Part-Of-Speech Tagset Guidelines for Chinese Treebank Project(draft)

Fei Xia

December 18, 1998

1 Introduction

The tagset is used in the Chinese Treebank project(CTB). The main criteria for the tagset are syntactic distributions(Note: Words with similar meanings might have different part of speeches. For example, 红 is an adjective, 红色 is a noun. Furthermore, Chinese does not have a rich inflectional system. Therefore, the syntactic distributions are the most important criterion for POS tagging.)

The main changes of this version(V2.0) are:

- We removed some tags: wh-tags(WD, WP, WA), MD, AN, PX, LS.
- We modified the definitions of some tags: LC, DT, CD

Currently, the tagset has 33 tags:

Nouns(3): NT, NR, NN

Localizer(1) : LC

Pronouns(1) : PN

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

Measure word(1): M

Verbs, adjectives(4): VA, VC, VE, VV

Adverbs(1): AD

Prep(4): P, LB, SB, BA

Conj(2): CC, CS

Particles, markers(7): DEC, DEG, DEV, DER, SP, AS, MSP

Interjection(1): IJ

Onomatopoeia(1): ON

Others(4): JJ, PU, FW, X

Among these tags:

- The following tags will be removed from the final product: BA, LB, X,
- The following might be removed from the final product: VC, VE, SB, DER, FW, merge DEC and DEG.
- Though unlikely, we might add tags for the following words: modals(or auxiliary verbs), negators, wh-tags, 状态词, 处所词, and another tag for 的.

Notation:

- 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 means all noun tags (NT/NN/NR), tag V means all verb tags (VA/VV/VC/VE). "Det" means all determiner tags DT/OD/CD or a combination of them.
- "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".

2 List of parts of speech with corresponding tags

2.1 Verbs: VA, VC, VE, VV

Normally, a verb satisfies the following:

1. Verbs(except auxiliary verbs etc.) serve as the predicate of the clause (main clause or embedded clause).
2. Verbs can be negated by 不 or 没.
3. AS can attach to some verbs.
4. 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, it is tagged as N or JJ(according to the tests for N and JJ), not as V.

2.1.1 Predicative adjectives: VA

VA roughly corresponds to adjectives in English and stative verbs in the literature on Chinese grammar(the definition of stative verbs varies from system to system. Some include psych-verbs such as "喜欢/了解/怨恨". In our system, those psych-verbs are tagged as VV)[1]

Our VAs include two types:

- type 1: predicates that have no object and can be modified by 很.
- type 2: predicates derived from type 1 either through reduplication(e.g. 红彤彤) or through the pattern N + A meaning "as A as a N"(e.g. 雪白). This type of VAs don't have objects, but it might not be modified by 很 either because it already has the intensifying meaning built-in.

Note: when a word in set(VA) modifies N without 的, it is tagged as JJ etc, not as VA. When a word in set(VA) has an object, it is tagged as VV, not VA. e.g. These activities 丰富/VV le his life.

[1]Whether the adjective is a subclass of the verb in Chinese is still an open question. We will not get into the debate.

2.1.2 Copular: VC

Only 是(shi4) and 为(wei2) are tagged as VC.

shi4 has two main types of usage: link two NPs/Ss, and as a focus marker (e.g. in cleft-sentences). Currently, in both cases it is tagged as VC.

2.1.3 You3 as the main verb: VE

Only 有, 没{有}, and 无 are tagged as VE when they are the main verbs(including the possessive you3, existential you3, etc.). See section 5.2.5 for details.

2.1.4 Other verbs: VV

This includes the rest of the verbs, such as modals, raising predicates(e.g. 可能), control verbs (e.g. 要, 想), action verbs(e.g. 走), psych-verb(e.g. "like/know/hate"), etc.

2.2 Nouns: NR, NT, NN

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

- Nouns can NOT be modified by adverbs such as 不, 很.
- Many nouns can be modified by det-M structure.

Nouns either function as the head of an NP, or they can modify other nouns directly(i.e. without 的).

