THE CKIP CHINESE TREEBANK: GUIDELINES FOR ANNOTATION

KEH-JIANN CHEN[*], CHI-CHING LUO[**], ZHAO-MING GAO[***], MING-CHUNG CHANG[****], FENG-YI CHEN[*****], CHAO-JAN CHEN[******], CHU-REN HUANG[*******]

Abstract. This paper aims to present the methodology and guidelines for annotation in CKIP Chinese Treebank. Under the framework of the Information-based Case grammar (ICG), a lexical feature-based grammar formalism, which stipulates each lexical item containing both syntactic and semantic information, the potential phrasal heads of input are located and the semantic relations between words are also identified. Thus, not only phrasal categories but also thematic roles are both annotated. Incorporating with Head-Driven Principle, some guidelines are also implemented for more consistent annotation in such grammatical phenomenon as the constructions of coordinates, topicalization, and the construction with nominal predicate. In addition, we tag the CKIP Treebank with semantic categories to extract useful collocation among semantic classes of the bracketed constitutes, which is also supposed to further enhance the performance of our parsing model.

1. INTRODUCTION

After establishing a 5-million-word Sinica Corpus with part-of-speech tags, the CKIP Chinese Treebank project is now bracketing and annotating the Sinica corpus by human post-editing the computer parsing results of this corpus. This paper aims to present the methodology and guidelines for annotating CKIP Chinese Treebank. To maintain a better quality for the annotation, the following issues have to be preplanned before massive production. First, the representational issue, in order to make designated implicit linguistic information explicit by annotations, we may want to know what kind of tags and how many different types of tags have to be annotated and in what way. In section 2, our tree representational models are proposed, which have nice features of both dependency grammar and conventional phrase structure grammar. The second issue is how to maintain the consistency and error free of the annotations without jeopardizing the efficiency of the annotation process. In section 3, the methodology to construct a tree bank and the steps to build up the Treebank are presented. In section 4, some guidelines for problematic construction are also

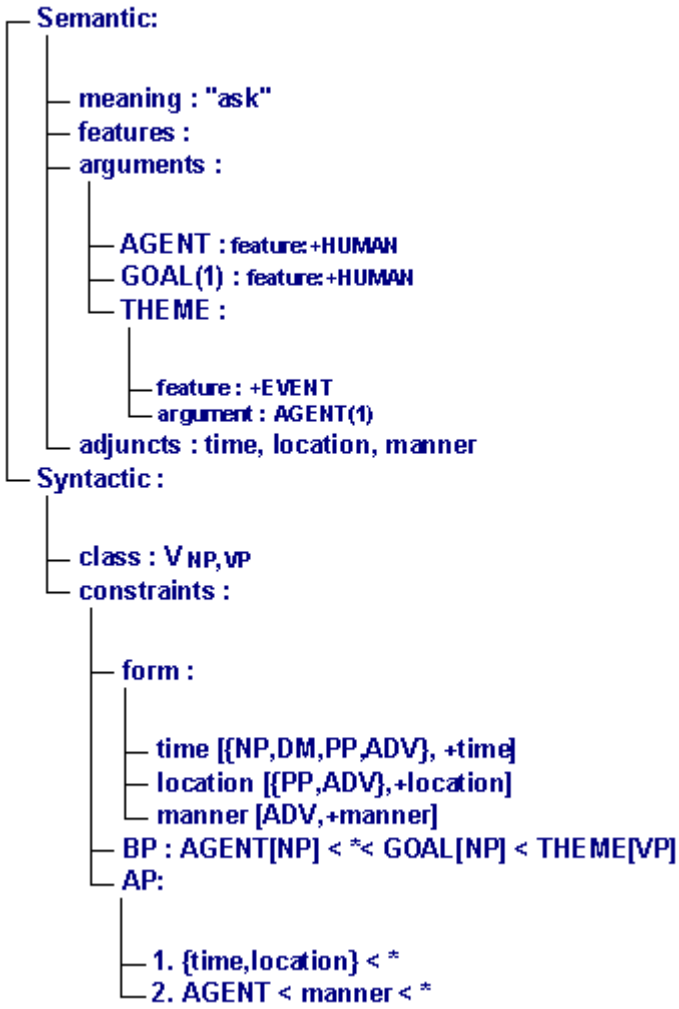
presented. In section 5, conclusion and the knowledge extraction from Chinese Treebank and its application are discussed.

