Creating a treebank

Lecture 3: 7/15/2011

Ambiguity

- Phonological ambiguity: (ASR)
  - “too”, “two”, “to”
  - “ice cream” vs. “I scream”
  - “ta” in Mandarin: he, she, or it
- Morphological ambiguity: (morphological analysis)
  - unlockable: [[un-lock]-able] vs. [un-[lock-able]]
- Syntactic ambiguity: (parsing)
  - John saw a man with a telescope
  - Time flies like an arrow
  - I saw her duck

Ambiguity (cont)

- Lexical ambiguity: (WSD)
  - Ex: “bank”, “saw”, “run”

- Semantic ambiguity: (semantic representation)
  - Ex: every boy loves his mother
  - Ex: John and Mary bought a house

- Discourse ambiguity:
  - Susan called Mary. She was sick. (coreference resolution)
  - It is pretty hot here. (intention resolution)

- Machine translation:
  - “brother”, “cousin”, “uncle”, etc.

Motivation

- Treebanks are valuable resources for NLP:
  - Word segmentation
  - POS tagging
  - Chunking
  - Parsing
  - Named entity detection
  - Semantic role labeling
  - Discourse
  - Co-reference
  - Event detection
  - …
- Problem: Creating treebanks is still an art, not a science.
  - what to annotate?
  - how to annotate?
  - who is in the team?

My experience with treebanks

- As a member of the Chinese Penn Treebank (CTB) project: 1998-2000
  - Project manager
  - Designed annotation guidelines for segmentation, POS tagging, and bracketing (with Nianwen Xue).
  - Organized several workshops on Chinese NLP

- As a user of treebanks
  - grammar extraction
  - POS tagging, parsing, etc.

Current work

- RiPLes project:
  - To build mini-parallel-treebanks for 5-10 languages
  - Each treebank has 100-300 sentences

- The Hindi/Urdu treebank project (2008-now):
  - Joint work with IIIT, Univ of Colorado, Columbia Univ, and UMass

Outline

- Main issues for treebanking

- Case study: the Chinese (Penn) Treebank

The general process

- Stage 1: get started
  - Have an idea
  - The first workshop
  - Form a team
  - Get initial funding
- Stage 2: initial annotation
  - create annotation guidelines
  - train annotators
  - manual annotation
  - train NLP systems
  - initial release
- Stage 3: more annotation
  - The treebank is used in CL and ling communities
  - Get more funding
  - Annotate more data
  - Add other layers

Main issues

- Creating guidelines
- Involving the community
- Forming a team
- Selecting data

- Role of processing NLP tools
- Quality control
- Distributing the data
- Future expansion of the treebanks

Guideline design: Highlights

- Detailed, “searchable” guidelines are important
  - Ex: the CTB’s guidelines have 266 pages

- Guidelines take a lot time to create, and revising the guidelines after annotation starts is inevitable.
  - An important issue: How to update the annotation when the guidelines changes?

- It is a good idea to involve the annotators while creating the guidelines

- Define high-level guiding principles, which lower-level decisions should follow naturally

  ➔ reduce the number of decisions that annotators have to memorize

A high-quality treebank should be

- Informative: it provides the info needed by its users
  - Morphological analysis: lemma, derivation, inflection
  - Tagging: POS tags
  - Parsing: phrase structure, dependency relation, etc.
  - ...

- Accurate and consistent: these are important for
  - training
  - evaluation
  - conversion

- Reasonable annotation speed

- Some tradeoff is needed:
  - Ex: walked/VBD vs. walk/V+ed/pastTense

An example: the choice of the tagset

- Large tagset vs. small tagset

- Types of tags:
  - POS tags: e.g., N, V, Adj
  - Syntactic tags: e.g., NP, VP, AdjP
  - Function tags: e.g., -TMP, -SBJ
    - Temporal NPs vs. object NPs
    - Adjunct/argument distinction
  - Empty categories: e.g., *T*, *pro*
    - Useful if you want to know subcategorization frames, long-distance dependency, etc.

When there is no consensus

- Very often, there is no consensus on various issues

- Try to be “theory-neutral”: linguistic theories keep changing

- Study existing analyses and choose the best ones

- Make the annotation rich enough so that it is easy to convert the current annotation to something else

Two common questions for syntactic treebanks

- Grammars vs. annotation guidelines

- Phrase structure vs. dependency structure

Writing grammar vs. creating annotation guidelines

- Similarity:
  - Both require a thorough study of the linguistic literature and a careful selection of analyses for common constructions

- Differences:
  - Annotation guidelines can leave certain issues undecided/uncommitted.
    - Ex: argument / adjunct distinction
  - Annotation guidelines need to have a wide coverage, including the handling of issues that are not linguistically important
    - Ex: attachment of punctuation marks

- The interaction between the two:
  - Treebanking with existing grammars
  - Extracting grammars from treebanks

Treebanking with a pre-existing grammar

- Ex: Redwoods HPSG treebank

- Procedure:
  - Use the grammar to parse the sentences
  - Correct the parsing output

- Advantage:
  - The analyses used by the treebank are as well-founded as the grammar.
  - As the grammar changes, the treebank could potentially be automatically updated.

- Disadvantage:
  - It requires a large-scale grammar.
  - The treebank could be heavily biased by the grammar

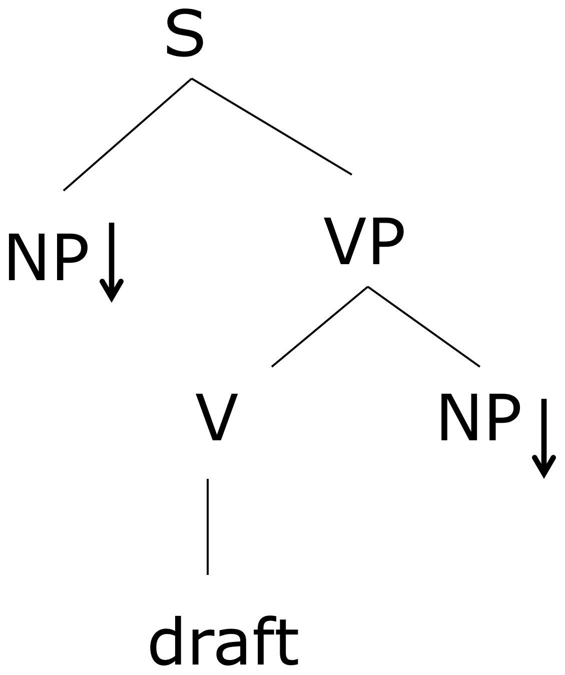
Extracting grammars from treebanks

- A lot of work on grammar extraction
  - Different grammar formalisms: e.g., CFG, LTAG, CCG, LFG

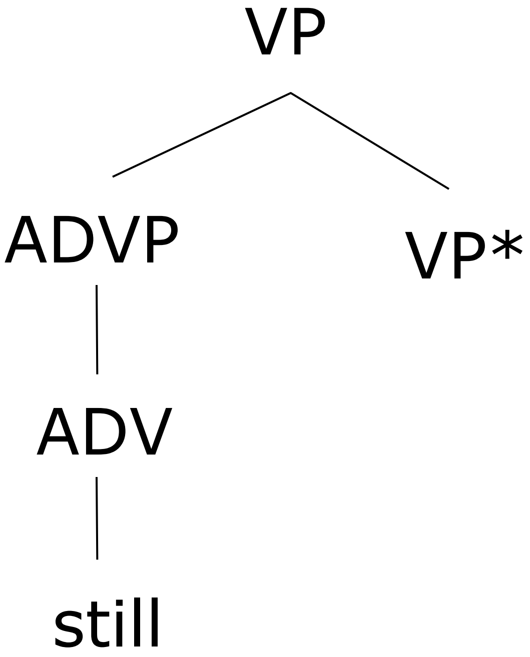
- Compared to hand-crafted grammars
  - Extracted grammars have better coverage and include statistical information, both are useful for parsing.
  - Extracted grammars are more noisy and lack rich features.