In other tagsets, the NT is called 时间词, and 名词 only includes NR and NN.

2.2.1 Proper Nouns: NR

Proper Nouns(NRs) are a subclass of nouns. A 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 not modified by a Det-M. For further information, see the Multilingual Entity Task (MET) Guidelines for Chinese.

The names of the following are NRs:
international 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. Russian), race(e.g. White), title(e.g. Prof. X),
disease, occupation, organ(e.g. lung), instrument(e.g. violin), game(e.g. soccer), flower, etc.

2.2.2 Temporal Noun - NT

Temporal Nouns can be the objects of 在 or 到 or 等到. They can be the argument of some verbs and they can also modify VP/S directly.

Temporal Nouns are either the names of the time(e.g. 1990年, 一月, 汉朝) or formed by X+LC, where X is a PN or N or even DT.

Some examples: 同时, 最后, 何时, 今后

2.2.3 Other Nouns: NN

Includes all other nouns. NNs normally can not modify VPs with or without 地.

2.3 Pronouns: 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.

PNs include personal pronouns(e.g. 我,你), demonstratives when used alone as NPs(e.g. 这, 此), possessive pronouns(e.g. 其), and "anaphora"(e.g. 我自己, 自己).

2.4 Adverbs: AD

The adverb is a big class. The behaviors of adverbs differ a lot. The main function of adverbs is: modify VP/S/numbers. This includes manner adverbs, frequency adverbs, degree adverbs, number-modifiers, etc. Some adverbs also have conjunctive function(they are called conjunctive adverbs).

Ex: 'still, yet': 仍然,
'very': 很,
'most': 最 , 'greatly, enormously': 大大.
'again': 又, 'number-modifier': 整, 约

2.5 Prepositions: LB, SB, BA, P

2.5.1 bei4 in long bei-construction: LB

This only includes 被 and 叫 (in spoken language) when they occur in the long bei-construction, i.e. NP0 + LB + NP1 + VP.

Note: 叫 is tagged as VV when it is used as a telescopic verb.

2.5.2 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.

Note: 给 has other tags: P and VV.

2.5.3 ba3 in ba-construction: BA

This only includes 把 and 将 when they occur in the ba-construction, i.e. NP0 + BA + NP1 + VP.

Note: 将 has other tags: AD and VV.

2.5.4 Other Preposition: P

Prepositions other than the words mentioned above.

Ex: 从, 对, 给, 靠

2.6 Determiners and numbers: DT, CD, OD

2.6.1 "Determiners": DT

This includes demonstratives(e.g. 这, 那, 该) and words such as "每, 各, 前, 后" etc.

DTs does NOT include cardinal numbers and ordinal numbers.

See section 6.11 for all the DTs.

2.6.2 Cardinal Numbers: CD

It includes cardinal numbers and 概数词 such as 来, 多, 好几, and words such as 好些, 若干, 半, 许多, 很多.

Ex: 1245, 一百

2.6.3 Ordinal Numbers: 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

2.7 Measure words: M

Measure words(or classifiers in some systems) follow determiners to form det-M structure to modify nouns or verbs. This includes 个体量词(e.g. 个), group measure words(e.g. 群), and words such as 公里, 升, etc.

Some 个体量词 can be modified by a limited set of adjectives. 临时量词 can be modified by nouns and adjectives, e.g. 一/CD 铁/NN 箱子/M 书/NN.

2.8 Conjunctions: CC, CS

2.8.1 Coordinating conjunctions: CC

Words that conjoin a construction that has two or more centers, each of which has approximately the same function as the whole construction are tagged as a coordinating conjunction (CC).
The patterns for CCs are: XP {,} CC XP; CC1 XP CC2 XP.

Ex: 与, 和, 或, 或者, 还是, 至, 到, 兑.

2.8.2 Subordinating conjunctions: 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.

Ex: 如果/CS ... 就/AD ... ; 因为/CS ... 所以/AD ...; 由于/CS ...

