

Introduction to Chinese Treebanks

Yuqing Guo
NCLT, DCU
1 Sep. 2005






- Penn Chinese Treebank
- Sinica Chinese Treebank
- Harbin Treebank
- Summary



Penn Treebank – Background

- Developer: The logo for the University of Pennsylvania, featuring a shield with three open books and the word "Penn" in a serif font, with "UNIVERSITY OF PENNSYLVANIA" written below it.
- Scale
 - v5.0: 18,782 sentences & 507,222 words (890 files)
 - v4.0: 15,162 sentences & 404,156 words (838 files)
- Source
 - Newswire
 - Xinhua (1994 and 1998), HKSAR (1997) & Sinorama Taiwan (1996-1998 & 2000-2001)
- Format
 - SGML



- Syntactic categories
 - Part-Of-Speech tags: 33
 - Bracket tags: 23
 - Clause level: 2
 - Phrasal level: 15
 - Verb compounds: 6
- Functional categories: 26
 - Clause types: 2
 - Grammatical roles: 7
 - Adverbials: 11
 - Others: 6
- Empty categories: 7



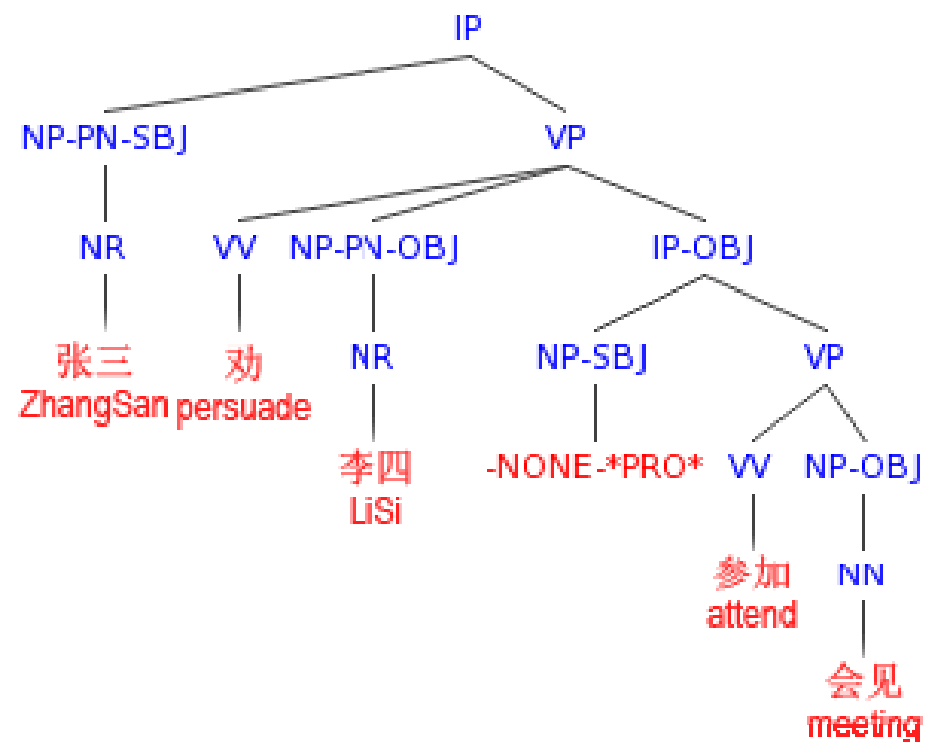
Penn Treebank – Sample

张三劝李四参加会见 (ZhangSan persuaded LiSi to attend the meeting)

```

(IP (NP-PN-SBJ (NR 张三))
  (VP (VV 劝)
    (NP-PN-OBJ (NR 李四))
    (IP-OBJ (NP-SBJ (-NONE- *PRO*))
      (VP (VV 参加)
        (NP-OBJ (NN 会见)))))))