Extracting LTAGs from Treebanks

Initial tree:

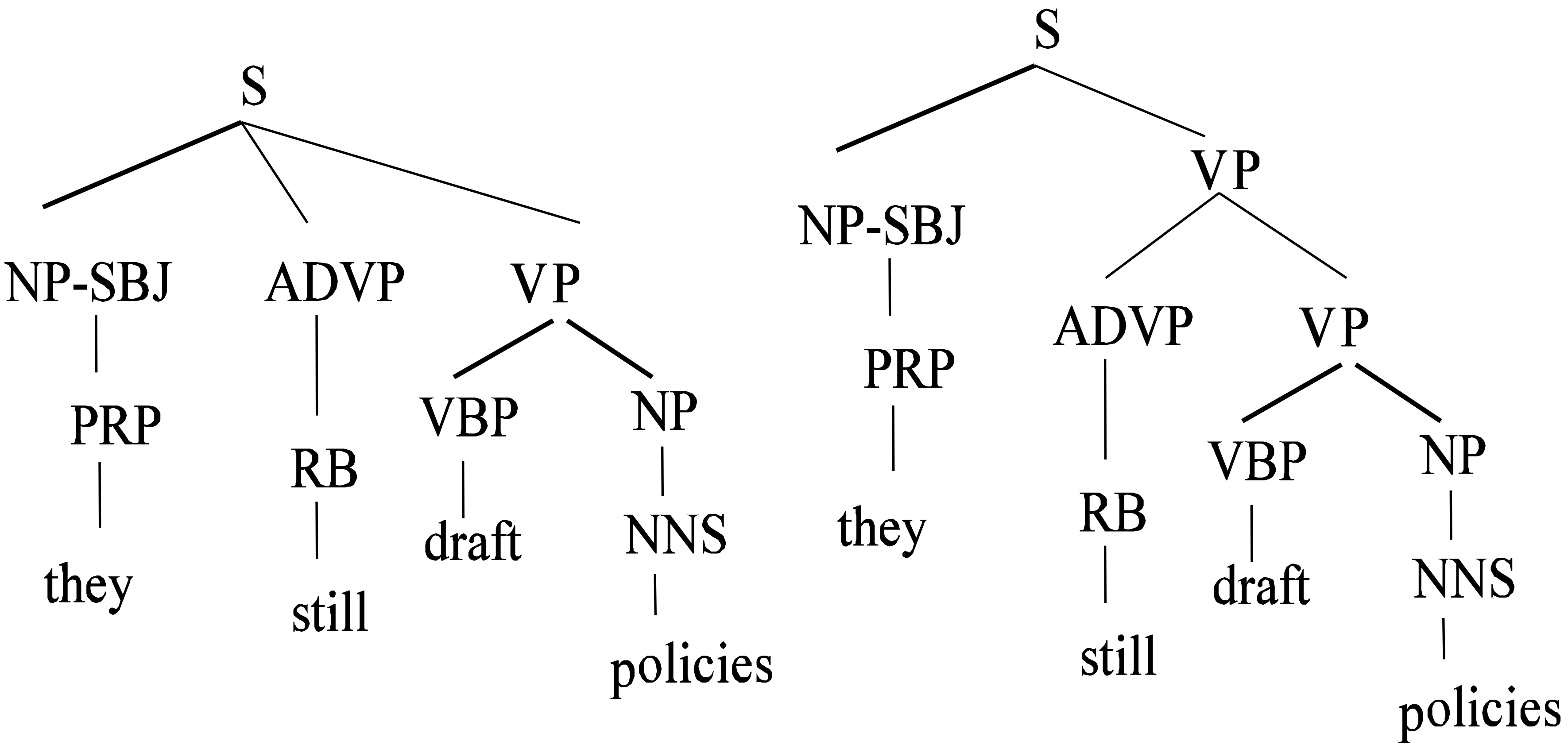


Auxiliary tree:



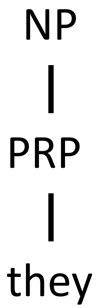
➔Arguments and adjuncts are in different types of elementary trees

The treebank tree

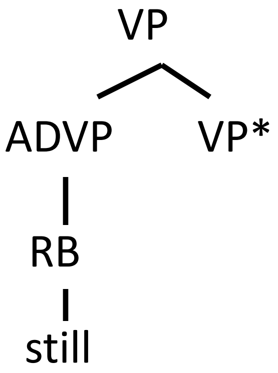


Extracted grammar

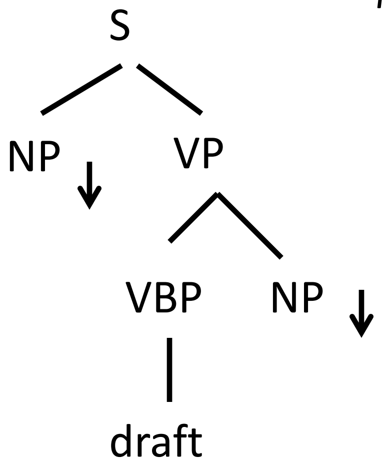
#1:



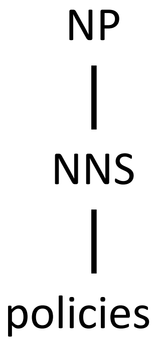
#2:



#3:



#4:



We ran the system (LexTract) to convert treebanks into the data that can be used to train and test LTAG parsers.

Two common questions

- Grammars vs. annotation guidelines
  - Grammars and treebank guidelines are closely related.
  - There should be more interaction between the two.

- Phrase structure vs. dependency structure

Information in PS and DS

	PS (e.g., PTB)	DS (some target DS)
POS tag	yes	yes
Function tag (e.g., -SBJ)	yes	yes
Syntactic tag	yes	no
Empty category and co-indexation	Often yes	Often no
Allowing crossing	Often no	Often yes

PS or DS for treebanking?

- PS treebank is good for phrase structure parsing
- Dependency treebank is good for dependency parsing.
- Ideally, we want to have both. But annotating both would be too expensive.

- Conversion algorithms between the two have been proposed, but they are far from perfect.