2.9 Localizers: LC

Many nouns alone can not be the argument of prepositions such as "在, 到". The main function of localizers is to attach to the preceding NP/S so that the whole phrase can act as the argument of those prepositions. Some localizers can stand alone as the arguments of the prepositions/verbs.

Localizers are of two types:

- mono-syllabic localizers: e.g. 前, 后, 里, 外, 内
- bisyllabic localizers: they are formed by
  - mono-syllabic localizers plus morphemes such as 边, 面, 头, and 以, 之 etc. e.g. 之间
  - two mono-syllabic localizers: e.g. 前后, 左右, 上下
  - a couple of other bisyllabic localizers: 周围, 期间, 附近

2.10 Markers: DEC, DEG, DEV, DER, SP, AS, MSP

2.10.1 DEC: de5 as a complementizer or a nominalizer

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

Note: 的 also has other tags: DEG, SP, and AS.

2.10.2 DEG: de5 as genitive marker and associative marker

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 also 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 (He run 得 very fast.)

Note: Some collocations ending with 得 are not V-de constructions. They are verbs. e.g. 记得, 获得.

2.10.4 Manner de5: DEV

This only includes 地 when it occurs in "XP 地 VP", where XP describes the manner of the VP. In some old literature, 的 is used instead. In that case, we will tag that 的 as DEV.

Ex: 高兴/VA 地/DEV 说/VV

2.10.5 Sentence-final particles: SP

This includes 了,呢,吧,啊,呀,吗 and negation words(不 不是 没 没有) in VP-not etc.

2.10.6 Aspect Particle: AS

Verbal particles that indicate aspect are tagged as aspect particles (AS). This category ONLY includes:
了(le5)-perfective aspect,
着(zhe5)-durative aspect,
过(guo4)-indefinte past aspect
的(de5)- past tense in cleft-sentence.

2.10.7 Other particles: MSP

This includes all other particles: e.g. 所, 来,以, 去, 而.

See section **??** for the complete list of MSPs.

2.11 Interjection: IJ

Interjections appear in the sentence-initial position, i.e. "IJ, S". It's a closed set.

Ex: 啊

2.12 Onomatopoeia: ON

ON is used to describe sounds. It is often followed by DEV to modify VPs or occur in the pattern "ON 一声"

Ex: 哗啦啦, 咯吱

2.13 Others: JJ, PU, FW, X

2.13.1 JJ

JJs include the following three types:

- type 1: 区别词(非谓形容词): Those words modify nouns in the pattern JJ+的+{N} or JJ+N, but they can not be the predicate of a sentence without the help of 的.

  The patterns: JJ + 的/DEG {+ N}, JJ+N.

  Ex: 共同/JJ {的/DEG} 目标/NN, 她 是 女/JJ 的/DEG.

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

  The pattern: JJ+N.

  Ex: 留美/JJ scholar, 随军/JJ 妓女/NN,

- type 3: adjectives: 新/JJ 消息/NN.

  The pattern: JJ+N

  Ex: 新/JJ 消息/NN.

  Note: when 的 is inserted between the adjective and the noun, the adjective is tagged as VA.

2.13.2 Punctuations: PU

Punctuations as words are tagged as PU. If they are inside words, they are not tagged.

Ex: Mary ,/PU John and Mike; 123,456/CD

2.13.3 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/NN, A型/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 in an extreme case such as citing a text in another language.

2.13.4 Unknown: X

This is used only when the annotator has no clues what the tag should be. If the annotator has some candidates, but he is not sure which is the right one, he should give a list of tags by the order of preference(e.g. tag1 | tag2).

3 Dash-tags and tags for non-words

3.1 Dash-tags

The dash-tags are optional. They provide additional information about how the word is formed.

- -short: the word is the short form of some string.
- -phr: the string is a phrase in non-modern Chinese, the phrasal rules of the string are no longer productive. We don't analyze a phrase if the phrase is totally fixed and doesn't comply with the rules of modern Chinese.
- -fw: the word is a foreign word. e.g. 我 喜欢 John/NR-fw. FW by definition is -fw.
- -p1: the first part of a word. e.g. 进出口(conjunction reduction), 喜不喜欢(A-not-A), potential form(V-neg-R).
- -p2: the second part of a word.

