is no overlap between set(AD) and set(CS).

See Section 4.6 for a list of CSs.

3.1.4 AD or JJ

In the pattern: X+NP, X is either AD or JJ.

A few ADs can modify NPs (e.g., 又[again/another] and 才[only]). There are no overlap between set(JJ) and set(AD) in this position.

An easy test to distinguish ADs and JJs is to insert 的[DEG] between the word and the NP. If the new phrase is still valid, the word is a JJ; otherwise, it is an AD.

3.1.5 AD or NN

In the pattern: X+VP, X is either an AD or an NN.

Temporal nouns and "location" nouns[1] can modify VP/S directly. In those cases, they are still tagged as nouns. For other nouns, if we can insert the preposition "在[at]" before the word without changing the meaning or validity of the sentence, tag the word as NN; otherwise, tag it as AD.

. 政治[politically]/AD (e.g., 政治[politically]/AD 解决[resolve]/VV 这[this] 个/M 问题[problem]⟨to resolve this problem through political means⟩)

.重点[emphatically]/AD (e.g., 重点[emphatically]/AD 抓[focus on]/VV 生产[production]).

3.1.6 AD or NT

In the pattern: X+VP, X is either an AD or an NT.

If the word X can be the head of an NP, tag it as NT. For example, 昨天[yesterday] is an NT, not an AD.

ADs: 早日[sooner]/AD 实现[achieve]/VV.

NTs: 目前[at present]/NT, 今后[from now on]/NT.

3.1.7 AD or VA

The patterns are: X+VP, X+地+VP.

We assume that a VA cannot modify VP directly, It needs the help of a DEV. We also assume that both VA and AD can occur in X+DEV+VP. In the following patterns:

[1]We can roughly define "location" nouns as nouns that indicate the location and can be the argument of the preposition zai4 without localizers.

- w 地 VP: if the meaning of w/AD and w/VA are the same, tag w as VA, otherwise, tag w according to the meaning in the context.
- w VP: tag it as AD.

AD: 大大[greatly]/AD 提高[increase].

VA: 高兴[happy]/VA 地/DEV ⟨happily⟩, 紧密[tight]/VA 地/DEV ⟨tightly⟩.

3.1.8 AD or VV

The pattern is: X+VP/S.

VVs can occur in A-not-A, ADs cannot.

"modal"-like ADs: 大概[probably], 将[will], 一定[definitely].

AD-like VVs: 是否[whether or not]/VV

3.1.9 AS or VV

In the pattern: V+X.

Currently, we have only four AS's. Some R in V-R compounds (e.g., 完[finish] in 写[write] 完[finish] ⟨finish writing⟩, 起来[start to do something] in 跑[run] 起来[start to do something] ⟨start to run⟩) have similar functions as AS's. We treat them as VV, not as AS.

3.1.10 CC or P

Some CCs (e.g., 与[and], 和[and], 跟[and], and 同[and]) are also prepositions.

In "NP0 X NP1", X is either CC or P:

- If NP0 and NP1 are permutable, then X is a CC.
- If X is preceded by any modifying adverbial, it is a P.
- If NP0 shows higher topicality and/or empathy than NP1, then X is a P.
- (If NP0 is eliminated, then X+NP1 must be eliminated), then X is a CC.

3.1.11 CS or P

The argument of a CS must be an S, whereas the argument of a P can be either an NP or an S.

Both CS and P can take a clause as its argument. However, there are some differences between "CS S1, S2" and "P S1, S2". Here we assume that the subjects of S1 and S2 refer to different entities:

- In "P S1, S2", P+S1 can be inserted between the subject of S2 and the VP of S2.

. 对[to]/P 我[I] 不能[cannot] 按时[on time] 来[come], 他[he] 感到[feel] 很[very] 失望[disappointed] ⟨I cannot come on time, and he is very disappointed about this⟩.

. 他[he] 对[to]/P 我[I] 不能[cannot] 按时[on time] 来[come] 感到[feel] 很[very] 失望[disappointed] ⟨I cannot come on time, and he is very disappointed about this.⟩

. 虽然[although]/CS 我[I] 不能[cannot] 去[go], 你[you] 还是[still] 应该[should] 去[go] ⟨Although I cannot go, you still should go⟩.

. *你[you] 虽然[although]/CS 我[I] 不能[cannot] 去[go] 还是[still] 应该[should] 去[go].

- In "CS S1, S2", CS can be inserted between the subject of S1 and the VP of S1.

. 虽然[although]/CS 他[he] 不能[cannot] 上场, 这[this] 个/M 队[team] 还是[still] 会[will] 赢[win] ⟨Although he cannot play this time, this team will still win the game⟩.

. 他[he] 虽然[although]/CS 不能[cannot] 上场, 这[this] 个/M 队[team] 还是[still] 会[will] 赢[win] ⟨Although he cannot play this time, this team will still win the game⟩.

. 对[to]/P 我[I] 不能[cannot] 按时[on time] 来[come], 他[he] 感到[feel] 很[very] 失望[disappointed] ⟨I cannot come on time, and he is very disappointed about this⟩.