```



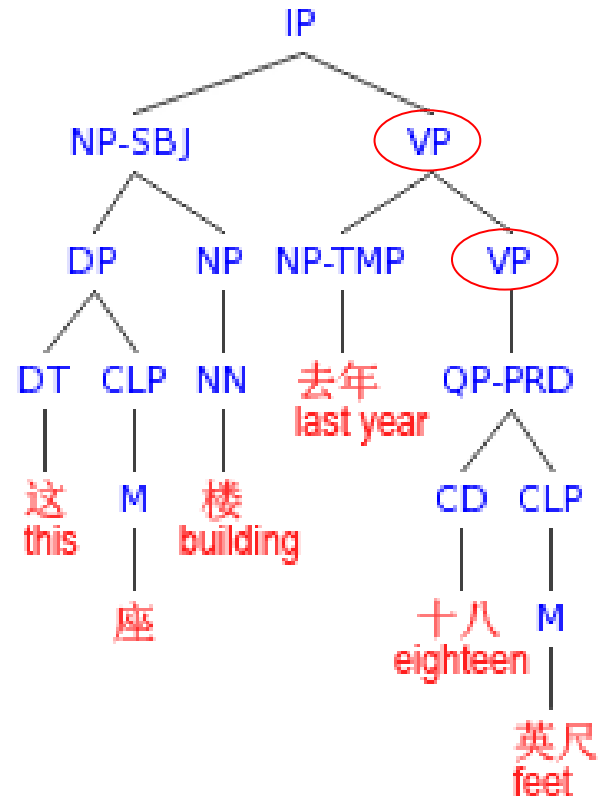


Penn Treebank – Sample

- Non-verbal predicates

去年这座楼十八英尺
This building was eighteen feet high last year

(IP (NP-SBJ (DP (DT 这)
(CLP (M 座)))
(NP (NN 楼)))
(VP (NP-TMP 去年)
(VP (QP-PRD (CD 十八)
(CLP (M 英尺))))))





Sinica Treebank – Background