3.2 Tags for non-words

These tags are needed only when we tag the internal structure of words.

- AFF: affix
- MORP: morpheme
- NMOR: non-morpheme

4 List of tags with corresponding part of speech

Table 1: Our POS tagset in alphabetical order

AD	adverbs	还
AS	aspect marker	着
BA	把 in ba-const	把, 将
CC	coordinating conj	和
CD	cardinal numbers	13
CS	subordinating conj	虽然
DEC	的 for relative-clause etc.	的
DEG	associative 的	的
DER	得 in V-de const. and V-de-R	得
DEV	地 before VP	地
DT	determiner	这
FW	foreign words	el, al
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 particles	所
NN	common nouns	书
NR	proper nouns	美国
NT	temporal nouns	今天
OD	ordinal numbers	第一
ON	onomatopoeia	哈哈 哗哗
P	prepositions excl.被 and 把	从
PN	pronouns	他
PU	punctuation	
SB	被 in short bei-const	被 给
SP	sentence-final particle	吗
VA	pred adjective	红
VC	是	是
VE	有 as the main verb	有
VV	other verbs	走
X	undecided	some 了

5 Problematic cases

5.1 Confusing parts of speech

5.1.1 AD or AS

zai4(在) before the verb is treated as AD, not as AS because other adverbs can intervene between zai4 and the verb.

5.1.2 AD or CC

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

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

The only words with both tags are: 又.

See section 6.4 for a complete list of CCs.

5.1.3 AD or CS

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

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

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

See section 6.6 for the complete list of CSs.

5.1.4 AD or JJ

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

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

Also, an easy test to tell them apart is to insert 的 between the word and the NP. If the new phrase is still valid, the word is a JJ, otherwise, it is an AD.

The following are tagged as JJs when they modify nouns: 上述: 上述/JJ 三/CD 国/NN

5.1.5 AD or NN

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

Temporal nouns and "location" nouns[2] can modify VP/S directly. In those cases, they are still tagged as nouns. For other nouns, if we can insert the preposition "在" before the word without changing the meaning or validity of the sentence, tag the word as NN, otherwise, tag it as AD.
Ex: 政治/AD 解决/VV (#196), 重点/AD 抓/VV ...

5.1.6 AD or NT

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

If the word X can be the head of an NP, it is tagged as NT. e.g. 昨天 is a NT, not an AD.

ADs: 早日/AD 实现/VV,

NTs: 目前/NN (#57),

5.1.7 AD or VA

The patterns are: X+VP, X+地 VP.

We assume that a VA can not modify VP directly, It needs the help of a DEV. We also assume that both VA and AD can occur in X+DEV+VP. As in the pattern:

- 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 context.
- w VP: tag it as AD.

Ex: 高兴/VA 地/DEV, 紧密/VA 地/DEV, 大大/AD 提高,

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

5.1.8 AD or VV

The pattern is: X+VP/S.

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

We assume that a verb can not be modified by another verb directly.

VV-like ADs: 联合/AD 组成/VV (#50),
"Modal"-like ADs: 大概, 将, 一定.

5.1.9 AS or VA/VV

In the pattern: V+X.

Currently, we have only four AS's. Some morphemes like R in V-R have similar functions to AS's. e.g. 来(in 拿来), 完(in 写完), 起来(in 跑起来). We treat them as V for the time being.

5.1.10 CC or P

Some CCs(e.g. 与 和 跟 同) 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 must be eliminated), then X is a CC.

5.1.11 CS or P

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

There is no overlap between set(CS) and set(P) WHEN the argument is a S.

Ex: 由于/P 美国/NR 的/DEC 压力/NN, 由于/CS 美国/NR 施加/VV 压力/NN

5.1.12 DT or JJ

The pattern is: X+N.