. *我[I] 对[to]/P 不能[cannot] 按时[on time] 来[come], 他[he] 感到[feel] 很[very] 失望[disappointed].

- In "P S1, S2", the S1 can be replaced by an NP.

. 对[to]/P 我[I] 不能[cannot] 按时[on time] 来[come], 他[he] 感到[feel] 很[very] 失望[disappointed] ⟨I cannot come on time, and he is very disappointed about this⟩.

. 对[to]/P 这[this] 件/M 事[incident], 他[he] 感到[feel] 很[very] 失望[disappointed] ⟨He is very disappointed about this incident⟩.

. 虽然[although]/CS 我[I] 不能[cannot] 去[go], 你[you] 还是[still] 应该[should] 去[go] ⟨Although I cannot go, you still should go⟩.

. *虽然[although]/CS 这[this] 件/M 事[incident], 你[you] 还是[still] 应该[should] 去[go].

CSs: 虽然[although]/CS.

Ps: 对[to]/P, 由于/P, 因为/P.

3.1.12 CD or DT

DT includes demonstratives and non-quantitative quantifiers. CD includes quantitative quantifiers. When CD and DT co-occur in an NP, the DT precedes the CD. CD can be used to answer the question "how many/much?"

Words such as 全体[all], 全部[all] and 一切[all] are tagged as DTs now. They can easily be re-tagged if we find other tags are more appropriate.

CDs: 许多[many], 若干[several], 个把[a few].

DTs: 各[each], 全[all], 某[certain/some], 这[this].

3.1.13 CD or JJ

Both CD and JJ can modify nouns. A J describes an attribute that can be used to classify the individuals in a set, whereas a CD cannot (CDs are "quantifiers").

Also, for some JJs, "JJ+{的[DEG]}+N" implies "N 是 JJ 的".

CDs: 一些[some], 大批[a large batch of], 好几[several], 不少[a few].

JJs: 共同[mutual], 女[female].

3.1.14 CD or NT

A word such as 一九九一[1991] or 九一 is tagged as NT when it means 一九九一年[year 1991] (e.g., 一九九一[1991]/NT 至[to] 一九九五年[year 1995]/NT).

3.1.15 DT or JJ

The pattern is: X+N.

DTs are either demonstratives or quantifiers. The quantifiers may have the scoping effect. JJs describe attributes of the nouns. See Section 4.11 for a list of DTs.

The word 前 with both tags:

- JJ: 美国[US] 前[former]/JJ 总统[president] 尼克松[Nixon].
- DT: 前[the previous]/DT 7 个[M] 月[month].

3.1.16 DT or OD

The word 首[the first] is an OD, not a DT (e.g., 首[the first]/OD 届/M).

3.1.17 DT or PN

The patterns are X as an NP, X+NP, and X+的+NP.

Only the demonstratives, such as 这[this] and 那[that], are in both set(PN) and set(DT).

- Demonstratives may occur in all three patterns. They are tagged as PN in "X" and "X+的[DEG]+NP", as DT in "X+NP".
- Words in set(PN) except the demonstratives are tagged as PN in all the three patterns.
- Words in set(DT) except the demonstratives can occur in "X+的[DEG]+NP" and 'X+NP', In both cases, they are tagged as DTs.

Examples:

- Words with both DT and PN tags: 这[this], 此[this].
  - 这[this]/DT 本[M] 书[book] 很[very] 好[good] ⟨This book is good⟩.
  - 这[this]/PN 很[very] 好[good] ⟨This is good⟩.
- Words with only the PN tag: personal pronouns, etc.
  - 他[he]/PN 爸爸[father]/NN ⟨his father⟩, 其[his/her/its]/PN
- Words with only the DT tag:
  - 所有[all]/DT 的/DEG 东西[thing]/NN.

3.1.18 JJ or NN

A JJ cannot be the head of an NP; NN can.[2]

For the sake of consistency and simplicity, we treat the following as NNs, not JJs: N+形[shape], N+状[form/shape], N+制[system].

N+形[shape] as NN: 椭圆形[oval], V形[V shape].
N+状[form/shape] as NN: 带状[belt shape], 颗粒状[small and roundish in shape].
N+制[system] as NN (when 制 means 制度[system]): 货币制[monetary system], 股份制[joint-stock system].

3.1.19 JJ or P

The following words have both tags:

The word 有关[about/P, concerned/JJ] has both tags:

- JJ: 有关[concerned]/JJ 单位[organization] ⟨the organization concerned⟩.
- P: 有关[about]/P 撤军[withdrawing troops]/NN 的/DEG 报告[report]/NN ⟨the report about withdrawing the troops⟩.

3.1.20 LC or NN

There are no overlap between LCs and NNs. See Section 4.17 for a list of LCs. One function of the LC is to attach to a NP/S, so the whole part can be the argument of prepositions or modify a VP/S directly.