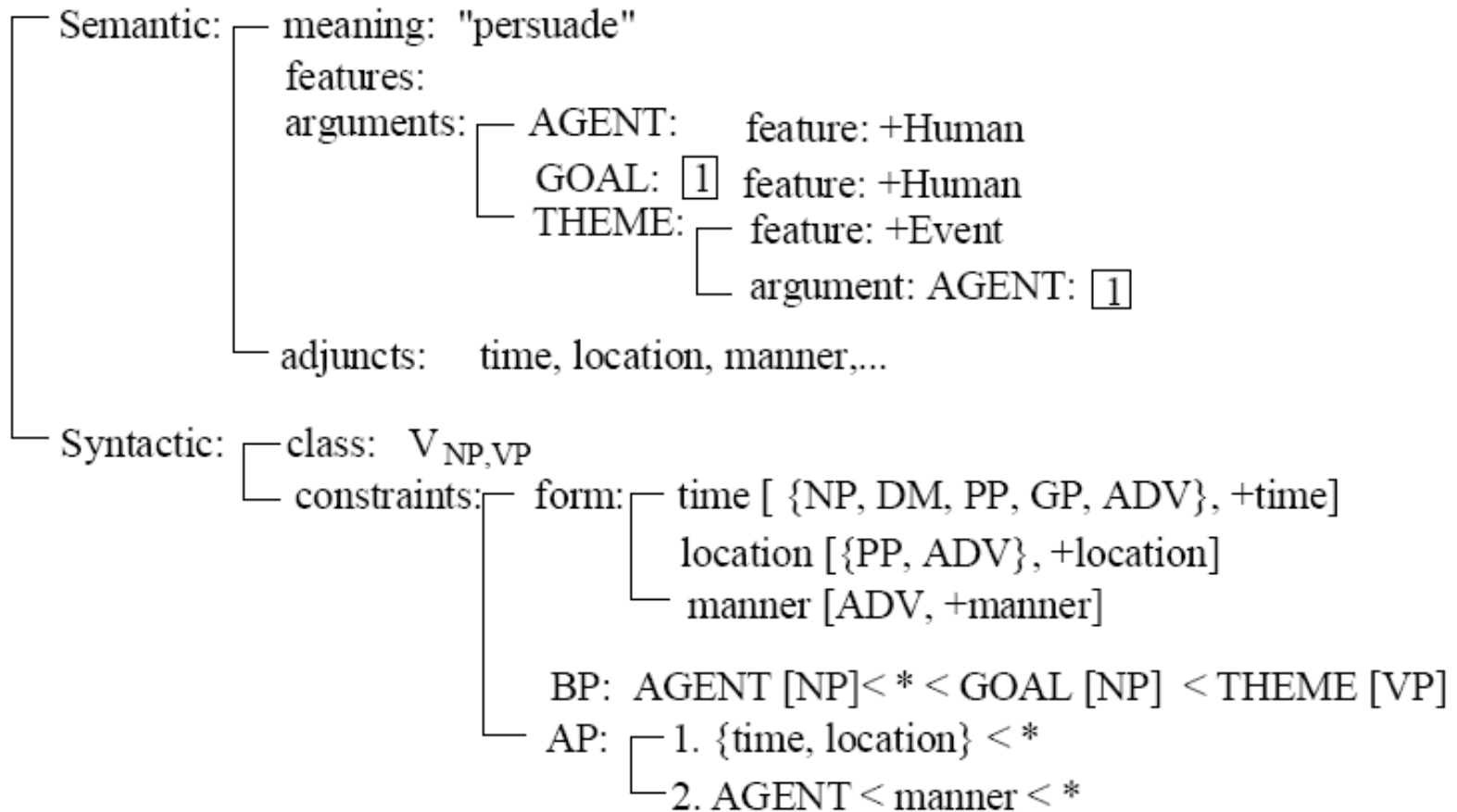
- Developer:  
- Scale
 - v2.1: 54,902 trees & 290,144 words (23 files)
- Source: Sinica Corpus
 - A Balanced Corpus of Modern Chinese
 - 5 million words
 - Word segmented and POS tagged
- Process
 - Automatically parsed
 - Manually checked