- Remedy: Make annotations (just) rich enough to support both.
  - Ex: mark the head in PS

PS ➔ DS

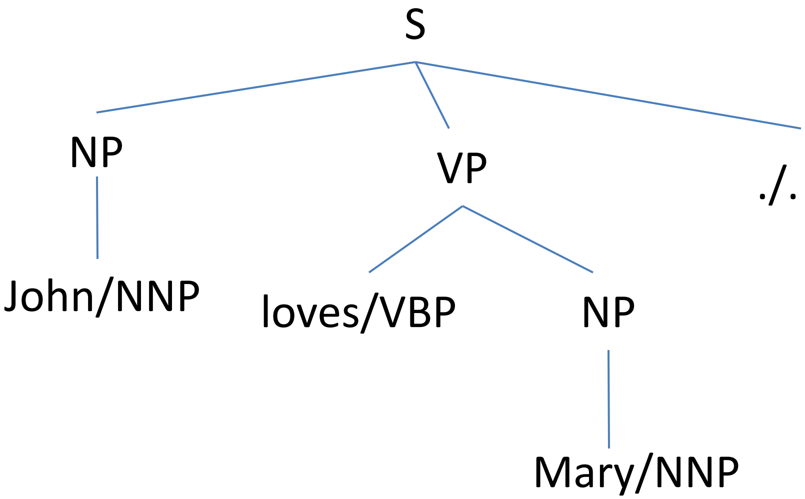
- For each internal node in the PS

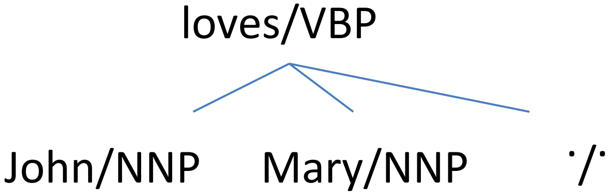
  (1) Find the head child

  (2) Make the non-head child depend on head-child

- For (1), very often people use a head percolation table and functional tags.

An example





Use a head percolation table:

(S, right, S/VP/….)
(NP, right, NP/NNP/NNPS/CD/…)
(VP, left, VP/VBP/VBD/…)

The approach is not perfect.

DS ➔ PS

- (Collins, Hajič, Ramshaw and Tillmann, 1999)
- (Xia and Palmer, 2001)
- (Xia et al., 2009)
- All are based on heuristics.
- Need to handle non-projectivity and ambiguity.

Main issues

- Creating guidelines
- Involving the community
- Forming the team
- Selecting data

- Role of processing NLP tools
- Quality control
- Distributing the data
- Future expansion of the treebanks

Community involvement

- Before the project starts, find out
  - what the community needs
  - whether there are existing resources (guidelines, tools, etc.)
- During the project, ask for feedback on
  - new guidelines
  - annotation examples
  - tools trained on preliminary release
- Don’t be discouraged by negative feedback

Forming the team

- Computational linguists:
  - Create annotation guidelines
  - Make/use NLP tools for preprocessing, final cleaning, etc.

- Linguistics experts
  - Help to create annotation guidelines

- Annotators
  - Training on linguistics and NLP is a big plus

- Advisory board: experts in the field

Annotators

- Linguists can make good annotators!

- Training annotators well takes a very long time

- Keeping trained annotators is not easy
  - Full time is good (combo annotation and scripting, error searching, workflow, etc.)

- Good results are possible:
  - Ex: IAA for CTB is 94%

Selecting data

- Permission for distribution

- The data should be a good sample of the language.

- Data from multiple genres?
  - Ex: 500K words from one genre, 250K from one genre and 250K from another, or other combinations?

- Active learning
  - To select the hardest sentences for annotation. Good idea?

Roles of tools

- Annotation tools

- Preprocessing tools

- Other tools:
  - Corpus search tools: e.g., tgrep2
  - Conversion tools:
  - Error detection tools:

Preprocessing tools (e.g., taggers, parsers)

- Use pre-existing tools or train new ones:
  - train a tool with existing data
  - preprocess new data with the tool
  - manually check and correct errors
  - Add the new data to the training data
  - Repeat the procedure

- It can speed up annotation and improve consistency

- However, the tools introduce a big bias to the treebanks, as annotators often fail to correct the mistakes introduced by the tools.

- Quality control is essential.

Quality control

- Human errors are inevitable

- Good guidelines, well-trained annotators, easy-to-use annotation tools, search tools, …

- Inter-annotator agreement should be monitored throughout the project.

- Detecting annotation errors using NLP tools

- Feedback from the user
  - From parsing work
  - From PropBank work
  - From grammar extraction work
  - …

Inter-annotator agreement

- Procedure:
  - Randomly select some data for double annotation
  - Compare double annotation results and create gold standard
  - Calculate annotation accuracy (e.g., f-measure) and inter-annotator agreement

- Possible reasons of the disagreement:
  - Human errors
  - Problems in annotation guidelines

    ➔ modify the guidelines if needed

Distributing the data

- Find a good collaborator: e.g., LDC

- Multiple releases
  - Preliminary releases for feedback
  - Later release with more data and/or fewer errors

- Presentations at major conferences

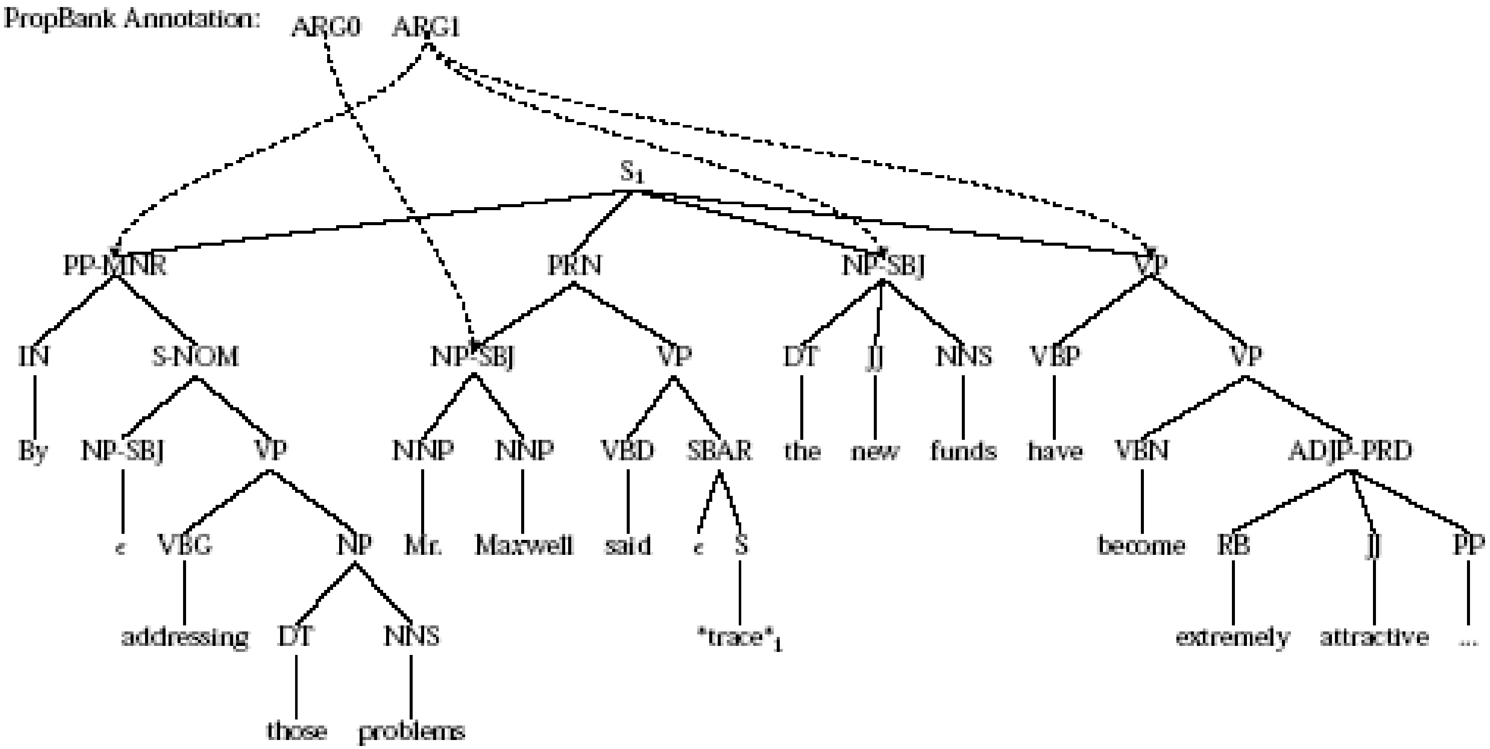
Expanding the treebank

- More data

- More genres

- Other layers of information
  - Ex: PropBank, NomBank, Discourse Treebank on top of treebanks
  - The choice made by the treebank could affect new layers