[2]This test, however, sometimes is hard to use for two reasons. First, the annotator has to decide whether he can make a sentence where the word in question is the head of an NP; Second, the annotator has to decide whether the same word in the modifier position and in the head position have the same meaning. Neither decision is easy for some words. As a result, the (JJ, NN) pair is one of the hardest pairs to distinguish.

3.1.21 LC or MSP

The word 来 have both tags:

- LC: 近年[recent years] 来[since]/LC ⟨in recent years⟩.
- MSP: 用[use] 暴力[violence] 来/MSP 维持[maintain]

3.1.22 M or NN

The pattern is: CD+X.

If an M can fill the position between the CD and the X without changing the meaning, tag X as NN; otherwise tag it as M. 临时量词 (the measure words that are temporarily borrowed from nouns and which can be modified by other nouns) is tagged as M, too.

Ex: 一[one]/CD 学生[student]/NN, 一[one]/CD 年[year]/M, 一[one]/CD 箱子[box]/M 书[book]/N.

3.1.23 NN or NR

NRs are the names of persons, organizations, countries, and so on. They normally cannot be modified by Det+M. [3]

Words ended with the following are tagged as NNs, not NRs: 方[side], 军[troop], 人[people/race/ethnicity], 会[council/meeting], 队[team], 厅[hall].

Ex: 美方[U.S. side]/NN, 美军[U.S. troop]/NN, 美国人[American]/NN, 办公厅[office]/NN.

3.1.24 NN or NT

The NP headed by an NT can modify S/VP directly and can answer the question "at what time". NNs other than "location" nouns cannot modify S/VP directly.

3.1.25 NN or VA

If the word can be modified by 很[very] in THAT context, tag it as VA/VV.

Ex: 表示[express/show]/VV 乐观[optimistic]/VA.

3.1.26 NN or VV

If the word X is in the head(NP) position, tag it as NN. If it is in head(VP) position, tag it as VV. To be more specific:

- X as head(NP) (e.g., when X appears in Det+M+X): tag X as NN.
- X modifies N in X+N: if X is the predicate of the clause that modifies N such as in a relative or complement clause without 的/DEC, then tag X according to the clause (most likely to be a VA/VV); otherwise, tag X as NN.

[3] As mentioned in Section 2.2.1, the tag NR is created mainly for pragmatic reasons. As a result, it is very hard to find good syntactic tests to distinguish it from NN. We can treat the NR as NN with some additional information.

- adverbial-phrase can modify X: If X in that context can be modified by YP+地[DEV] or by words that can only modify verbs, tag X as VV.
- X is the head of an argument of the verb V: tag X according to the subcategorization of the V.

  If the V takes NP as its argument, then X is an NN. If the V takes VP/S as its argument, then X is a VV. If the V can take either, tag X according to the reading.

  One way to decide whether a verb's argument X is an NP or an S is to check whether or not the argument X can be followed by its own argument. If the argument X cannot be followed by its own argument, then the argument X is an NP, not an S.

  For example, 进行[carry out] 对[to] 宗教[religion] 的/DEG 改革[reform], *进行[carry out] 改革[reform] 宗教[religion], so the object of 进行[carry] in this context is an NP, not an S. Therefore, 改革[reform] in this context is an NN.

  The object of the following Vs is an NP, not a VP:
  - 进行[carry out]: 进行[carry out]/VV 密切[close]/JJ 合作[collaboration]/NN.
  - 加以[supplement with]: 加以[supplement with]/VV 推进[moving forward]/NN).
  - 受到[be given]: 受到[be given]/VV 批评[criticism]/NN.
  - 给以[give]: 给以[give]/VV 奖励[reward]/NN.
  - 予以[give]: 予以[give]/VV 表彰[praise]/NN.

  Note: 得以[so that...can be...] can take a VP/S object; for example, 得以[so that...can be...] 提高[improve]/VV 教学[teaching] 水平 [level] ⟨so that teacher's performance can be improved⟩.

3.1.27 NN or PN

A word of the form DT+N/PN+N has the properties of both a PN and an NP. For consistency, we treat all of them as NNs or NTs.

NN: 本人[oneself]/NN, 本校[our school]/NN, 全球[whole world]/NN, 当地[the place mentioned]/NN, 我校[my school]/NN.

NT: 当今[present time]/NT, 何时[when]/NT.

3.1.28 P or VV

Most prepositions in Chinese come from verbs, and many of them can still be used as verbs in some context; therefore, sometimes it is not easy to distinguish them.

NP0 X NP1 YP, X is a P or a VV:

- If there is no other verb in a complete sentence, X is a verb.

  Ex: 他[he]/PN 在[be at]/VV 家[home]/NN ⟨He is at home⟩.

- If the NP1 is absent or moved (as in short answer or VP-not-V question), then X is a verb.[4]

  Ex: 他[he] 在[be at]/VV 家[home] 不[not] 在[be at]/VV ⟨Is he at home⟩?

- If AS can follow X, then X is a verb.

  Note: some prepositions (e.g., 为了[for]/P) end with 着[AS] or 了[AS], so be careful.

- Some words have both tags, but the meanings might be different.