There is no overlap between set(DT) and set(JJ). See the section 6.11 for the complete list of DTs.

DTs are either demonstratives or quantifiers. The quantifiers may have the scoping effect.

5.1.13 DT or PN

In the patterns: X as a NP, X+NP, X+的+NP.

Only the demonstratives are in both set(PN) and set(DT), they are tagged as DTs when they modify a noun(i.e. X+NP), otherwise, tag them as PNs(i.e. in X).

- Demonstratives may occur in all three patterns. They are tagged as PN in "X" and "X+的+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 might occur in "X+的+NP" and 'X+NP', In both cases, they are tagged as DTs.

Ex:

- 这/DT book is very good.
- 这/PN is what I wanted.

- 他/PN father,
- 所有/DT 的/DEG 东西/NN.

5.1.14 JJ or NN

JJ can not be the head of a NP, NN can.

Some common JJs: N+性: 全国性
Some JJ-like NNs: 国际,

5.1.15 JJ or P

The following words have both tags:

. 有关: 有关/P 撤军/NN 的/DEG 报告/NN, 有关/JJ 单位.

5.1.16 LC or NN

There are no overlap between LCs and NNs. See the section 6.15 for the complete list of LCs. The main function of the LC is to attach to a NP/S, so the whole part can be the argument of prepositions.

5.1.17 M or NN

The pattern is: CD+X.

If no M can fill the position in between, tag X as M, otherwise tag it as NN. 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/CD {个} 学生/NN, 一/CD 年/M, 一/CD 箱子/M 书/N.

5.1.18 NN or NR

NRs are the names of persons, organizations, countries, etc. They normally can not be modified by Det-M.

5.1.19 NN or NT

The NP headed by an NT can modify S/VP directly and can answer the question "at what time".

5.1.20 NN or VA

If the word can be modified by 很 in THAT context, tag it as VA.
Ex: 表示/VV 乐观/VA.

5.1.21 NN or VV

If a word X in set(VV) occurs in head(NP) position, it is tagged as NN.

If X in that context can be modified by Det-M, tag X as N. If X can be modified by YP+地 or words that can only modify verbs. tag X as V.

The head of the object of the following verbs is a N, not a verb: 进行(e.g. 进行/VV 密切/JJ 合作/NN),

Some verb can take either a VP or a NP as its object, choose the ones that corresponding to the preferred reading.

5.1.22 P or VV

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

NP0 X NP1 YP, X is Prep or V:

- if there is no other verb in a complete sentence, X is a verb.
  e.g. He 在/V home.
- if the NP1 is absent or moved (as in short answer or VP-not-V question), then X is a verb.
  (because Chinese does not allow prep stranding).
  e.g. He 在/V home bu4 在/V ?

  Note: P can occur in A-not-A VP. e.g. He 从 没 从 Beijing 出发 .
  P-not-P might be a compound prep, so P-not-P is not an example of preposition stranding.
- if AS can follow X, then X is a verb.
  Note: some prepositions end with 着 or 了, so be careful.
- the meaning of prep/V might be different, e.g. 靠 ('close to'/P, "depend on"/VV).
- If it is still ambiguous after applying the tests above, tag the word as P.

VV-like Ps: 作为 in #39,

5.1.23 VA or VV

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

Some words have both VA and VV tags, such as feng1fu4.

Note: if the object of the word is preposed as a P+NP and the NP can be moved back to the postverbal position,, tag the word as VV, e.g. 他对我很关心/VV.

5.2 Specific words and collocations

This section will be expanded a lot during the project.

5.2.1 deng3, deng3deng3

deng3(等):
XP deng3 NP: tag it as CC. (#79)
XP deng3: tag it as AD.

deng3deng3(等等):
XP deng3deng3: tag it as AD.

5.2.2 de5

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

5.2.3 lai2

lai2(来):

- as main verb("VV"): He 来/VV le.
- as R in V-R construction, or DIR in V-DIR("VV"): He 拿来/VV a book.

