Sinica Treebank:
Design Criteria, Annotation Guidelines, and On-line Interface

Chu-Ren Huang[1], Feng-Yi Chen[2], Keh-Jiann Chen[2], Zhao-ming Gao[3], & Kuang-Yu Chen[2]

churen@sinica.edu.tw, apple@iis.sinica.edu.tw, kchen@iis.sinica.edu.tw, zmgao@ccms.ntu.edu.tw, sasami@iis.sinica.edu.tw

[1]Institute of Linguistics, Academia Sinica, Taipei, Taiwan
[2]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[3]Dept. of Foreign Languages & Literatures, National Taiwan University, Taipei, Taiwan

Abstract

This paper describes the design criteria and annotation guidelines of Sinica Treebank. The three design criteria are: Maximal Resource Sharing, Minimal Structural Complexity, and Optimal Semantic Information. One of the important design decisions following these criteria is the encoding of thematic role information. An on-line interface facilitating empirical studies of Chinese phrase structure is also described.

1. Introduction

The Penn Treebank (Marcus et al. 1993) initiated a new paradigm in corpus-based research. The English Penn Treebank has enabled and motivated corpus and computational linguistic research based on information extractable from structurally annotated corpora. Recently, the research has focused on the following two issues: first, when and how can a structurally annotated corpus of language X be built?

Second, what information should or can be annotated? A good sample of issues in these two directions can be found in the papers collected in Abeille (1999).

The construction of the Sinica Treebank deals with both issues. First, it is one of the first structurally annotated corpora in Mandarin Chinese. Second, as a design feature, the Sinica Treebank annotation includes thematic role information in addition to syntactic categories. In this paper, we will discuss the design criteria and annotation guidelines of the Sinica Treebank. We will also give a preliminary research result based on the Sinica Treebank.

2. Design Criteria

There are three important design criteria for the Sinica Treebank: maximal resource sharing, minimal structural complexity, and optimal semantic information.

First, to achieve maximal resource sharing, the construction of the Sinica Treebank is bootstrapped from existing

Chinese computational linguistic resources. The textual material is extracted from the tagged Sinica Corpus (http://www.sinica.edu.tw/ftms-bin/kiwi.sh, Chen et al. 1996). In other words, the tasks and issues involving tokenization / word segmentation and category assignment are previously resolved. It is worth noting that the segmentation and tagging of Sinica Corpus have undergone vigorous post-editing. Hence the precision of category-assignment is much higher than with an automatically tagged corpora. In addition, since the same research team carried out the tagging of Sinica Corpus and annotation of Sinica Treebank, consistency of the interpretation of texts and tags are ensured. For structure-assignment, an automatic parser (Chen 1996) is applied before human post-editing.

Second, the criterion of minimal structural complexity is motivated to ensure that the assigned structural information can be shared regardless of users' theoretical presupposition. It is observed that theory-internal motivations often require abstract intermediate phrasal levels (such as in various versions of the X-bar theory). Other theories may also call for an abstract covert phrasal category (such as INFL in the GB theory for Chinese). In either case, although the phrasal categories are well-motivated within the theory, their significance cannot be maintained in the context of other theoretical frameworks. Since a primary goal of annotated corpora is to serve as the empirical base of linguistic investigations, it is desirable to annotate structure divisions that are the most commonly shared among theories. We came to the conclusion that the minimal basic level structures are the ones that are shared by all theories. Thus our annotation is designed to achieve minimal structural complexity. All abstract phrasal levels are eliminated and only canonical phrasal categories are marked.

Third, a critical issue involving Treebank construction as well as theories of NLP is how much semantic information, if any, should be incorporated. The original Penn Treebank took a fairly straightforward syntactic approach. A purely semantic approach, though tempting in terms of theoretical and practical considerations, has never been attempted yet. A third approach is to annotate partial semantic information, especially those pertaining to argument-relations. This is an approach shared by us and the Prague Dependency Treebank (e.g. Bohmova and Hajikova 1999). In this approach, the thematic relation between a predicate and an argument is marked in addition to grammatical category. Note that the predicate-argument relation is usually grammatically instantiated and generally considered to be the semantic relation that interacts most closely with syntactic behavior. This allows optimal semantic

information to be encoded without going too beyond the partially automatic process of argument identification.