  Ex: 靠[against/P, depend on/VV]:
  - 他[he] 靠[against]/P 墙[wall] 站着[stand] ⟨He is standing against the wall⟩.
  - 你[you] 不[not] 能[can] 总[always] 靠[depend on]/VV 父母 [parents] ⟨You cannot always depend on your parents⟩.

- If it is still ambiguous after applying the tests above, tag the word as P.

3.1.29 VA or VV

VAs do not have objects and all VAs except the ones with "absolute" meaning can be modified by adverbs such as 很(very).

Some words have both VA and VV tags, such as 丰富[enrich/VV,rich/VA], 少[miss/VV, few/VA], 多[have extra/VV, plenty of/VA]. For example, 这儿[here] 少[miss]/VV 了/AS 一[one] 本/M 书[book] ⟨A book is missing here⟩, 这[here] 书[book] 很[very] 少[few]/VA ⟨There are very few books here⟩.

Note: the term "object" needs further quantification.

- Given a verb and an NP, if both P+NP+V and V+NP are grammatical, treat the NP as the verb's object and tag the verb as VV.
  Ex: 他[he] 对[to] 我[I] 很[very] 关心[care]/VV ⟨He cares about me⟩.

- If the NP or Det-M that follows a verb describes the degree, it is not considered the object of the verb.
  Ex: 鞋[shoe] 大[big]/VA 了[AS] 一号[one size] ⟨The shoes are one size bigger⟩.

For consistency, we treat as VVs all four-character "idioms" which function as VPs.

Ex: 堂堂正正[impressive or dignified in personal appearance], 欣欣向荣 [flourishing; prosperous], 有条不紊[in an orderly way], 扎扎实实[solid], 不得人心[be unpopular], 熟视无睹[to ignore], 微不足道[trivial], 危机重重[crisis-ridden], 闻名于世[world-famous], 无人不晓[renowned], 变幻无常[volatile], 别具一格[having a unique style], 并行不悖[not mutually exclusive], 波澜壮阔[surging forward with great momentum], 出人意料[exceeding one's expectations], 死灰复燃[dying embers glowing again], 腾空而起[rise to the sky], 同室操戈 [family members drawing swords on each other], 万事如意[everything is fine], 稳中求进[strive for further progress while going steadily], 握手言和[shake hands and make peace with each other], 息息相关[be closely linked], 相依为伴[depend on each other], 相映成辉[each shining more brilliantly in

[4]We believe that Chinese does not allow preposition stranding. P can occur in "P-not-P VP". For example, 他[he] 从[from] 没[not] 从[from] 北京[Beijing] 出发[start off] ⟨Did he start off from Beijing⟩?
P-not-P might be a compound preposition, so we don't consider P-not-P as an example of preposition stranding.

the other's company], 一分为二[one divides into two], 因地制宜[suit measures to local conditions].

Note: four-character "idioms" that function as NPs are tagged as NNs.

Ex: 无名之辈[a nobody]/NN, 有识之士[a man of insight]/NN.

3.1.30 VE or VV

Only 有[have], 没[not]{有[have]}, and 无[not have] are tagged as VE when they are the main verbs (including the possessive you3, existential you3, etc.). The main reason that we assign those verbs with VE, not VV, is because the treatment of existential sentences is controversial. Giving these verbs a different tag will make it easy to find the existential sentences in the corpus. Therefore, verbs such as 具[have], 具有[have], 拥有[possess], 富有[rich] are tagged as VV, rather than VE, because they cannot appear in existential sentences.

3.2 Specific words and collocations

3.2.1 de5

de5(的) has four tags: DEC, DEG, AS, and SP.

- DEC: 我[I] 买[buy] 的/DEC 书[book] ⟨the book I bought⟩.
- DEG: 我[I] 的/DEG 书[book] ⟨my book⟩
- SP: 他[he] 是/VC 一定[definitely] 要[should] 来[come] 的/SP ⟨He should/will definitely come⟩.
- AS: 他[he] 是/VC 在[at] 这里[here] 下[get off] 的/AS 车[bus] ⟨He got off the bus right here⟩.

3.2.2 deng3, deng3deng3

deng3(等) and deng3deng3(等等) are tagged as ETCs when they appear in

- XP 等 NP: 科技[science and technology] 、文教[culture and education] 等/ETC 领域
- XP 等/等等: 科技[science and technology] 、文教[culture and education] 等等/ETC

deng3(等) has other POS tags:

M : 二[two] 等[class]/M 大[big] 国[country] ⟨second-class big country⟩.

VV : 他[he] 会[will] 等[wait]/VV 你[you] 的/SP ⟨He will wait for you⟩.