  He [走/VV 上来/VV]/VV.
- sentence-ending particle(SP): He 拿出 book 来/SP.
- adjoin to VP, similar to 以("MSP"):
  He 用/P this 来/MSP prove his innocence.
- between NUM and CL("CD"): 30来/CD 个
- LC: after a time string, as a localizer, similar to 以来("LC"): 3 年 来/LC

5.2.4 lian2

lian2(连) in the following examples are tagged as AD(conjunctive adverb):
He lian2 Mary dou1 not know.
He lian2 gen1 his wife dou1 lie.
He lian2 cry dou1 cry bu4 cheng2.

5.2.5 you3

you3(有) is tagged as VE when it is the main verb.

mei2-you3(没有) is tagged as VE when it is the main verb. It is tagged as VV when it modifies VP. It is tagged as SP in VP-mei2you3.

mei2 alone is tagged as VE when it is the main verb, It is tagged as AD when it modifies VP and when it occurs in A-not-A. It is tagged as SP in VP-mei2.

you3-mei2-you3 is tagged as "[you3/VV mei2you3/VV]/VV" when it modifies a VP, and as "[you3/VE mei2you3/VE]/VE" when it is the main verb.

5.2.6 zhe4yang4

zhe4yang4(这样):
NN(?): 这样/NN 的/DEG 伙伴/NN 关系/NN. (#171).
AD: He 这样/AD 做/VV ....
VV: 你 别 这样/VV.

6 Lists of words for each part of speech tag

If the set for a POS tag is closed, we will list all such words which appear in our 100K data. For open-set tags, we will 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". The number enclosed in the paretheses is the number of the sentence where the word occurs.

6.1 AD

The followings are ADs:

- Conjunctive adverbs:

  otherwise: 否则 (#175),
  therefore: 所以, 因此, 因而
  however: 却,
  then, as a result: 那么(#176), 就, 便, 结果, 则, 这样(#153)
  in addition: 另外, 此外
  furthermore: 进而
  later: 随后, 然后
  as well: 也
  so that: 以便(#139), 从而
  "for example": 例如/AD(#157), 如/AD(#161)
  "that is": 即

- NT-like adverbs: 届时(#176)

- frequency adverbs: 多次(#180, 多次/AD 发生/VV), 一下, 一下子

- manner adverbs: 互利(#186, 互利/AD 合作/VV),

- Modal-like adverbs: 将

- VA-like adverbs: 更好(#139, 更好/AD 地/DEV 造福/VV),

- ADs that modify numbers:

  - AD + number: e.g. 近/AD, 不足(#153), 上(上/AD 千/CD),

  - NUM + M + AD: 整, 多,
    Note: 3点整 is tagged 3点/NT 整/AD.

- ADs that can precede NPs:
  又(又/AD 一/CD 个/M 参与者/NN),
  不到: 5 minutes 不到/AD, 不到/AD 5 minutes.

- Negative ADs: 未(also VE), 不, 没(also VE)

- phrase-word: 进一步(#82), 越来越(#189), 尤其是(#155),

- other: 一起/AD(#144), 等(#146, also CC), 等等(#157), 另/AD, 正在/AD,
  凡(#176), 才

6.2 AS

closed set: 了, 着, 过, 的

6.3 BA

closed set: 把, 将

6.4 CC

closed set: the list is not complete yet.

'and': 与, 和, 跟, 同, 及, 以及, 又, 并, 而(大/VA 而/CC 全/VA), 并且, 兼, 而且

'or': 或, 或者, 还是(also AD),

others: 至, 到, 兑, 兼

paired-CCs: 既/CC .. 又/CC(#186), 无论/CC ... 还是/CC, 不管/CC ..还是/CC.

others: 等(also AD), 兼(also VV)

6.5 CD

It includes cardinal numbers and quantitative quantifiers such as 几, 许多/DT(#180), 若干(#188), 数(#134), 大部分(#150), 部分/CD, 一些/CD, 大批/CD, 好几/CD.

6.6 CS

closed set: the list is not complete yet.