3. Annotation Guidelines I: Category and Hierarchy

The basic structure of a tree in a treebank is a hierarchy of nodes with categorical denotation. As in any standard phrase structure grammar, the lexical (i.e. terminal) symbols are defined by the lexicon (CKIP 1992). And following the recent lexicon-driven and information-based trends in linguistic theory, linguistic information will be projected from encoded lexical information. Please refer to CKIP (1993) for the definition of lexical categories that we followed. We will give below the inventory of the restricted set of phrasal categories used and their interpretation. This set defines the domain of expressed syntactic information (instead of projected or inherited information). Readers can also consult Chen et al.'s (2000) general description of how the Sinica Treebank is constructed for a more complete list of tags as well as explanation in Chinese.

3.1. Defining Phrasal Categories

There are only 6 non-terminal phrasal categories annotated in the Sinica Treebank.

(1) Phrasal Categories

1. S: An S is a complete tree headed by a predicate (i.e. S is the start symbol).
2. VP: A VP is a phrase headed by a predicate. However, it lacks a subject and cannot function alone.
3. NP: An NP is headed by an N.
4. GP: A GP is a phrase headed by locational noun or locational adjunct. Since the thematic role is often determined by the governing predicate and not encoded locally; nominal phrases are given a tentative role of DUMMY so that it can inherit the correct role from the main predicate.
5. PP: A PP is headed by a preposition. The thematic role of its argument is inherited from the mother, hence its argument is marked with a DUMMY.
6. XP: A XP is a conjunctive phrase that is headed by a conjunction. Its syntactic head is the conjunction. However, since the actual category depends on the interactive inheritance from possibly non-identical conjoined elements, X in XP stands for an under-specified category.

3.2. Defining Inheritance Relations

Following unification-based grammatical theories, categorical assignments in Sinica Treebank are both lexicon-driven and head-driven. In principle, all grammatical information is lexically encoded. Structurally heads indicate the direction of information inheritance and define possible predicate-argument relations. However, since the notion 'head' can have several

different linguistic definitions, we attempt to allow at least the discrepancy between syntactic and semantic heads. In Sinica Treebank, three different kinds of grammatical heads are annotated.

(2) Heads

1. **Head**: indicates a grammatical head in an endocentric phrasal category. Unless a different semantic head is explicitly marked, a Head marks a category that serves simultaneously as the syntactic and semantic heads of the construction.
2. **head**: indicates a semantic head which does not simultaneously function as a syntactic head. For instance in constructions involving grammaticalized 'particles,' such as in the 'VP-de' construction, the grammatical head ('de' in this case) does not carry any semantic information. In these cases, the head marks the semantic head ('VP' in this case) to indicate the flow of content information.
3. **DUMMY**: indicates the semantic head(s) whose categorical or thematic identity cannot be locally determined. The two most likely scenarios involving DUMMY are (a) in a coordination construction, where the head category depends on the sum of all conjuncts. And (b) in a non-NP argument phrase, such as PP, where the semantic head carries a thematic role assigned not by the immediate governing syntactic head ("P" in this case), but by a higher predicate. In these cases, DUMMY allows a parser to determine the correct categorical / thematic relation later, while maintaining identical local structures.

3.3. Beyond Simple Inheritance

When simple inheritance fails, the following principles derived from our design criteria serve to predict the structural assignments of a phrasal category: default inheritance, sisters only, and left most.

3.3.1. Default Inheritance

This principle deals primarily and most effectively with coordinations and conjunctions. The theoretical motivation of this account follows Sag et al.'s (1985) proposal. In essence, the category of a conjunctive construction must be inherited from its semantic heads. However, since conjunctions are not restricted to same categories, languages must have principled ways to determine the categorical identity when different semantic heads carry different information.

First, in the trivial case when all head daughters are of the same category, the mother will inherit that category.

Second, when the different head daughters are an elaboration of the same basic category (e.g. both Nd and Ne are elaboration of N), then the basic category is the default inheritance category for the mother. This can be illustrated by (3).

(3) [[[*da4*]VH11[*er2*]Caa [*yuan2*]VH13]]
VP
big and round

Third, when other inheritance mechanisms fail to provide a clear categorical choice, the default inheritance is activated. There are two default hierarchies. The first one deals with when the head daughters are all lexical categories (4a), and the second one deals with when they are all phrasal categories (4b). If there is a disparity between lexical and phrasal categories, then a lexical category will be expanded to a phrasal category first.