3.2.3 lai2

lai2(来) has several tags:

- VV:
  - 他[he] 来[come]/VV 了/AS ⟨He is coming⟩.
  - 他[he] 拿来[bring]/VV 了/AS 一[one] 本[M] 书[book] ⟨He brought a book⟩.
  - 他[he] 走[walk]/VV 上来[up]/VV ⟨He walked up⟩.
- SP: 他[he] 拿出[take out] 书[book] 来/SP ⟨He took out his book⟩.
- MSP: 他[he] 用[use] 这[this] 个/M 来/MSP 证明[prove] 他[he] 是/VC 无辜[innocent] 的[SP] ⟨He uses this to prove that he is innocent⟩.
- LC: 3 年[M] 来[since]/LC ⟨for the past three years⟩.
- part of a CD: 30来[over thirty]/CD 个[M].

3.2.4 lian2

lian2(连) in the 连[even] ... 都[all]/也[also] is tagged as AD.

Ex: 他[he] 连[even] 我[I] 都[also] 不[not] 认识[know] ⟨He does not even know me (let alone others)⟩.

3.2.5 you3

- you3 (有[have]) is tagged as VE when it is the main verb.
- mei2-you3 (没有[not have/VE, not/VV]):

  - VE: when it is the main verb (e.g., 他[he] 没有[not have]/VE 书 [book]) ⟨He does not have a book⟩.

  - VV: when it modifies VP (e.g., 他[he] 没有[not]/VV 来[come]/VV ⟨He did not come.⟩).
- mei2 (没[not have/VE, not/VV]):

  - VE: when mei2 alone is the main verb (e.g., 他[he] 没[not have]/VE 书[book] ⟨He does not have a book⟩).

  - AD: when mei2 modifies VP (e.g., 他[he] 没[not]/AD 来[come]/VV ⟨He didn't come.⟩) and when it occurs in A-没-A (e.g., 他[he] 来[come]/VV 没[not]/AD 来 [come]/VV ? ⟨Did he come?⟩).
- you3-mei2-you3 (有没有):

  - VE: it is tagged as "(you3/VE mei2you3/VE)/V" when it is the main verb.
  - VV: it is tagged as "(you3/VV mei2you3/VV)/V" when it modifies a VP.

3.2.6 zhe4yang4

zhe4yang4(这样[this]):

- PN: 这样[this]/PN 的/DEG 伙伴[partnership]/NN 关系[relation]/NN ⟨this kind of partnership relation⟩.
- AD: 你[you] 这样[this way]/AD 做[do]/VV 不[not] 对[right] ⟨It's not right to do it this way⟩.
- VV: 你[you] 别[do not] 这样[behave like this]/VV ⟨Don't behave like this⟩.

Chapter 4

Lists of Words for Each POS Tag

For each POS tag X, we lists a few words with that tag. If a word w listed for POS tag X is also in set(Y), we mark it as *w (also Y)*. For each set, we list some words with that tag, especially the ones that are often mistaken for other tags. For example, if a word with tag X is often incorrectly tagged as Y, we list the word under the section for X. The list starts with the string "Y-like X".

4.1 AD

The followings are some ADs occured in the corpus:

- Conjunctive adverbs:

  otherwise: 否则
  therefore: 所以, 因此, 因而
  however: 却
  then, as a result: 那么, 就, 便, 结果, 则, 这样
  in addition: 另外, 此外
  furthermore: 进而
  later: 随后, 然后
  as well: 也
  so that: 以便, 从而
  "for example": 例如, 如
  "that is": 即

- Temporal adverbs: 届时[at the appointed time], 即将[be about to], 紧接着 [right after].
- Frequency adverbs: 多次[many times] (e.g., 多次[many times]/AD 发生[happen]/VV).
- Degree adverbs: 极为[very], 较[comparatively], 较为[comparatively].
- Manner adverbs: 互利[mutually beneficial] (e.g., 互利[mutually beneficial]/AD 合作[collaborate]/VV).
- Modal-like adverbs: 将[will].
- ADs that modify numbers:
  Ex: 近[approximately]/AD, 不足[less than].

- ADs that can precede NPs:
  又[again/another] (又[another]/AD 一/CD 个/M 参与者[participants]/NN).

- Negation ADs: 未[not](also VE), 不[not], 没[not](also VE).
- Phrase-word: 进一步[a step further], 越来越[more and more], 尤其是 [especially], 据称[according to reports]/AD, 据悉[it is reported that]/AD.
- Others: 一起[together]/AD, 另[in addition]/AD, 正在[be in the process of] /AD, 凡[every], 才[then and only then].

4.2 AS

Closed set: 了, 着, 过, 的.

4.3 BA

Closed set: 把, 将.

4.4 CC

Closed set.

'And': 与 (also P [with]), 和 (also P[with]), 跟 (also P[with]), 同 (also P[with]), 及, 以及, 并 (also AD), 并且, 而 (大[big]/VA 而[and]/CC 全[complete]/VA), 而且, 且.

'Or': 或, 或者, 还是 (e.g., 去[go] 还是[or]/CC 不[not] 去[go], also AD).

Paired-CCs:[1] 既[both]/CC .. 又[and]/CC, 又[both]/CC .. 又[and]/CC 不仅[not only]/CC ... 而且[but also]/CC.