2. REPRESENTATIONAL MODEL

In order to have a rich and simple representation model for sentence structures, we adopt the grammar representation in the Information-based Case Grammar (ICG), a lexical feature-based grammar formalism (Chen 1996). The ICG representation unifies the thematic structures, phrasal structures and head-daughter dependency relations into a single uniform representational structure. In addition to the structure bracketing, this representation retains the syntactic information of phrasal structure and registers the thematic role for each constituent. It adopts the head-daughter dependency relational structure, which makes shallower structure. Thus the grammar representation has less ambiguous structures. In addition, an ICG-based parser is available to produce consistent parsing trees before human post-editing.

It is very important and difficult to identify word relations in parsing Chinese. To express the complex constraint relations between words, each lexical item is encoded both syntactic and semantic information in ICG. Thus, for a phrasal head, the syntactic and semantic constraints of the phrase are expressed by the formation rules and liner ordering rules for the thematic roles, which is illustrated in the following example (1). The grammatical information is simplified for easy of illustration.

(1) *jiao* "ask"



The above grammatical representation matches the sentences taking *jiao* 'ask' as the head verb. For instance, it matches the sentence (2a) and the matched tree structure is as in (2b).

(2a) *Ta jiao Li-si jian chio.*
He ask Lisi pick ball.
"He asked Lisi to pick up the ball."

(2b) S (agent: NP(Head: Nhaa: *Ta* 'He') |Head: $V_{NP,VP}$: *jiao*'ask' |goal: NP(Head: Nba: *Li-si)* |
theme: VP (Head: VC2: *jian* 'pick' | goal: NP (Head: Nab: *chio*'ball')))

The above representation of the tree structure has the advantages of maintaining both the phrase structure rules and the syntactic and semantic dependency relations. The matching processes are achieved by a head-driven parser. It begins with the

identification of the potential phrasal heads of input, which is guided by phrasal patterns registered in the heads. Once the heads are located, the syntactic and semantic restrictions between words are also identified. The parsing and identification of thematic roles are done simultaneously (Chen, 1996).

3. STEPS TO BUILD UP A TREE BANK

In order to keep the consistency of the annotation as well as the production efficiency, we develop our tree-bank by first parsing the pos-tagged corpus by computers and then post-editing the parsing results via a tree-editing tool by human annotators. In the following, we will present steps to build up our Chinese Treebank. Starting from the material we use, the second step begins with the parsing process to bracket and to assign the labels of phrasal categories and thematic roles. To be more consistent and more efficient, we set up some guidelines and derive a tool for manual post-editing in the final step.

3.1. Material

The material we use is the 5-million-word Sinica Corpus with part-of-speech tags. The tagged Chinese corpus, developed by CKIP, is balanced in various topics, and each text is marked according to five criteria, namely, genre, style, mode, topic and source (Chen, Huang et al, 1996). The ongoing effort to construct tree structures has completed 29,331 sentences. And the adopted texts consist of such topic as business, politics, travel and sports.

3.2. Bracketing and Annotating

For a string of input, the parsing process begins with the word identification and the initialization of lexical information from a built-up lexicon encoding syntactic and semantic information in each lexical item. Next, the head-driven parser begins to identify the potential heads of the input (Chen, 1996). Therefore, in addition to the bracketing, not only syntactic categories but also thematic roles are annotated. The sentence in example (3a) is parsed to be (3b).

(3a) *tamen da che dao Wu Lai kao rou*
they take vehicle to Wu Lai roast meat
"They go to Wu Lai by bus to have a BBQ."

(3b) S(agent: NP(Head: Naeb: tamen ‘they’) | Head:VC2: da ‘take’ | goal: NP (Head: Nab: che ‘vehicle’) | complement: VP(Head: VC1: dao ’to’ | goal: NP(Head: Nca: Wulai ‘Wulai’) | complement: VP(Head: VC2: kao’ roast’ | goal: NP(Head: Naa: rou ’ meat’))))

The conditions for assigning the labels of phrasal categories are summarized in the following. As for the complete lists of syntactic categories and thematic roles in CKIP Treebank, please see Appendix 1 and 2.