Some examples: 因为, 如果, 由于, 即使, 不管, 尽管, 虽然, 只要, 只有, 一旦,

. 别看: 别看/CS 小海龟在沙滩上爬行往往显得笨拙, 一旦/CS 到了海里，就/AD 变得灵活自如。/PU

6.7 DEC

closed set: 的, 之

6.8 DEG

closed set: 的, 之

6.9 DER

closed set: 得

6.10 DEV

closed set: 地

6.11 DT

closed set: the list is not complete yet.

- Demonstrative determiners:
  这, 此, 该, 本, 那, 上, 下, 同, 前, 另, 其余, 个别(#157): 个别/DT 发达/JJ 国家/NN, 某, 头(头/DT 7/CD 个/M 月/NN), 这些
- Quantifiers(excl. quantitative quantifiers): "every, all, any" etc:
  . 各, 每, 何, 所有(#158),
  . 整: 整/DT 个/M 欧洲,
  . 全: 全/DT 省/NN;
  . 全体: 全体/DT 外交/NN 官员/NN;
  . 同: 同/DT 一/CD 个/M 原因/NN);
  . 一切: 一切/DT 努力/NN,

. 有的: 有的 书,

6.12 IJ

closed set: the list is not complete yet.

Some examples: 啊, 嘿

6.13 JJ

Some examples: 共同, 双边, 很大, 高科技, 有关(有关/JJ 国际/NN 公约/NN), 老牌/JJ(老牌/JJ 军工/JJ 企业/NN),
. 上述: 上述¡***¿/AD 三/CD 国/NR

6.14 LB

closed set: 被, 叫

6.15 LC

closed set: the set is not complete yet.

monosyllabic LCs: 初: 七十年代/NT 初/LC
bisyllabic LCs: 之间, 在内, 以来, 以后(also a NT), 期间, 左右

6.16 MSP

- LC-like MSP:
  - 来说: 对/P ... 来说/MSP;
  - 的话: 如果/CS ... 的话/MSP;
  - 开始: 从/P 十二月/NT 九日/NT 开始/MSP,
  - 起: 从/P 十二月/NT 起/MSP
  - 般: 雷鸣/NN 般/MSP 掌声/NN
  - 为止/MSP:

- conj-like MSP: connect a PP/VP and a VP.
  - 以: 以/MSP 增强/VV 总体/JJ 竞争/NN 实力/NN
  - 而: 为/P 生存/VV 下去/VV 而/MSP 不得不/VV 采取/VV 的/DEC 行动/NN

- the word 所:

6.17 NN

phrase-word: 之一(#188, 目的/NN 之一/NN)

localizer-like NN: 附近/NN(#181), 国内

6.18 NT

Some examples: 1990年, 最后/NT,
Examples of N+LC as a NT: 战后/NT, 赛前/NT, 今后, 日前/NT, 何时/NT
Examples of PN+LC as a NT: 此后/NT,

6.19 ON

Some examples: 刷, 哗啦啦

6.20 P

closed set: the list is not complete.

Some examples:

VV-like prep: 经过/P(#191), 作为/P(#195), 截止, 有关, 离

CS-like prep: 随着, 沿着, 鉴于, 除了, 为了

AD-like prep: 就/P 机制/NN 问题/NN

6.21 PN

closed set: the list is not complete.

Some examples:

- personal pronoun: 他,
- demonstratives alone as a NP(also DTs): 这, 此, 这里
- possessive pronoun: 其

- "anaphora": 他自己, 自己

6.22 PU

closed set: the list is not complete.

6.23 SB

closed set: 被, 给

6.24 SP

closed set: the list is not complete.