- Representation Model
 - Head-Driven Principle
 - Information-based Case Grammar (ICG, Chen 1990, 1996)
- Syntactic categories
 - Lexical categories
 - 46 main categories
 - 192 sub-categories
 - Phrasal categories
 - 6 common categories
 - 3 special categories
- Thematic roles



勸 Quan "persuade":





Sinica Treebank – Thematic roles

Verb		agent, addition, alternative, aspect, avoidance, benefactor, causer, companion, comparison, complement, concession, conclusion, condition, conjunction, contrast, conversion, degree, deixis, deontics, duration, epistemics, evaluation, exclusion, experiencer, frequency, goal, Head, hypothesis, imperative, inclusion, instrument, interjection, listing, location, manner, negation, nominal, particle, purpose, range, reason, recipient, rejection, restriction, result, selection, source, standard, target, theme, time, topic, uncondition, whatever
Noun	general	predication, property, possessor, apposition, quantifier, quantity
	nominalization	time, location, predication, property, quantifier, nominal, agent, theme, experiencer, goal, negation
Preposition		dummy
Conjunction		dummy1, dummy2

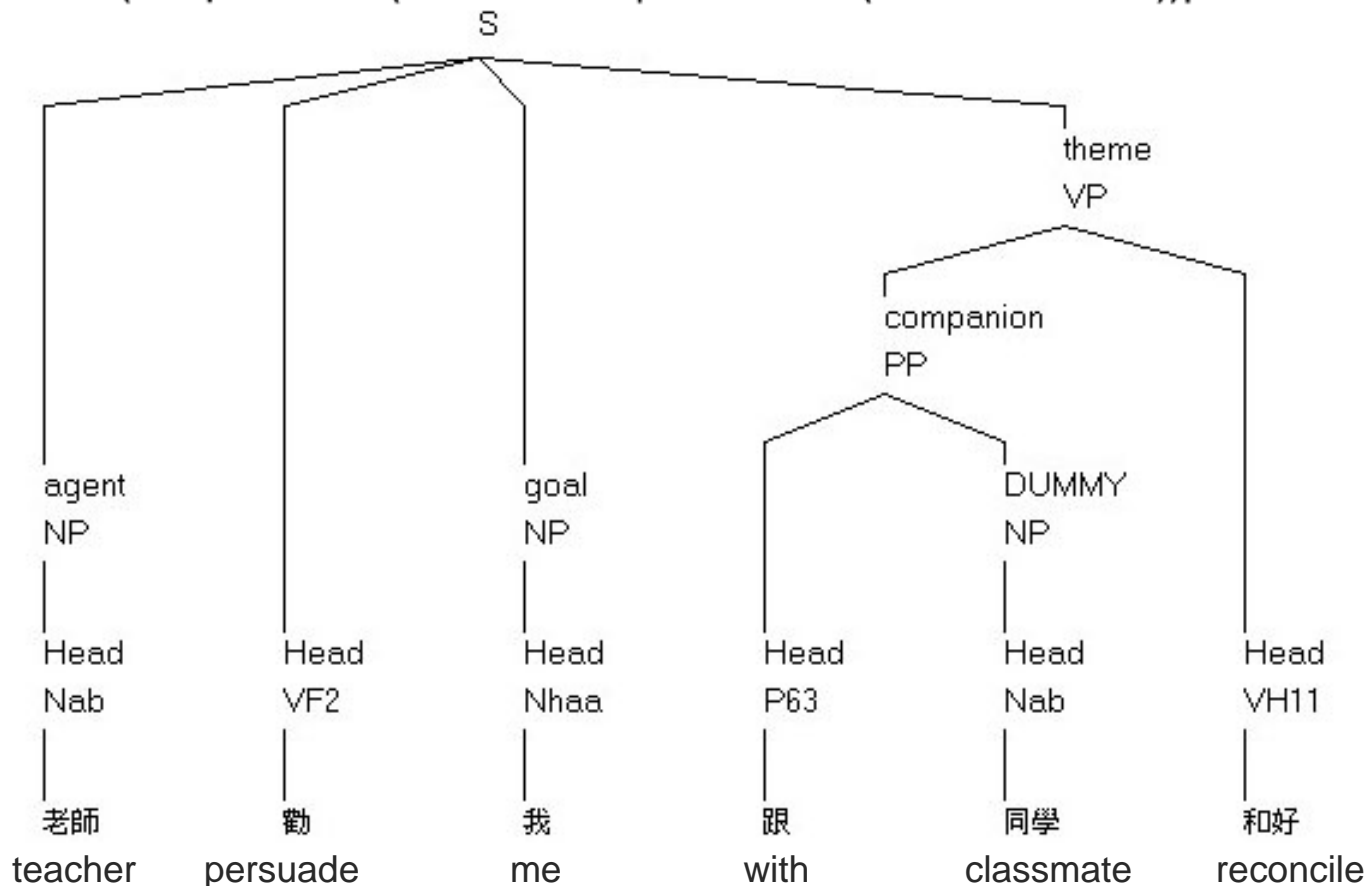


Sinica Treebank – Sample

老師勸我跟同學和好

The teacher persuaded me to be reconciled with the classmates

S(agent:NP(Head:Nab:老師)|Head:VF2:勸|goal:NP(Head:Nhaa:我)|
theme:VP(companion:PP(Head:P63:跟|DUMMY:NP(Head:Nab:同學))|Head:VH11:和好))





Harbin Treebank – Background

- Developer:  哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY
- Scale:
 - 10,000 sentences & 129,707 words
- Syntactic categories
 - Lexical categories
 - 42 word tags
 - 10 punctuation tags
 - Phrasal categories
 - 30 tags