(4) List of Phrasal Categories

S, V, VP, N, NP, DM (Determiner Measure phrase), GP (Postpositional phrase, or localizer phrases), PP, A • {de, di, zhi}, N • {de, di, zhi}, V • {de, di, zhi}, S • {de, di, zhi}, DM • {*de, di, zhi*}, GP • {d*e, di, zhi*}, NP • {*de, di, zhi*}, PP • {*de, di, zhi*}, VP • {*de, di, zhi*}, ADV • {*de, di, zhi*}, *dao* • {NP, VP, S}, *ge* • VP, *de2* • {NP, VP, S}

- Only phrases that have a thematic role are labeled XP.
- There are no empty categories in our framework. Thus, verbs taking verbal or sentential objects like juede 'feel' can subcategorize for a VP or a S.
- There is no clear-cut borderline between lexical and phrasal categories in our tree structures. For instance, the node governing two conjoined nouns is labeled N rather than NP. The labeling of an XP is delayed until the assignments of thematic roles are determined.
- The node VP can occur whenever one of three conditions is met.

    a. It is a subcategorized argument.
    b. It is in a clause which lacks a subject NP.
    c. It is conjoined with a VP.

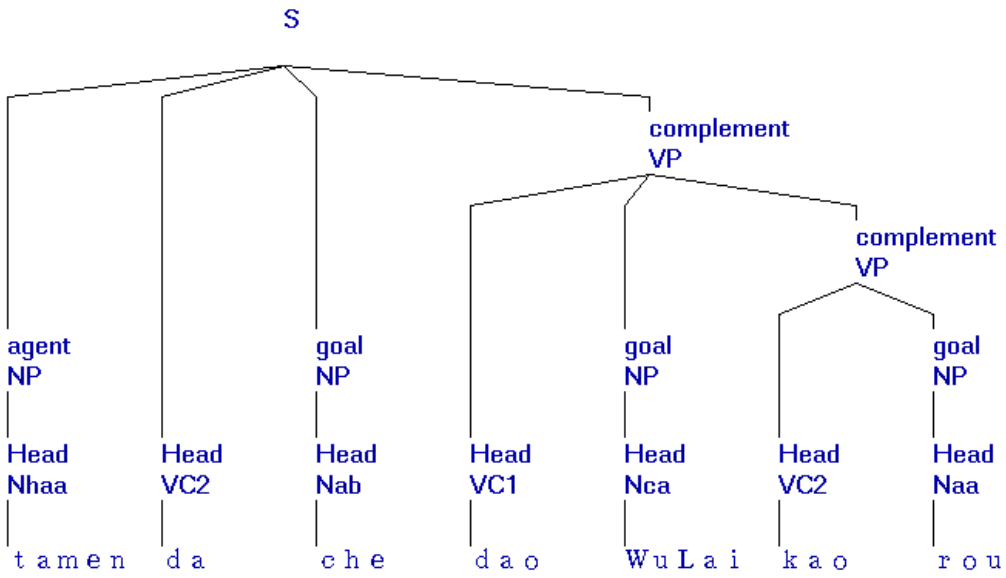
- *Sentential subjects and clausal complements are labeled as S rather then NP.*
- *Under the node GP, localizer is as its head and dummy as its argument.*
- *Like GP, the node PP takes dummy as its argument.*
- *Clitics such as de, di, de2, ge, dao and their preceding phrases can form a phrase whose phrase names are expediently expressed as {de, di, de2, dao, ge} •{NP, VP, S} or {A, N,V, S, DM, GP, NP, PP, VP, ADV} •{de, di, zhi}. Therefore, AP and ADVP are not included.*

3.3. The Manual Post-editing

After automatic bracketing and annotating of input, many unwanted parsed sentences may be generated, so we need manual post-editing to modify them or pick up one of the best possible structures. To be more consistent in manual editing, we thus set up some guidelines to be followed for some special constructions, which will be discussed in section 4. Moreover, to do the post-editing efficiently, we derive an on-line editing tool. Three main functions are included in this tool.

a. To avoid making mistakes and to work more efficiently in editing, the first function of this tool is to change the liner form of parsed sentences into graphic tree structures. Therefore, the liner form of example (3b) may convert into the tree structure as in example (5).
b. It also provides users a more convenient way to modify. They can move or modify the whole sub-trees directly, if needed.
c. The final function, which is under construction, is error correction. The function aims to provide users to do grammar or format checking, and finally do the auto correction with ill-formed constructions.