(4) Default Inheritance Hierarchy for Categories
a) Lexical Categories: V > N > P > Ng
b) Phrasal Categories: S> VP> NP> PP> GP

When phrasal conjuncts are involved, S is the privileged category since it is the start symbol of the grammar. VP comes next since its structural composition is identical to that of S. If the structure involved is not a predicate (i.e. head of a sentence), then it must be a role. For argument roles, NP's are more privileged than PP's, and PP's are more privileged than GP's. (5) is an instance of the application of this default hierarchy.

(5) [[*da4liang4*]Neqa [*er*2]Caa [*feng1sheng4*]VH11]V]VP
big-quantity and bountiful
'bountiful and of big quantity'

When lexical conjuncts are involved, the same principle is used. The priority is given to the predicate head of the sentence. Among possible argument roles, the nominal category is the default. An illustrative example can be found in (6).

(6) [[*wei4lan2 de tian1kong1*]NP [*yu3*]Caa[*zhu1qun2biao1han4*]S]S
aqua-blue DE sky and people ferocious
'That the sky being aqua blue and that the people being ferocious…'

3.3.2 Sisters Only

Following most current linguistic theory, argument roles and adjunct complements must be sisters of a lexical head. However, driven by our design criteria of minimal structural complexity, no same level iteration is allowed. Thus these arguments and adjuncts can be located by the straightforward definition of sisterhood: that they share the same mother-daughter relation with the head. The result is a flat structure.

3.3.3 Left First

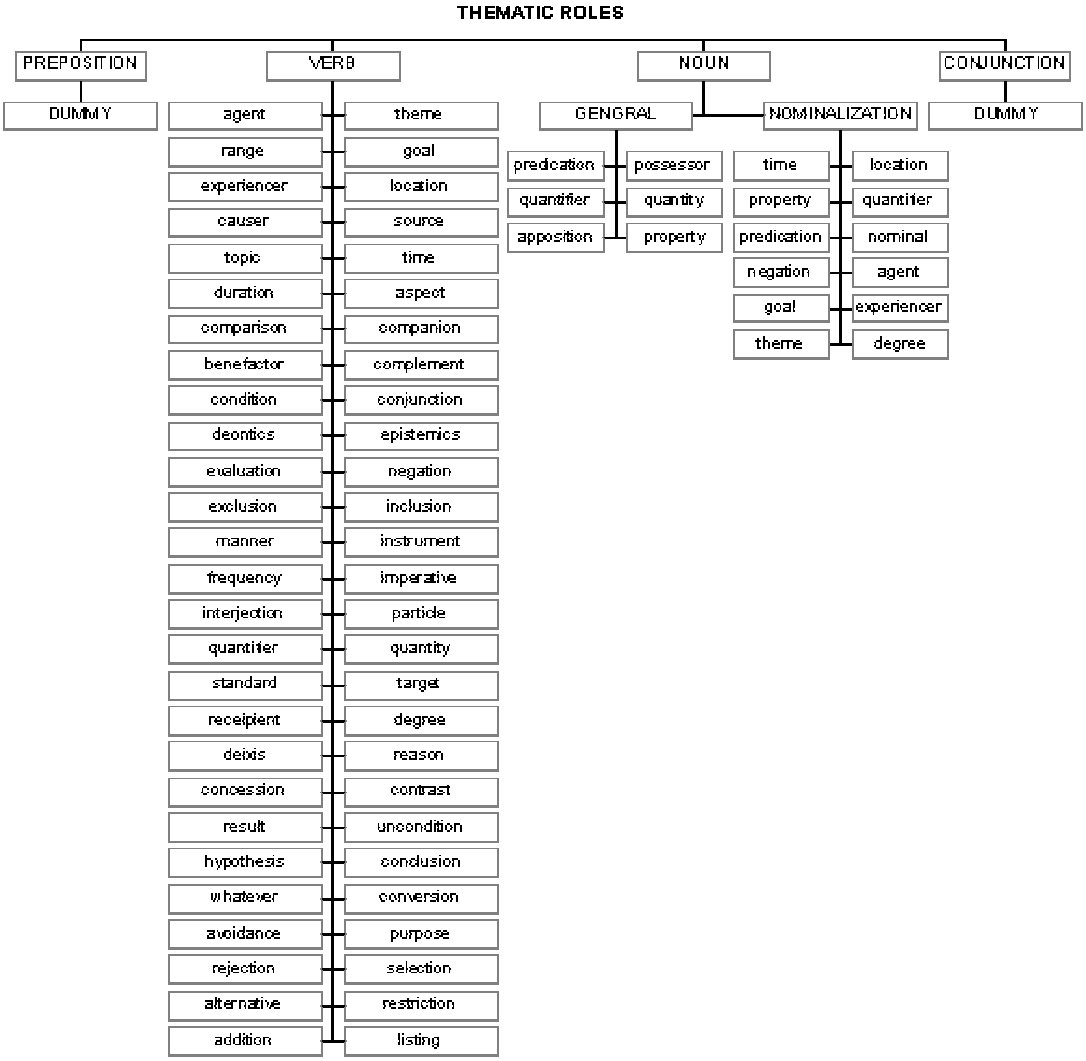
This principle is designed to account for possible internal structure when there are more than two sisters without having to add on hierarchical complexity. Hence, the default interpretation of internal structure of multiple sisters is that the internal association starts from left to right.