Treebank-PropBank Reconciliation



Problem: One PropBank argument can involve many parse nodes

Solution: Single argument – single parse node analysis

Outline

- Main issues for treebanking
- Case study: The Chinese Penn Treebank

CTB: overview

- Website: http://verbs.colorado.edu/chinese
- Started in 1998 at Penn, later in CU and Brandeis Univ.
- Supported by DOD, NSF, DARPA
- Latest version, v7.0, 1.2M-word Chinese corpus
  - Segmented, POS-tagged, syntactically bracketed
  - Phrase structure annotation
  - Inter-annotator agreement: 94%
  - On-going expansion, another 1.2M words planned
- Additional layers of annotation
  - Propbank/Nombank, Discourse annotation

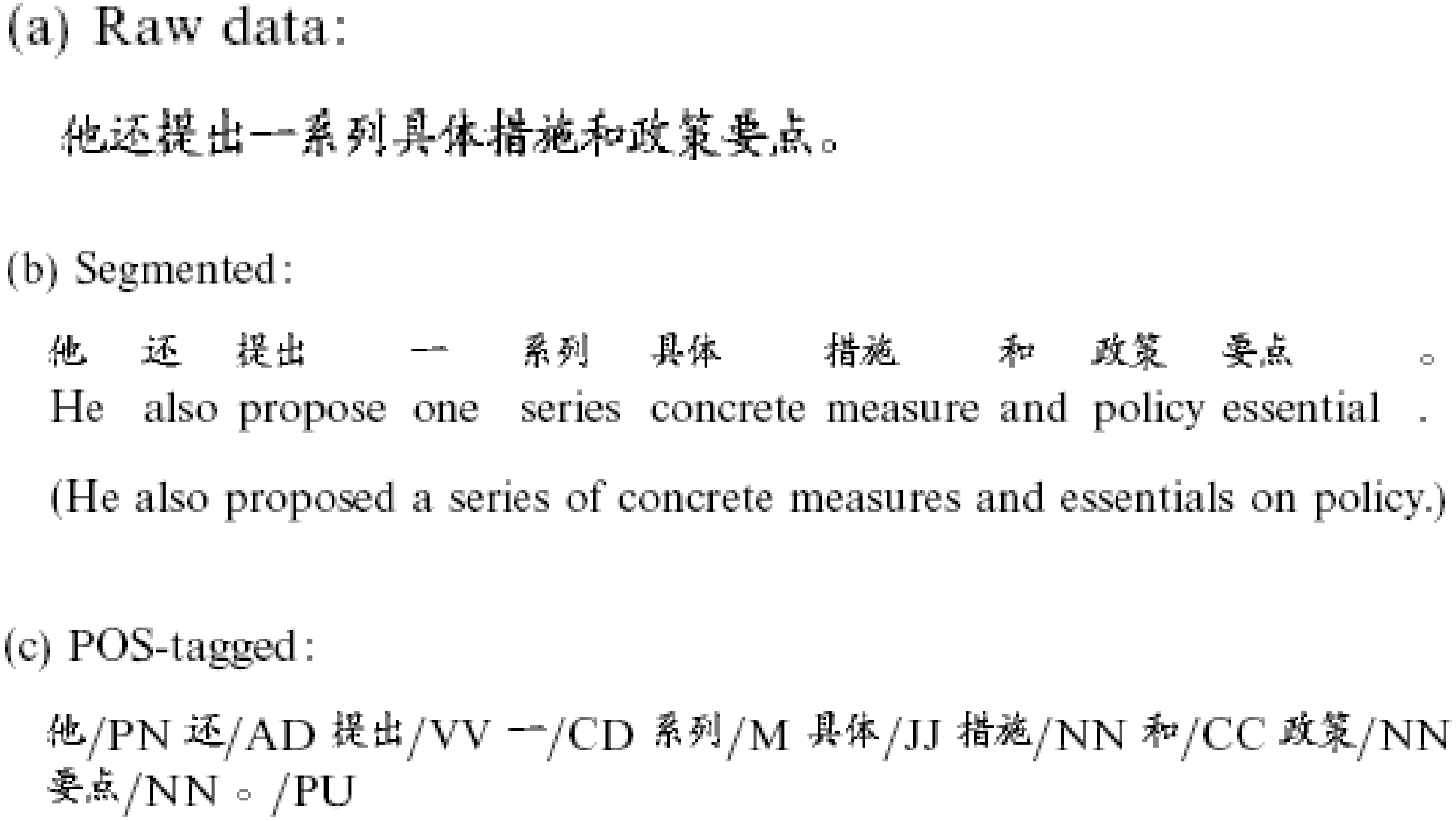
Timeline

- Stage 1 (4/1998-9/1998): get started
  - 4/98: meeting with a funding agency
  - 7/98: the first workshop
    - Existing annotation guidelines
    - Community needs
  - 9/98: form a team:
    - team leader
    - guideline designers
    - linguist experts
    - annotators
  - ?/98: Get funding for annotating 100K words

Timeline (cont)

- Stage 2 (9/1998- early 2001): initial annotation
  - One of the guideline designers, Nianwen Xue, was also an annotator
  - finish three sets of annotation guidelines
  - preliminary release and 1st official release: CTB 1.0
  - Several workshops to get community feedback
- Stage 3 (early 2001 - now): more annotation:
  - syntactic treebank:
    - 100K words => 1.2M words
    - Domains: Xinhua News, Hong Kong data, Taiwan magazine, etc.
  - PropBank: finish 1.2M words
  - Discourse treebank: in process
  - The treebank has been used in numerous NLP studies.

A treebank example



(a) Raw data:

他还提出一系列具体措施和政策要点。

(b) Segmented:

他	还	提出	一	系列	具体	措施	和	政策	要点	。
He	also	propose	one	series	concrete	measure	and	policy	essential	.

(He also proposed a series of concrete measures and essentials on policy.)