Harbin Treebank – Problems

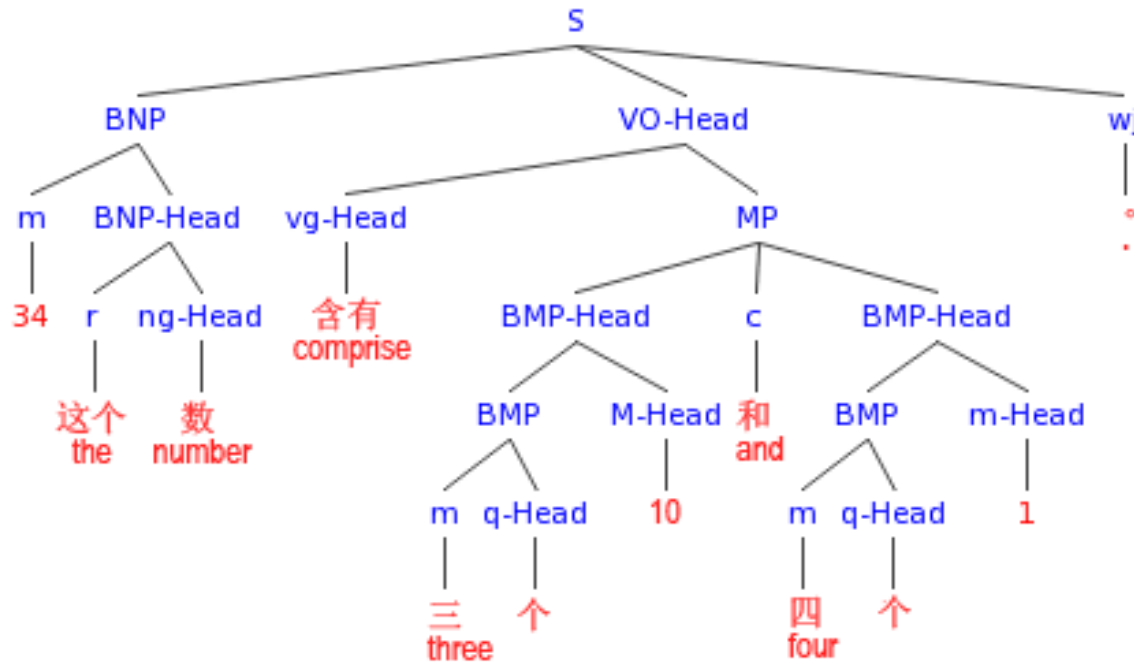
- Phrasal bracketing
- Variant sentence length
 - from 2 to 120 words
 - no clause level tagging
- Inconsistent annotation
- Flexible format
- Unmatched brackets
- Errors



Harbin Treebank – Sample

34这个数含有三个10和四个1。
 The number of 34 comprises three 10 and four 1.

BNP[34/m *BNP[这个/r *数/ng]]*VO[*含有/vg MP[*BMP[BMP[三/m *个/q] *10/M]
 和/c *BMP[BMP[四/m *个/q]*1/m]]]。 /wj





	Penn	Sinica	HIT
Coding	GB	Big5	GB
Sentence Length	29	6	13
POS Tag	33	192	52
Phrasal Tag	23	9	30
Functional Tag	26	60	1
Purpose	f-structure annotation	golden standard	supplement



Tree Graphing Tool

The screenshot shows the TBViewer application window. On the left, a file tree displays the structure of the file 'chtb_002.fid', including a title and two paragraphs. Paragraph 4 is highlighted, containing the sentence: '去年实现进出口总值达一千零九十八点二亿美元，占全国进出口总值的比重由上年的百分之三十七提高到百分之三十九。' On the right, a tree graph visualizes the syntactic structure of this sentence. The root node is 'IP', which branches into 'NP-SBJ' and 'VP'. 'NP-SBJ' further branches into 'CP' and 'WHNP-1'. 'CP' branches into '-NONE-' and '*OP*'. 'WHNP-1' branches into 'IP', which then branches into 'NP-SBJ'. 'NP-SBJ' branches into 'NP-TMP' and 'NP-OBJ'. 'NP-TMP' branches into 'NT' (去年). 'NP-OBJ' branches into 'VP' and '-NONE-' '*T*-1'. 'VP' branches into 'VV' (实现). 'NP' branches into 'NN' (进出口) and 'NN' (总值). 'VP' branches into 'VV' (达). 'QP-OBJ' branches into 'CD' (一千零九十八点二亿) and 'CLP' (M 美元). The root 'IP' also branches into 'PU'.

**Thanks !
&
Any Questions?**