SPs: 了,呢,吧,啊,呀,吗, 不, 不是, 没, 没有

6.25 VA

one word: 半真半假,

6.26 VC

closed set: 是, 为

6.27 VE

closed set: 有, 没, 没有, 无: 无/VE 大/VA 新闻/NN

6.28 VV

VVs:

- MD-like VVs: 要/VV(#188), 愿意/VV(#144),
- phrase-word: 在座/VV(#134), 报以/VV(#134), 为期(#163),处于/VV (ART182 #3)

7 Common Collocations

7.1 Linking elements

We list the POS of the common linking-element pairs:

- CC ... CC:
  - 既/CC .. 又/CC, 不管/AD ..还是/CC,
- CS ... AD:
  - 一旦/CS ... 就/AD ...

7.2 Others

- CD: 几
- DT: 这些(also PN)
- NN: 其中, 国内
- NT: 同时, 最后, 何时, 这时, 今后

8 Comparison with Other Tagsets

Table 2-6 illustrate the similarities and differences between our tagset and the ones used in Rocling, Peking University, and the English Treebank project.

Table 2: Comparison between ours and Rocling's tagset

	Our tag	Rocling tag
total tags	33	46
nouns	3	6
temporal noun	NT	Nd
verbal noun	NN	V[+nom]
proper noun	NR	Nb
other noun	NN	Na, Nc, Ncdb
localizer	1(LC)	1(Ncda)
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, Dfa, Dfb)
prepositions	4	1
被	LB, SB	P
把	BA	P
other prep	P	P
DP-related	4	6
deteminer	DT	Nes, Nep
number	CD, OD	Neu(ND)
measure word	M	Nf
conjunctions	2	4
coord. conj	CC	Caa
other conj	CS	Cbb

Table 3: Comparison between ours and Rocling's tagset(ctd)

	Our tag	Rocling tag
markers	7	3
aspect marker	AS	Di
的	DEC, DEG, AS, SP	DE
地	DEV	DE
得	DER	DE
sent-final part.	SP	T
other particles	MSP	-
others	6	4
foreign words	FW	FW
interjection	IJ	I
noun-modifier	JJ	A
sound word	ON	??
punctuation	PU	PUNCT?
undecided	X	??

Table 4: Comparison between ours and PKU's tagset

	Our tag	PKU's tag
total tags	33	18
nouns	3	3
temporal noun	NT	t
verbal noun	NN	V[+nom]
proper noun	NR	n
other noun	NN	n, s
localizer	LC	f
pronouns	1(PN)	1(r)
verbs	4	3
shi4	VC,	v
you3	VE, VV	v
other verbs	VV, VA	v, a, z
adverbs	1(AD)	1(d)
prepositions	4	1
被	LB, SB	p
把	BA	p
other prep	P	p
DP-related	4	2
deteminer	DT	r
number	CD, OD	m
measure word	M	q
conjunctions	2	1
coord. conj	CC	c
subord. conj	CS	c

Table 5: Comparison between ours and PKU's tagset(ctd)

	Our tag	PKU's tag
markers	7	3
aspect marker	AS	u
的	DEC, DEG, AS, SP	u
地	DEV	u
得	DER	u
sent-final part.	SP	y
other particles	MSP	u
others	6	3
foreign words	FW	?
interjection	IJ	e
noun-modifier	JJ	b
sound word	ON	o
punctuation	PU	?
undecided	X	?

Table 6: Comparison between ours and Penn Treebank tagset

	Our tag	Penn Treebank tagset
total tags	33	36
nouns	3	4
temporal noun	NT	NN, NNS
verbal noun	NN	NN, NNS
proper noun	NR	NP, NPS
other noun	NN	NN, NNS
localizer	1(LC)	0
pronouns	1(PN)	4(PP, PP, 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	4	1*
prep	LB, SB, BA, P	IN
DP-related	4	4
deteminer	DT	DT, WDT, PDT
number	CD, OD	CD
measure word	M	-
conjunctions	2	2*
coord. conj	CC	CC
subord. conj	CS	IN
particles	7	0
others	6	11
foreign words	FW	FW
interjection	IJ	UH
noun-modifier	JJ	JJ, JJR, JJS
sound word	ON	-
punctuation	PU	-
listed item	-	LS
misc tags	X	RP, SYM, TO, EX, POS