4. Annotation Guidelines II: Structural Annotation of Thematic Information

A thematic relation contains a compact bundle of syntactic and semantic information. Although thematic relations are lexically encoded on a predicate, they can only be instantiated when that information is projected to phrasal arguments. In other words, the only empirical evidence for the existence of a thematic relation is a realized argument. However, a realized argument cannot by itself determine the thematic relation. The exact nature of the relation must be determined based on the lexical information from the predicate as well as checking of the compatibility of that realized argument. Since structural information alone cannot determine thematic relations, prototypical structural annotation, such as in the original Penn Treebank, does not include thematic roles since they contain non-structural information.

On the other hand, in theories where lexical heads drive the structural derivation / construction (e.g. ICG and HPSG and LFG), thematic relations are critical. Hence, we decided to encode realized thematic relations on each phrasal argument. The list of thematic relations encoded on the head predicate is consulted whenever a phrasal argument is constructed, and a contextually appropriate relation sanctioned by the lexical information is encoded. It is worth noting that in our account, we not only mark the thematic relations of a verbal predicate, but we also mark the thematic relations governed by a deverbal noun, among others. Also note that an argument of a preposition is marked as a placeholder DUMMY. This is because a preposition only governs an argument syntactically, while its thematic relation is determined by a higher verb.

(7) Thematic Roles: Classification and Inventory



5. Current Status of the Sinica Treebank and On-line Interface

Following the above criteria and principles, we have already finished Sinica Treebank 1.0. It contains annotations of 38,725 Chinese structural trees containing 239,532 words. It covers subject areas that include politics, traveling, sports, finance, society, etc. This version of the Sinica Treebank will be released in the near future as soon as the licensing documents are cleared by the legal department of Academia Sinica. A small subset of it (1,000 sentences) is already available for researchers to download from the website *http://godel.iis.sinica.edu.tw/CKIP/trees1000.htm*. A searchable interface is also being developed and tested for researchers so that they can directly access the complete treebank information.

As an annotated corpus, one of the most important roles that a treebank can play is that it can serve as a shared

source of data for linguistic, especially syntactic studies. Following the example of the successful Sinica Corpus, we have developed an on-line interface for extraction of grammatical information from the Sinica Treebank. Although the users that we have in mind are theoretical linguists who do not necessarily have computational background; we hope that non-linguists can also benefit from the ready availability of such grammatical information. And of course, computational linguists should be able to use this interface for quick references before going into a more in-depth study of the annotated corpus.

Currently, the beta site allows users specify a variety of conditions to search for structurally annotated sentences. Conditions can be specified in terms of keywords, grammatical tags (lexical or phrasal), thematic relations, or any boolean combination of the above elements. The search result can be presented as either annotated structure or simply the example sentences. Simply statistics, based on either straightforward frequency count or mutual information, are also available. For linguistically interesting information, such as the heads of various phrasal constructions, a user can simply look up the explicitly syntactic **Head** or semantic **head**; as well as **DUMMY** when it serves as a head placeholder. The website of this interface, as well as the general release of the Sinica Treebank 1.0, is scheduled to be announced at the second ACL workshop on Chinese Language Processing in October 2000.

6. Conclusion

The construction of the Sinica Treebank is only a first step towards application of structurally annotated corpora. Continuing expansion and correction will make this database an invaluable resource for linguistic and computational studies of Chinese.