(c) POS-tagged:

他/PN 还/AD 提出/VV 一/CD 系列/M 具体/JJ 措施/NN 和/CC 政策/NN 要点/NN 。/PU

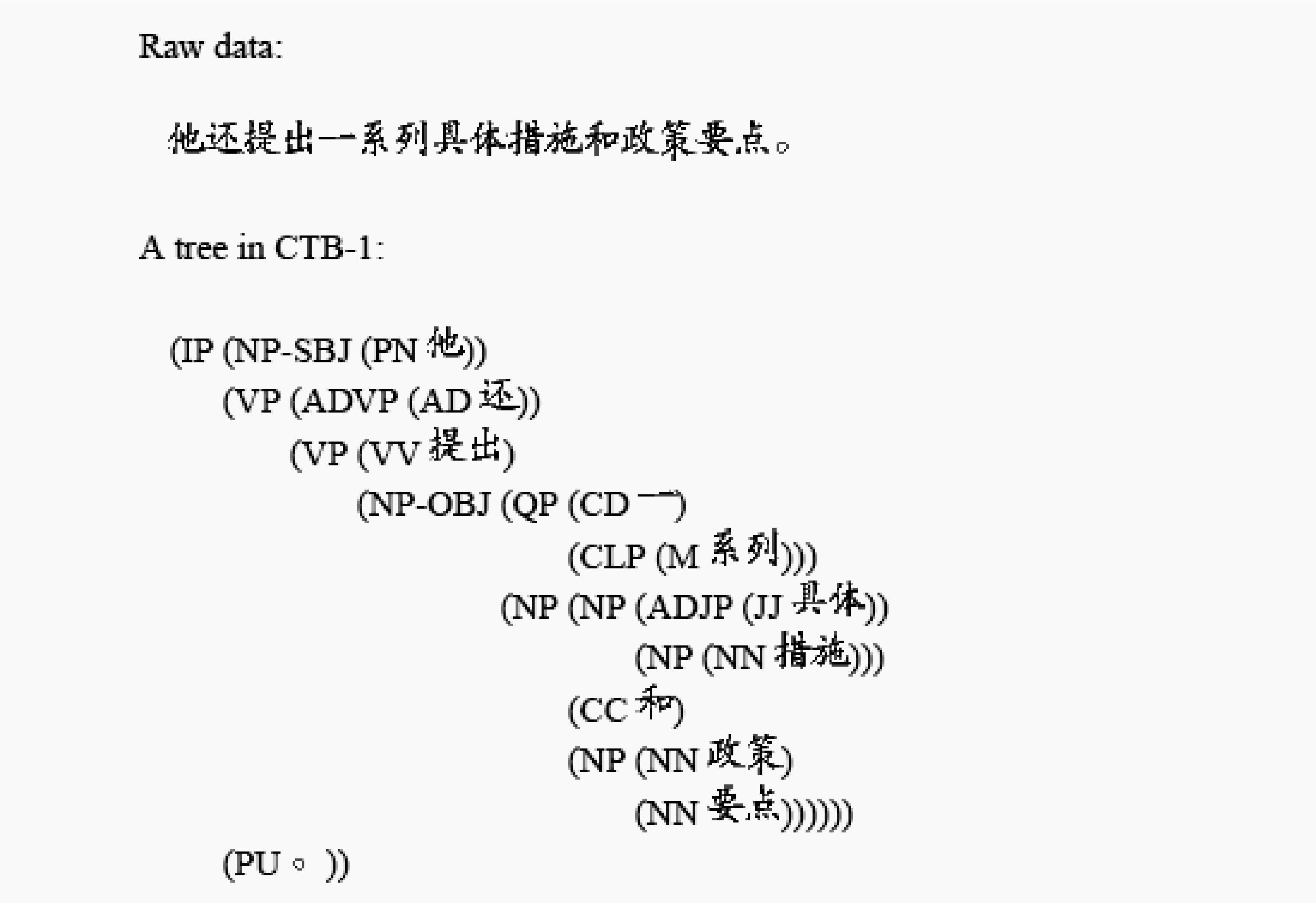
(d) Bracketed:

```
(IP (NP-SBJ (PN 他/he))
    (VP (ADVP (AD 还/also))
        (VP (VV 提出/propose)
            (NP-OBJ (QP (CD 一/one)
                        (CLP (M 系列/series)))
                    (NP (NP (ADJP (JJ 具体/concrete))
                                (NP (NN 措施/measure)))
                        (CC 和/and)
                        (NP (NN 政策/policy)
                            (NN 要点/essential))))))
    (PU 。))
```

CTB: Milestones

Version	Year	Quantity (words)	Source	Propbank/ Nombank	Discourse annotation
CTB1.0	2001	100K	Xinhua	yes	Pilot
CTB3.0	2003	250K	+HK News	yes	no
CTB4.0	2004	400K	+Sinorama	yes	no
CTB5.0	2005	500K	+Sinorama	yes	no
CTB6.0	2007	780K	+BN	yes	no
CTB7.0	2010	1.2M	+BC, WB	yes	no

An example



Raw data:

他还提出一系列具体措施和政策要点。

A tree in CTB-1:

```
(IP (NP-SBJ (PN 他))
    (VP (ADVP (AD 还))
        (VP (VV 提出)
            (NP-OBJ (QP (CD 一)
                            (CLP (M 系列)))
                        (NP (NP (ADJP (JJ 具体))
                                    (NP (NN 措施)))
                            (CC 和)
                            (NP (NN 政策)
                                    (NN 要点))))))
    (PU 。 ))
```

CTB-1

- The tasks:
  - Laying the good foundation for the whole project: creating guidelines, forming the team, getting feedback from the community, etc.
  - Annotating 100K-word Xinhua News

- Main steps:
  - Step 0 (6/98 - 8/98): Feasibility study
  - Step 1 (9/98 – 3/99): Word segmentation and POS tagging.
  - Step 2 (4/99 – 9/00): Bracketing
  - Step 3 (6/00 – 12/00): Preliminary release of CTB-1

The team for CTB1

- PIs: Martha Palmer, Mitch Marcus, Tony Kroch
- Project managers and guideline designers: Fei Xia, Nianwen Xue
- Annotators: Nianwen Xue, Fu-dong Chiou
- Programming support: Zhibiao Wu
- Linguistic consultants: Tony Kroch, Shizhe Huang

Community involvement

- Two workshops:
  - 06/1998: 3-day workshop at UPenn
  - 10/2000: 1-day workshop at Hong Kong (during ACL-2000)

- Three meetings:
  - 08/1998: At ACL-1998 in Montreal, Canada
  - 11/1998: At ICCIP-1998 in Beijing, China
  - 06/1999: At ACL-1999 in Maryland, US

- Two preliminary releases: in 6/2000 and 12/2000 by LDC

Challenges in designing guidelines for Chinese

- No natural delimiters between words in written text

- Very little, if any, inflectional morphology
  - Ex: No (explicit) tense, gender, person, number, agreement morphology

- Many open questions about syntactic constructions

- Little consensus on standards and analyses within the Chinese linguistics/NLP community

Guidelines

- word segmentation

- POS tagging

- Bracketing

Word segmentation

日文章鱼怎么说？

日文 章鱼 怎么 说 ？
Japanese octopus how say
“How to say octopus in Japanese?”

日 文章 鱼 怎么 说 ？
Japan article fish how say
“? How to say fish in Japanese articles?”

What is a word?

- Some examples:
  - name: “Hong Kong” vs. “London”
  - noun compound: “classroom” vs. “conference room”, “salesman” vs. “sales person”, “kilometer” vs. “thousand yards”
  - verb particle: “pick up”, “agree on”, “put off”, “give in”, “give up”
  - affix: “pro- and anti-government”, “ex-husband”, “former president”
  - hyphen: “e-file”, “parents-in-law” vs. “New York-based company”
  - punctuation: $50, 101:97
  - “electronic mail”, “e-mail”, “email”

- Anna Maria Di Sciullo and Edwin Williams, 1987. “On the definition of word”:
  - orthographic word: “ice cream” is two words
  - phonological word: e.g., I’ll
  - lexical item (or lexeme):
  - morphological object
  - syntactic atom: e.g., Mike’s book
  - …

How often do people agree?