(5) Tree structure of example (3b)



they take vehicle to Wu Lai roast meat
"They go to Wu Lai by bus to have a BBQ."

4. GUIDELINES FOR ANNOTATING PROBLEMATIC CONSTRUCTION

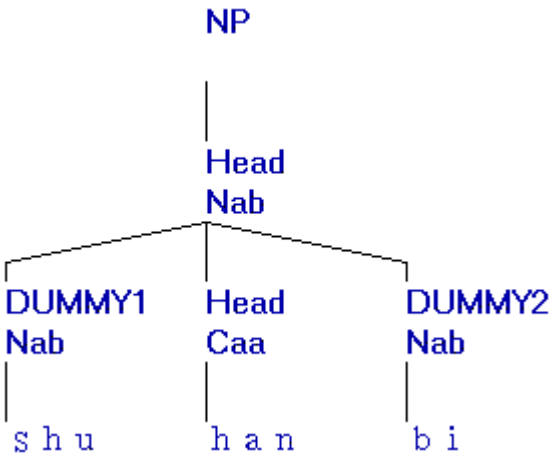
Incorporating with Head-Driven Principle, some guidelines are also implemented for more consistent annotation in such grammatical phenomenon as the constructions of coordinates, topicalization, and the construction with nominal predicate.

4.1. Coordinates

The conjoined nodes in the coordinate structures may be words with the same function and may be not. Thus, the idiosyncrasy between conjoined nodes poses the problem how to annotate the node governing the conjoined words or phrases. According to the following conditions, there are different annotations of the node governing the coordinates.

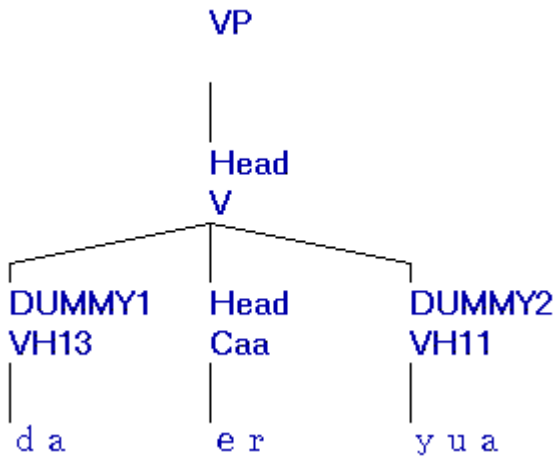
a. When the syntactic categories of conjoined nodes are the same, the label of higher node governing these elements does not change at all. Example is given in (6).
b. When the syntactic categories of conjoined nodes are similar and share with the same major function X, the higher node governing these elements are labeled as an X. Example is given in (7).
c. When the syntactic categories of conjoined nodes are different, the label of higher node varies for the right-hand node. It is illustrated in example (8).

(6)



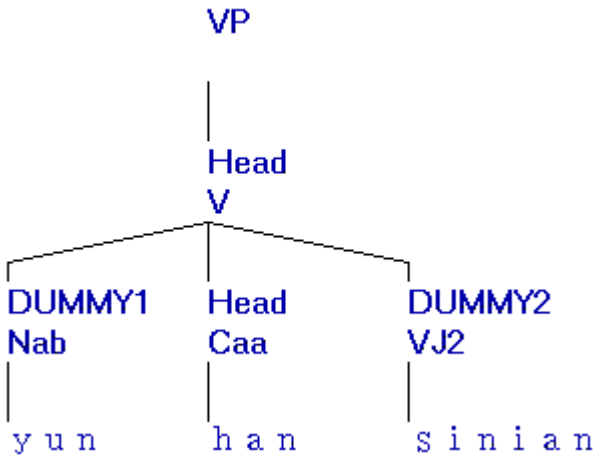
book and pen
'books and pens'