Others:[2]
- 至[to]: 九一年[1991] 至[to] 九五年[1995]
- 到[to]: 一月[January] 到[to]/CC 三月[March]
- 兼[and]: 国务[state affairs] 委员[committee member] 兼[and]/CC 科委[committee of science and technology] 主任[director])

4.5 CD

It includes cardinal numbers and quantitative quantifiers such as 几[several], 许多[some], 若干[a few], 数[several], 大部分[most of], 部分[part of], 绝大部分[a large portion of], 大多数[majority of], 大量[large

[1] Like similar collocations in English such as "either ... or" and "both ... and", it is possible that the first words in the pairs are not CCs. Nevertheless, as long as these words are consistently marked as CCs, it is easy to change them later if necessary.

[2] It could be argued that these words listed here are prepositions or verbs when they appear in the pattern "YP X YP". But the conversion of the POS tags and the corresponding structures are pretty easy.

quantity of], 大批[a large batch of], 多数[most of], 多少[how many/much], 多[many], 个把[a few], 很多[many], 一些[some], 好几[a few], 不少[not few], 诸多[many].

Note: 不少[not few] and 很多[many] are tagged as CD when they appear before measure words/nouns (i.e., in the pattern X + {M} + N.) They are tagged as AD+VA when they are predicates as in NP + X. For example, 很多[many]/CD 学生[student]/NN, 学生[student]/NN 很[very]/AD 多[many]/VA.

Note: 很[very] 少[few] and 不[not] 多[many] are not CDs.

4.6 CS

Closed set: 如果[if], 如[if], 若[if], 假如[if], 即使[even if], 不管[no matter], 不论[no matter], 无论[no matter],不但[not only], 尽管[even though], 虽然[although], 虽[although] 只要[as long as], 只有[only when], 一旦[as soon as],

4.7 DEC

Closed set: 的, 之.

4.8 DEG

Closed set: 的, 之.

4.9 DER

Closed set: 得.

4.10 DEV

Closed set: 地.

4.11 DT

Closed set:

- Demonstrative determiners:
  这[this], 这些[these], 此[this], 该[that], 本[our], 那[that], 那些[those], 上[the previous], 下[the next], 前[the previous] (前[the previous]/DT 7 个[M] 月[month]), 后[the later], 头[the first](头[the first]/DT 7 个[M] 月[month]), 另[another], 其余[the rest], 其他[the other], 其它[the other], 某[certain/some], 某些[some],

- Quantifiers (excluding quantitative quantifiers): "every, all, any" and so on:
  . 各[each], 诸[every], 每[every].
  . 何[any], 什么[any], 任何[any].
  . 整[whole]: 整[whole]/DT 个/M 欧洲[Europe].
  . 全[whole]: 全[whole]/DT 省[province]/NN.

. 全体[all], 全部[all]: 全体[all]/DT. 外交[foreign affairs]/NN 官员[official]/NN.
. 一切[all], 所有[all]: 一切[all]/DT 努力[effort]/NN.
. 同[same]: 同[same]/DT 一[one] 天[day] ⟨on the same day⟩.
. 有的[some], 有些[some]: 有的[some] 书[book].

4.12 ETC

Closed set: 等, 等等

4.13 FW

FW is used a dozen of times for words whose meaning we don't know.

4.14 IJ

Closed set: No IJ occur in our corpus.

Ex: 啊, 嘿.

4.15 JJ

Ex: 共同[mutual], 双边[both parties], 很大[great amount/degree/deal], 高科技[high-tech], 有关[concerned] (有关[concerned]/JJ 国际[international]/NN 公约[treaty]/NN ⟨the international treaty concerned⟩), 老牌[old brand]/JJ (老牌[old brand]/JJ 军工[military industry]/JJ 企业[business]/NN).

. 上述[aforementioned], 下列[the following]: 上述[aforementioned]/JJ 三[three]/CD 国[country]/NN

4.16 LB

Closed set: 被, 叫, 给, 为 (e.g., 交换机[switchboard] 市场[market] 为/LB 外国 [foreign] 产品[product] 垄断[monopolize] ⟨The market of the switchboard is monopolized by foreign products⟩).

4.17 LC

Closed set: there are 40+ LCs in our corpus. We just list a few LCs here.

- fang1wei4ci2(方位词):
  - Monosyllabic LCs: 中[among], 间[between], 内[in], 里[in], 外[out], 底[end/bottom], 上[above], 下[down], 前[before], 后[after], 末[end], 边[side], 旁[next to], 畔[riverside], 侧[side], 初[at the beginning], 北[north].
  - Bisyllabic LCs: 以上[above], 以下[below], 以后[after], 以前[before], 以内[within], 以外[beyond/outside], 之间[between], 之前[before], 之外[beyond/outside/except], 之后[after], 之中[among], 之内[within], 之际[at the time], 之初[at the beginning], 左右[around], 前后[around].

- Others: 来[ever since], 以来[since], 时[when], 起[since], 为止[until], 开始[since], 在内[including], 处[at the place of], 止[until].

4.18 M

There are about 130 Ms in our corpus.