- 100 sentences
- seven annotators
- no annotation guidelines are given
- pair-wise agreement:
  - Input: c1 c2 c3 c4 c5
  - sys: c1 c2 c3 | c4 c5
  - gold: c1 c2 | c3 | c4 c5
  - fscore = 2 * prec * recall / (prec + recall)
  - prec = ½, recall = 1/3, f-score = 0.4

Tests of wordhood

- Bound morpheme
- Productivity
- Frequency of co-occurrence
- Compositionality
- Insertion
- XP-substitution
- The number of syllables
- …

➔ None is sufficient.

Tests of wordhood

- Bound morpheme: e.g., “ex-husband”, “my ex”
- Productivity
- Frequency of co-occurrence: e.g., “one CL”
- Compositionality: e.g., kilometer
- Insertion: e.g., V1-not-V2
- XP-substitution
- The number of syllables
- …

➔ None is sufficient.

Our approach

- Choose a set of tests for wordhood
- Spell out the results of applying the tests to a string
- Organize the guidelines according to the internal structure of a string
  - Noun:
    - DT+N: e.g., ben3/this ren2/person ("I")
    - JJ+CD: e.g., xiao3/small sen1/three ("mistress")
    - N+N: e.g., mu4/wood xing1/star ("Jupiter")
    - V+N: e.g., zhen4ming2/proof xi4/letter ("certificate")
    - …
  - Verb:
    - reduplication: AA, ABAB, AABB, AAB, A-one-A, A-not-A, …
    - AD+V:
    - …

POS: verb or noun

美国 将 与 中国 讨论 贸易 赤字

U.S. will with China discuss/discussion trade deficit

“The U.S. will discuss trade deficit with China.”

美国 将 与 中国 就 贸易 赤字 进行 讨论

U.S. will with China regarding trade deficit engage discuss/discussion

“The U.S. will engage in a discussion on the trade deficit with china.”

Verb or preposition?

Google	用	30	亿	现金	收购	Double	Click
Google	use/with	30	100-million	cash	buy	Double	Click

Google used 3 billion cash to buy Double Click

Google bought Double Click with 3 billion cash

Main issue in POS tagging

Should POS tags be determined by distribution or by meaning?

Our approach:

- Use distribution (not meaning) for POS tagging
- Provide detailed tests for confusing tag pairs: e.g., (noun, verb)

Bracketing example:
Sentential complement or object control?

他	希望	她	做	作业
he/him	hope	she/her	do	homework

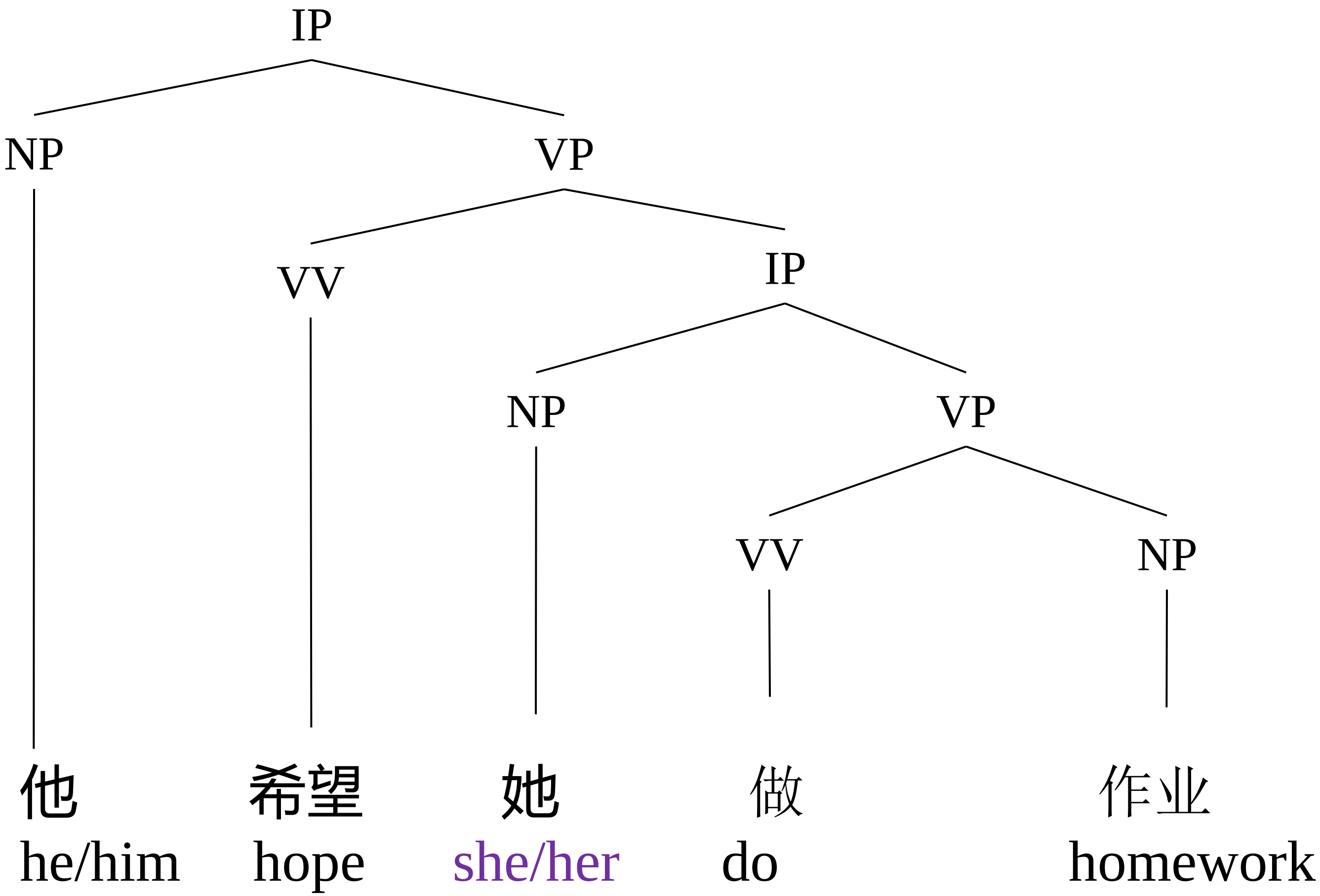
"He hopes that she will do her homework."

他	逼	她	做	作业
he/him	force	she/her	do	homework

"He forced her to do her homework."