(7)



big and round
'big and round'

(8)



cloud and miss
'cloud and missing'

4.2. Topicalization

Sometimes it is not easy to give a definite role name to those with multiple roles as in topic sentence. The topicalized elements may be the subcategorized argument of one verb and may be not. Therefore, there are two different labels which are coded according to the following conditions.

a. When the topic of one sentence is not the subcategorized argument, it is just coded as *topic*. Example is given in (9)

b. When one subcategorized argument of the head is topicalized, it is annotated as *topic* with additional feature, namely *topic [+thematic role]*. Example is given in (10).

(9) *zhe yi zhung shi wo de jingyan zuei fengfu*
this one kind thing my experience most abundant
'I am highly experienced for this kind of thing.'

In the sentence, *zhe yi zhung shi* 'this kind of thing' is not the subcategorized argument of the verb *fenfu* 'abundant' , so it is annotated as *topic.*

(10) *zhe yi zhung yu wo zuei xihuan chi*
this one kind fish I most like eat
'I like to eat this kind of fish most.'

In this sentence, the object of the verb *chi* 'eat' is fronted and topicalized, so it is annotated as *topic [+theme]*.

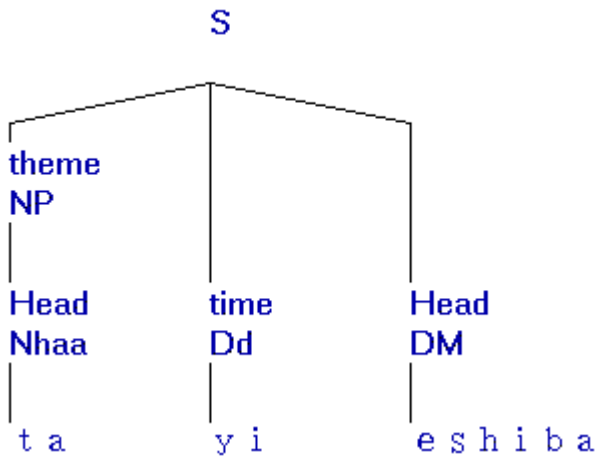
4.3. Construction with nominal predicate

There may be no verbal heads in Chinese sentences. In addition to being the head of NP, nominal constructions may function as predicates of sentences. Therefore, the construction with nominal head will be annotated as S but not NP when one of the following conditions is met.

a. The concept or meaning of the construction with nominal head is time.
b. The construction with nominal head taking adverbial modifiers, which is event related.

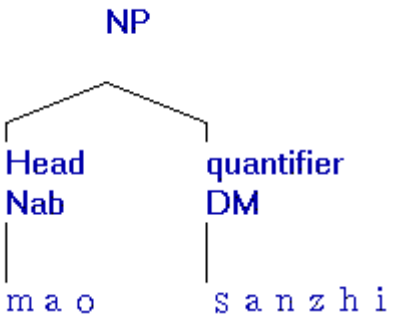
Compare example (10) with (11). The head of (10) is ' 28 years old', taking the concept of time. In addition, it takes a time adjunct, *yi* 'already'. On the contrary, (11) is coded as NP because it does not take any time related modifier with the construction.

(10)



He already 28
'He is 28 years old now.'

(11)



cat three CL
'three cats'

5. CONCLUSION AND FUTURE WORK

Under the framework of ICG, texts from 5-million-workd tagged corpus are bracketed and assigned to the labels of phrasal categories and thematic roles. After the manual post-editing, tree structures are constructed. In addition, we also intend to tag the CKIP Treebank with semantic categories in order to extract useful collocation among semantic classes of the bracketed constituents, which is supposed to further enhance the performance of our parsing model.

REFERENCES

CHEN, Keh-Jiann. (1996) "A Model for Robust Chinese Parser." *Computational Linguistics and Chinese Language Processing.* vol. 1, no.1. pp.183-204.