- classifiers: 个, 种, 批, 位, 条, 起, 组, 笔, 幢, 点.
- unit: 吨[ton], 公里[kilometer], 平方公里[square kilometer].
- currency: 马克[Mark], 澳元[Australian dollar].
- unit of time: 年[year], 天[date], 秒[second], 分钟[minute][3]
- compound measure word: 人次[number of people], 架次[number of flights], 排排[row].

4.19 MSP

MSPs occur in our corpus:
- 以: 以/MSP 增强[fortify]/VV 总体[overall]/JJ 竞争[competitiveness]/NN 实力[strength]/NN ⟨so as to fortify overall strength of competitiveness⟩
- 而: 为[for]/P 生存[survive]/VV 下去[continue]/VV 而/MSP 不得不[have no choice but to]/VV 采取[take]/VV 的/DEC 行动 [action]/NN ⟨the action which must be taken in order to survive⟩
- 来, 去: 用[use] ... 来/去 维持[maintain] ⟨use... to maintain⟩
- 所: 他[he] 所 需要[need] 的[DEC] ⟨The thing that he needs...⟩

The following are not MSPs: 的话[if so]/SP, 从而[so that]/AD, 以便[so that]/AD.

4.20 NN

Phrase-word: 之一[one of] (目的[purpose]/NN 之一[one of]/NN ⟨one of the purposes⟩)

NN with N+LC structure: 国内[domestic], 海外[oversea].

4.21 NR

Ex: 阿根廷[Argentina], 柏林[Berlin], 克林顿[Clinton].

4.22 NT

Ex: 1990年[year 1990], 最后[at last].

N+LC as a NT: 战后[after the war]/NT, 赛前[before the contest]/NT, 今后[from now on], 日前[the other day]/NT, 何时[when]/NT, 目前[at present]/NT.

PN+LC as a NT: 此后[afterwards]/NT.

[3]We tag them as Ms because no measure words can be inserted between them and the preceding CDs. According to the same test, we tag 小时[hour] and 月[month] as NNs.

4.23 OD

Ex: 第一[the first], 首[the first].

4.24 ON

No ON occurs in our corpus.

Ex: 刷, 哗啦啦.

4.25 P

Closed set: there are about 70 Ps in our corpus.

VV-like prep: 经过[through], 截止[until], 有关[about], 离[from].

CS-like prep: 随着[along with], 沿着[along], 鉴于[due to], 除了[except], 为了[in order to]

AD-like prep: 就[on]/P (就[on]/P 机制[system/mechanism]/NN 问题[question/issue]/NN).

4.26 PN

Closed set: there are about 30 PNs in or corpus.

Ex:

- Personal pronoun: 他[he], 我[I], 你[you], 您[you], 她[she], 它[it], 之[him/her/it], 我们[we], 你们[you], 她们[them (female)], 他们[them (male)], 它们[them (non-human)],
- Demonstratives alone as a NP (also DTs): 这[this], 这儿[here], 那[that], 此[this], 这里[here], 那里[there].
- Possessive pronoun: 其[her/his/its].
- Reflexives: 他自己[himself], 自己[self].
- Others: 彼此[each other], 大家[everyone], 对方[the other party], 双方[both parties], 自身[self].

4.27 PU

Closed set: there are 31 PUs in our corpus.

Some examples: —— ？ ， ！ … 、

4.28 SB

Closed set: 被, 给.

4.29 SP

Closed set: there are 6 SPs in our corpus.

SPs: 了, 的, 呢, 吧, 呀, 吗.

4.30 VA

There are about 350 VAs in our corpus.

Ex: 便宜[inexpensive], 不错[not bad], 方便[convenient].

4.31 VC

Closed set: 是[be], 为[be], 非[not be].

4.32 VE

Closed set: 有[have], 没[not have], 没有[not have], 无[not have].

4.33 VV

VVs:

- AD-like VV: 是否[whether or not]/VV.
- Phrase-word: 在座[be present]/VV, 报以[respond with]/VV, 为期[scheduled for a duration of time]/VV, 处于[be in a certain condition]/VV.
- Words such as 这样[this way] and 那样[that way] are tagged as VVs when they are followed by AS, or when there is no other verb in that clause. For example, 就[then] 这样[do...this way]/VV 吧[SP] ⟨Let's do it this way⟩.

Chapter 5

Common Collocations

5.1 Length, width, etc.

There are three patterns:

Pattern 1 : NP0 CD M X (non-comparative):
Ex: 他[he] 二[two] 米[meter] 高[tall] ⟨He is 2-meter tall⟩.

Pattern 2 : NP0 X CD M (non-comparative):
Ex: 这[this] 条[M] 船[boat] 长[length] 十[ten] 米[meter] ⟨The boat is 10-meter long⟩.

Pattern 3 : NP0 X CD M (comparative):
Ex: 他[he] 高[tall] 二[two] 寸[inch] ⟨He is 2 inches taller⟩.

X is tagged as VA in all three patterns.