References

1. ABEILLE, Anne. 1999. Ed. Proceedings of *ATALA Workshop – Treebanks.* Paris, June 18-19, 1999. Univ. de Paris VII.
2. BOHMOVA, Alla and Eva Hajicova. 1999. How Much of the Underlying Syntactic Structure Can be Tagged Automatically? In Abeille (Ed). 1999.31-40.
3. CHEN, Feng-Yi, Pi-Fang Tsai, Keh-Jiann Chen, and Chu-Ren Huang. 2000. Sinica Treebank. [in Chinese] *Computational Linguistics and Chinese Language Processing*. 4.2.87-103.
4. CHEN, Keh-Jiann. 1996. A Model for Robust Chinese Parser. *Computational Linguistics and Chinese Language Processing*. 1.1.183-204.
5. CHEN, Keh-Jiann, Chu-Ren Huang. 1996. Information-based Case Grammar: A Unification-based Formalism for Parsing Chinese. In Journal of Chinese Linguistics Monograph Series No. 9. Chu-Ren Huang, Keh-Jiann Chen, and Benjamin K. T'sou Eds. *Readings in Chinese Natural Language Processing.* 23-45. Berkeley: JCL.
6. CHEN, Keh-Jiann, Chu-Ren Huang, Li-Ping Chang, Hui-Li Hsu. 1996.

Sinica Corpus: Design Methodology for Balanced Corpora. *Proceedings of the 11th Pacific Asia Conference on Language, Information, and Computation (PACLIC II)*. Seoul Korea. 167-176.
7. CHEN, Keh-Jiann and Shing-Huan Liu. 1992. Word Identification for Mandarin Chinese Sentences. *Proceedings of COLING*-92.101-105.
8. CHEN, Keh- Jiann, Shing-Huan Liu, Li-Ping Chang, Yeh-Hao Chin. 1994. A Practical Tagger for Chinese Corpora." *Proceedings of ROCLING VII*.111-126.
9. CHEN, Keh-Jiann, Chi-Ching Luo, Zhao-Ming Gao, Ming-Chung Chang, Feng-Yi Chen, and Chao-Ran Chen. 1999. The CKIP Chinese Treebank: Guidelines for Annotation. In Abeille (Ed). 1999.85-96.
10. CKIP (Chinese Knowledge Information Processing). 1993. The Categorical Analysis of Chinese. CKIP Technical Report 93-05. Nankang: Academia Sinica.
11. HUANG, Chu-Ren, Keh-Jiann Chen, Feng-Yi Chen, and Li-Li Chang. 1997. Segmentation Standard for Chinese Natural Language Processing. *Computational Linguistics and Chinese Language Processing*. 2.2.47-62
12. Lin, Fu-Wen. 1992. Some Reflections on the Thematic System of Information-based Case Grammar (ICG). CKIP Technical Report 92-01. Nankang: Academia Sinica.
13. Marcus, Mitch P., Beatrice Santorini, and M. A. Marcinkiewiicz. 1993. Building a Large Annotated Corpus of English: The Peen Treebank. Computational Linguistics. 19.2.313-330.
14. SAG, Ivan, Gerald Gazdar, Thomas Wasow, and Steven Weisler. 1985. Coordination and How to Distinguish Categories. Natural Language and Linguistic Theories. 117-171.

Appendix

1. Lexical Categories

(1) NON-PREDCITIVE ADJVECTIVE: A
(2) CONJUNCTION: C
(3) ADVERB: D
(4) INTERJECTION: I
(5) NOUN: N
(6) DETERMINATIVES: Ne
(7) MEASURE WORD / CLASSIFIER: Nf
(8) POSTPOSITION WORD: Ng
(9) PRONOUN: Nh
(10) PREPOSITION: P
(11) PARTICLES: T
(12) VERB: V

2. Sample Sentence and Tree

那個挽髻的女人白髮之後
便不再理會庭前亭亭玉立的青草
nage wanji de nyuren baifa zhihou bian buzai lihui
that hair-style DE woman white-hair after then never pay-attention
ting qian tingting yuli de qingcao
courtyard front slender-ly standing-erect DE green-grass
'After her hair had turned white, that coiffured woman never paid any more attention to the nicely standing green grass in the front courtyard.'
S(agent:NP(quantifier:DM:那個|property:VP・的(head:VP(Head:VA4:挽髻)|Head:DE:的)|Head:Nab:女人)|time:GP(DUMMY:VP(Head:VH11:白髮)| Head:Ng:之後)|time:Dd:便|time:Dd:不再| Head:VC2:理會|goal:NP (property:VP・的(head:VP (location:NP(property:Ncb:庭|Head:Ncda:前)|Head:VH11:亭亭玉立)|Head:DE:的) | Head:Nab:青草))