CHEN, Keh-Jiann, Chu-Ren Huang, Li-Ping Chang, Hui-Li Hsu. (1996). "Sinica Corpus: Design Methodology for Balanced Corpra." *Proceedings of the 11th Pacific Asia Conference on Language, Information, and Computation (PACLIC II)*, Seoul Korea, pp.167-176.

CHEN, Keh- Jiann, Shing-Huan Liu, Li-ping Chang, Yeh-Hao Chin (1994). "A Practical Tagger for Chinese Corpora." *Proceedings of ROCLING VII*, pp.111-126.

CHEN, Keh-Jiann, Chu-Ren Huang (1994). "Features Constraints in Chinese Language Parsing." *Proceedings of ICCPOL '94*, pp. 223-228

CHEN, Keh-Jiann (1992). "Design Concepts for Chinese Parsers." *3rd International Conference on Chinese Information Processing*, pp.1-22.

APPENDIX 1 SYNTACTIC CATEGORY

*NON-PREDCITIVE ADJVECTIVE

A

*CONJUNCTION

Caa , Cab, Cba, Cbaa, Cbab, Cbb, Cbba, Cbbb, Cbc, Cbca, Cbcb

*ADVERB

Daa, Dab (quantity)

Dbaa, Dbab, Dbb, Dbc (modal)

Dc (negation)

Dd (time)

Dfa, Dfb (degree)

Dg (locative)

Dh (manner)

Di (aspect)

Dj (interrogative)

Dk (sentential adverb)

*INTERJECTION

I

*NOUN

Naa (Mass Noun)

Nab (Common Noun)

Nac (Abstract Noun, Countable)

Nad (Abstract Noun)

Naea, Naeb (Group Noun)

Nba, Nbc (Proper Noun)

Nca, Ncb, Ncc, Ncda, Ncdb (Location Noun)

Ndaaa, Ndaab, Ndaac, Ndaad, Ndaba, Ndabb, Ndabc, Ndabd, Ndabe, Ndabf, Ndc, Ndca, Ndcb, Ndcc (Time Noun)

*DETERMINATIVES

Neu, Nes, Nep, Neqa, Neqb

*MEASURE WORD OR CLASSIFIER

Nfa, Nfb, Nfc, Nfd, Nfe, Nff, Nfg, Nfh, Nfi

*POSTPOSITION WORD

Ng

*PRONOUN

Nhaa, Nhab, Nhac, Nhb, Nhc

*PREPOSITION

P01, P02, P03, P04, P05, P06, P07, P08, P09, P10, P11, P12, P13, P14, P15, P16, P17, P18, P19, P20, P21, P22, P23, P24, P25, P26, P27, P28, P29, P30, P31, P32, P33, P34, P35, P36, P37, P38, P39, P40, P41, P42, P43, P44, P45, P46, P47, P48, P49, P50, P51, P52, P53, P54, P55, P56, P57, P58, P59, P60, P61, P62, P63, P64, P65

*PARTICLES

Ta, Tb, Tc, Td

*VERB

*ACTIVE INTRANSITIVE VERB

VA11, VA12, VA13, VA2, VA3, VA4

*PSEUDO ACTIVE TRANSITIVE VERB

VB11, VB12, VB2

*ACTIVE TRANSITIVE VERB

VC1, VC2, VC31, VC32, VC33

*DITRANSITIVE VERB

VD1, VD2

*ACTIVE VERB WITH SENTENTIAL OBJECT

VE11, VE12, VE2

*ACTIVE VERB WITH VP OBJECT

VF1, VF2

*CLASSIFICATORY VERB

VG1, VG2

*STATIVE INTRANSITIVE VERB

VH11, VH12, VH13, VH14, VH15, VH16, VH17, VH21, VH22

*PSEUDO STATIVE TRANSITIVE VERB

VI1, VI2, VI3

*STATIVE TRANSITIVE VERB

VJ1, VJ2, VJ3

*STATIVE VERB WITH SENTENCETIAL OBJECT

VK1 , VK2

*STATIVE VERB WITH VP OBJECT

VL1, VL2, VL3, VL4

APPENDIX 2 THEMATIC ROLES