Appendix A

Summary of the Treebank Part-of-Speech Tagset

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	ＩＳＯ
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 A.1: Our POS tagset in alphabetical order

Appendix B

Comparison with Other Tagsets

Table B.1-B.5 illustrate the similarities and differences between our tagset and the ones used by Rocling [Chi95], Peking University [YZWZ98], and the English Penn Treebank [San90].

	Our tag	Rocling tag
total tags	33	46
nouns	3	5
temporal noun	NT	Nd
verbal noun	NN	V[+nom]
proper noun	NR	Nb
other noun	NN	Na, Nc, Ncd
localizer	1(LC)	1(Ng)
pronouns	1(PN)	1(Nh)
verbs	4	17
modals	VV	a subclass of D
shi4	VC	SHI
you3	VE, VV	V-2
other verbs	VV, VA	VA - VL
adverbs	1(AD)	5(D, Da, Dfa, Dfb, Dk)
prepositions	1(P)	1(P)
DP-related	4	6
determiner	DT	Nes, Nep
number	CD, OD	Neu
measure word	M	Nf
conjunctions	2	4
coord. conj	CC	Caa
subord. conj	CS	Cbb

Table B.1: Comparison between ours and Rocling's tagset

	Our tag	Rocling tag
particles	8	3
aspect marker	AS	Di
的	DEC, DEG, AS, SP	DE
地	DEV	DE
得	DER	DE
sent-final part.	SP	T
等	ETC	Cab
other particles	MSP	-
others	8	3
interjection	IJ	I
sound word	ON	??
punctuation	PU	??
noun-modifier	JJ	A
foreign words	FW	FW
被	LB, SB	P
把	BA	P

Table B.2: Comparison between ours and Rocling's tagset(ctd)

	Our tag	PKU's tag
total tags	33	26
noun	3	3
temporal noun	NT	t
verbal noun	NN	V[+nom]
proper noun	NR	n
other noun	NN	n, s
localizer	1(LC)	1(f)
pronoun	1(PN)	1(r)
verb	4	3
shi4	VC	v
you3	VE, VV	v
other verb	VV, VA	v, a, z
adverb	1(AD)	1(d)
preposition	1 (P)	1(p)
DP-related	4	2
determiner	DT	r
number	CD, OD	m
measure word	M	q
conjunctions	2	1
coord. conj	CC	c
subord. conj	CS	c

Table B.3: Comparison between ours and PKU's tagset

	Our tag	PKU's tag
particles	8	2
aspect marker	AS	u
的	DEC, DEG, AS, SP	u
地	DEV	u
得	DER	u
sent-final part.	SP	y
等	ETC	??
other particles	MSP	u
others	8	4
interjection	IJ	e
sound word	ON	o
punctuation	PU	w
noun-modifier	JJ	b
foreign words	FW	??
被	LB, SB	P
把	BA	P
tags for non-words	0	7

Table B.4: Comparison between ours and PKU's tagset(ctd)

	Our tag	Penn Treebank tag
total tags	33	36
nouns	3	4
temporal noun	NT	NN, NNS
verbal noun	NN	NN, NNS
proper noun	NR	NNP, NNPS
other noun	NN	NN, NNS
localizer	1(LC)	0
pronouns	1(PN)	4(PRP, PRP, WP, WP)
verbs	4	7
modals	VV	MD
other verbs	VV, VA, VC, VE	VB, VBD, VBG, VBN, VBP, VBZ
adverbs	1(AD)	4(RB, RBR, RBS, WRB)
prepositions	1(P)	1*(IN)
DP-related	4	4
determiner	DT	DT, WDT, PDT
number	CD, OD	CD
measure word	M	-
conjunctions	2	2*
coord. conj	CC	CC
subord. conj	CS	IN
particles	8	0
others	8	11
interjection	IJ	UH
sound word	ON	-
punctuation	PU	-
noun-modifier	JJ	JJ, JJR, JJS
foreign words	FW	FW
被	LB, SB	-
把	BA	-
misc tags	0	RP, SYM, TO, EX, POS, LS

Table B.5: Comparison between ours and the English Penn Treebank tagset

Appendix C

Dash Tags

The dash-tags are optional. They provide additional information about how the word is formed. Currently we use only one dash tag for words.

- -SHORT: the word is the short form of some words.
  Ex: 深/NR-SHORT (深 is a short form of 深圳/NR).

Bibliography

[Chi95] Chinese Knowledge Information Processing Group. An Introduction to the Academia Sinica Balanced Corpus (in Chinese). Technical Report 95-02, Taipei: Academia Sinica, 1995.

[Gon97] Qianyan Gong. *Zhongguo Yufaxue Shi (The History of Chinese Syntax). Yuwen* Press, 1997.

[San90] Beatrice Santorini. Part-of-Speech Tagging Guidelines for the Penn Treebank Project. Technical report, Dept of Computer and Information Science, University of Pennsylvania, 1990.

[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.

[YZWZ98] Shiwen Yu, Xuefeng Zhu, Hui Wang, and Yunyun Zhang. *The Grammatical Knowledge-base of Contemporary Chinese — A Complete Specification (in Chinese).* Tsinghua University Press, 1998.