NP	V	NP	V	NP

Sentential complement



IP

NP VP

VV IP

NP VP

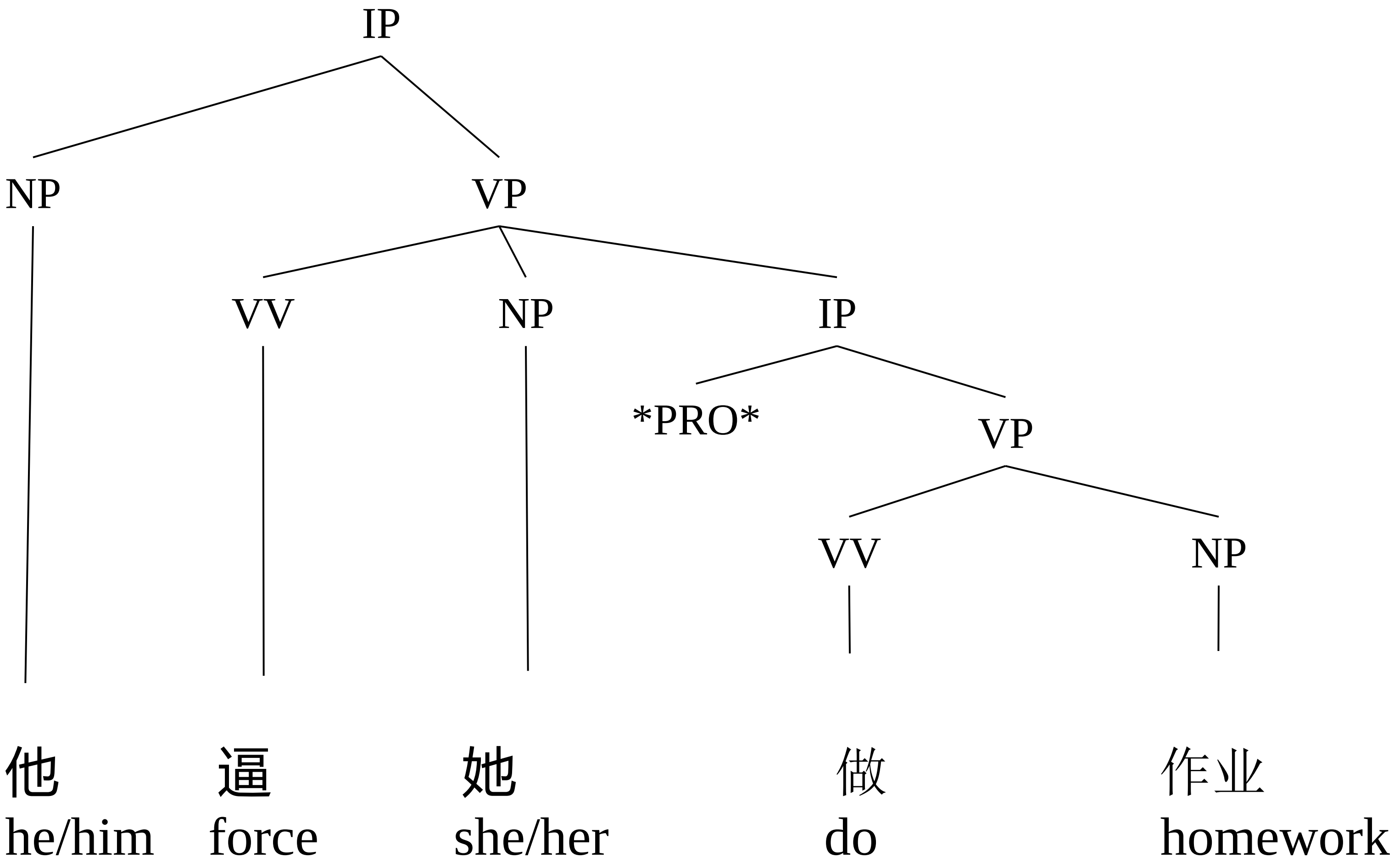
VV NP

他 希望 她 做 作业

he/him hope she/her do homework

“He hopes she will do her homework.”

Object control



IP

NP VP

VV NP IP

PRO VP

VV NP

他	逼	她	做	作业
he/him	force	she/her	do	homework

“He forced her to do her homework.”

Tests for sentential complement vs object control

For verb v1 in “NP1 v1 NP2 v2 NP3”:

- Can it take an existential construction as its complement?
- Can it take an idiom as its complement?
- Can it take a BEI construction as its complement?
- Can it take a topic construction as its complement?
- Can the complement clause have an aspectual marker?

Yes → Sentential complement

No → Object control

Good annotation guidelines

- Correctness / plausibility
- Convertibility
- Consistency
- Searchability
- Wide coverage
- Annotation speed

Revision of guidelines

- First draft before annotation starts

- Second draft after the 1st pass of annotation

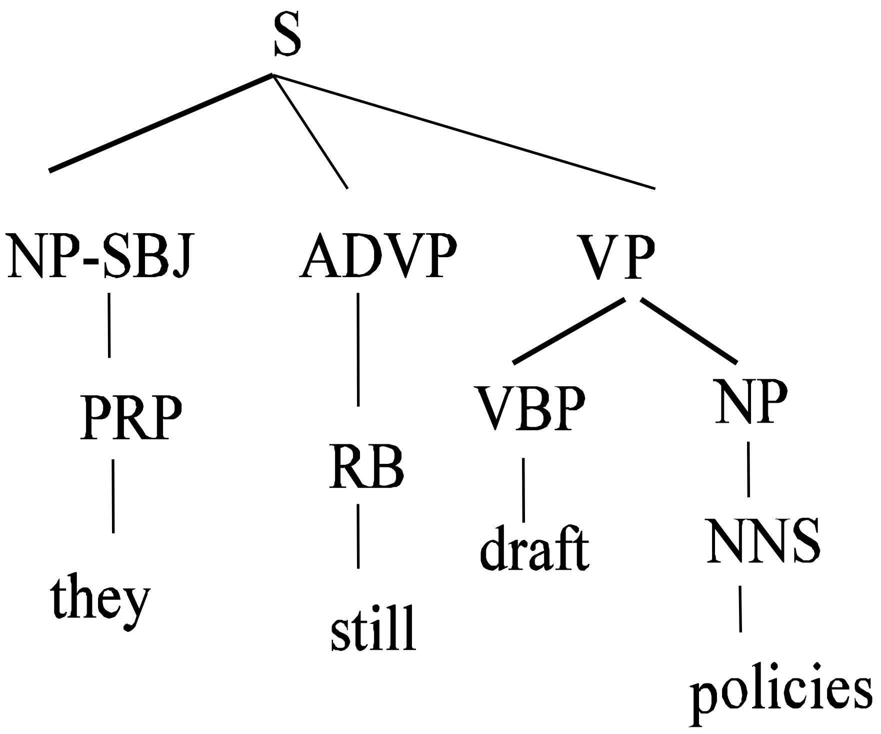
- Final version after the 2nd pass of annotation

- Three sets of guidelines
  - Segmentation: 31 pages
  - POS tagging: 44 pages
  - Bracketing: 191 pages

Quality control

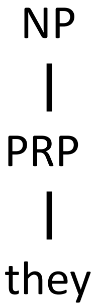
- Inter-annotator agreement:
  - Double annotation:
  - Inter-annotation agreement: 94%
  - Compared against the gold standard: 95-99%

The treebank tree

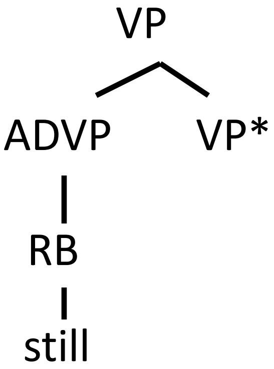


Extracted grammar

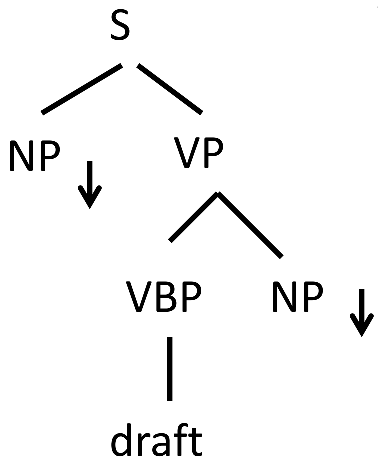
#1:



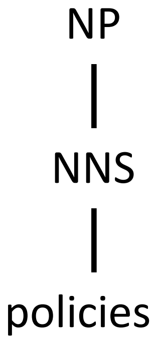
#2:



#3:



#4:



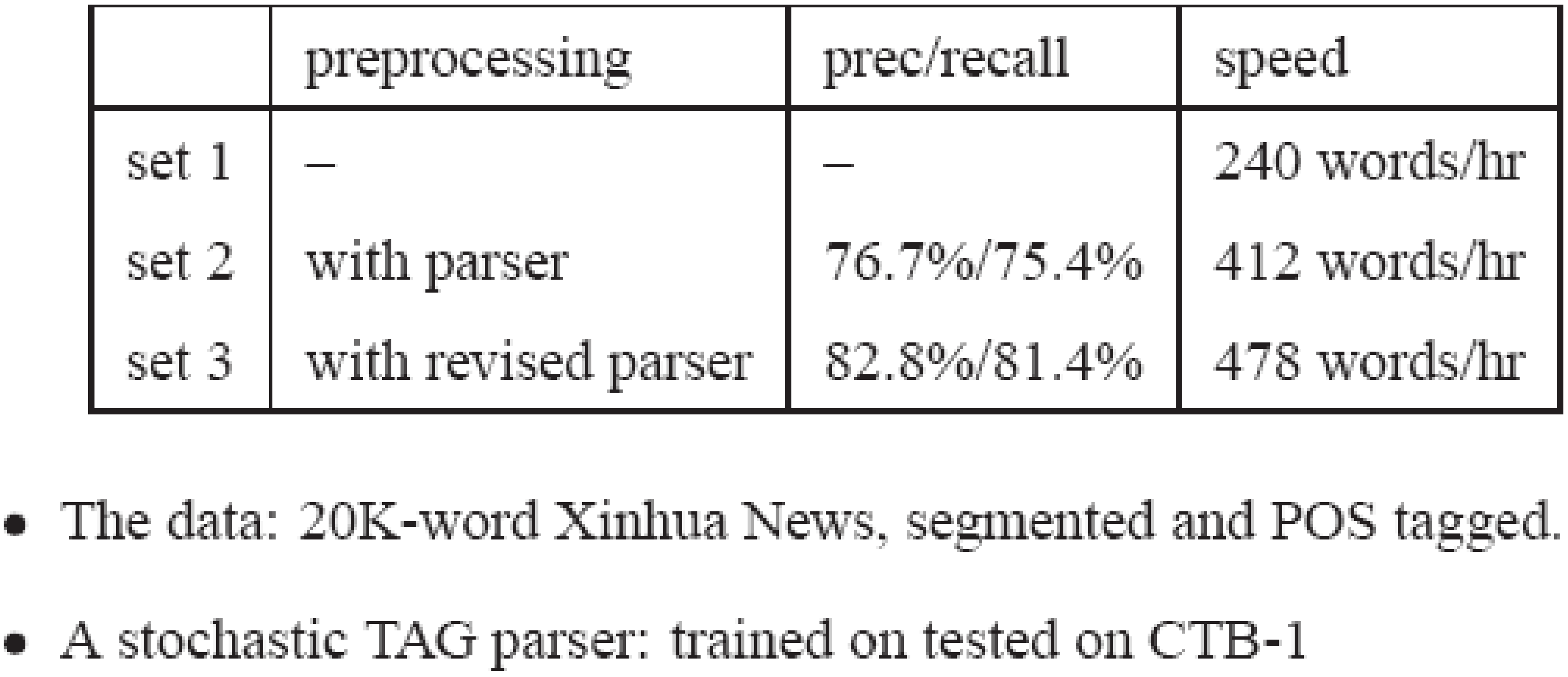
Detecting annotation errors using NLP tools

- A tool, LexTract, that extracts tree-adjoining grammars from treebanks

- Experiments:
  - run LexTract on the treebank and get a grammar G
  - mark each “rule” in G as correct or incorrect
  - correct trees in the treebank that generate the wrong “rules” in G

- Results:
  - Detect about 550 errors in CTB-1
  - A good grammar with statistical info

Preprocessing



	preprocessing	prec/recall	speed
set 1	–	–	240 words/hr
set 2	with parser	76.7%/75.4%	412 words/hr
set 3	with revised parser	82.8%/81.4%	478 words/hr

- The data: 20K-word Xinhua News, segmented and POS tagged.
- A stochastic TAG parser: trained on tested on CTB-1

Uses

- Segmentation
  - International Chinese word segmentation bake-offs: (2003, 2005, 2006, 2008)
- POS tagging
  - Tseng et al 2005, Hillard et al 2006, Xia and Cheung 2006, …
- BaseNP chunking
  - Liang et al 2006, Xu et al 2006, Chen et al 2006…
- Empty category recovery
  - Zhao and Ng 2007

More on uses

- Constituent structure parsing
  - Chiang and Bikel 2002, Levy and Manning 2003, Luo 2003, Hearne and Way 2004, Bikel 2004, Xiong et al 2005, Bod 2006, …

- Dependency structure parsing
  - Ma et al 2004, Jin et al 2005, Cheng et al 2006, Xu and Zhang 2006,Duan et al 2007, Wang 2007, Wang, Lin and Schuurmans 2007, Nivre 2007,…

More on uses

- Grammar extraction
  - Xia et al 2000; Burke et al 2004; Guo et al 2007
- Classifier Assignment
  - Guo and Zhong 2005
- Machine Translation
  - Wang, Collins and Koehn 2007,

The formation of SIGHAN

- A special interest group of ACL, formed in 2000
- A direct result of the two Chinese NLP workshops and three meetings in 1998-2000.
- 6 SIGHAN workshops, 4 bakeoffs so far
- A community consisting of researchers from all over the world

Chinese PropBank (CPB)

Version	CPB 1.0	CPB 2.0
CTB version	CTB 5.0	CTB 6.0
Date	2005	2008
Words	250K	500K
Total verbs framed	4,865	11,171
Total framesets	5,298	11,776

Future expansion

- Discourse relations
  - Pilot study (Xue 2005)
  - Need to start with sense tagging of discourse connectives
- Temporal and event

Conclusion

Annotation procedure

- Selecting data
- Creating guidelines
- Training annotators

- Tokenization / Word segmentation
- POS tagging
- Bracketing

- Quality control
- Preliminary and final release

➔ Use preprocessing tools to speed up annotation.
➔ Revision is needed at various stages

Lessons learned from treebanking

- Good annotation guidelines:
  - A treebank should be informative, and the annotation should be accurate and consistent.
  - More interaction is needed between grammar development and treebank development.

- Good, trained people:
  - Linguists for guideline design
  - Computational linguists for preprocessing and system support
  - Well-trained annotators
  - The large community for feedback

Lessons learned (cont)

- Quality control
  - Routine double annotation
  - Tools for detecting annotation errors
  - Feedback from parsing, PropBank, etc.

- Use of NLP tools
  - Preprocessing speeds up annotation, but could potentially biases the treebank.
  - Other tools: search, conversion, etc.

- There should be more coordination between different layers of annotation (e.g., treebank and PropBank)

The next step

- To build a multi-representational, multi-layered treebank

- Advantages:
  - It contains multiple layers: DS, PS, and PB
  - Certain annotation can be generated automatically (e.g., DS => PB, and DS => PS)
  - “Inconsistency” can be detected and resolved

- Disadvantages:
  - Coordination between various layers