Brandeis University

# Treebanking Chinese Text: What it's like

Nianwen Xue
Brandeis University

TLT'11
December 1, 2012

# Outline

- Overview
  - Background
  - Annotation procedure (old and new)
- Properties of Chinese and their impact on treebanking
  - Unreliable sentence boundaries
  - No word boundaries
  - Pervasive dropped elements
  - Fewer morpho-syntactic cues
- PTB and CTB
  - Similarities and differences

# Historical background

- Started in 1998 at Penn http://www.brandeis.edu/~clp/chinese
- Supported by DOD, NSF, DARPA
- Now a 1.5M word Chinese corpus
  - Word-segmented, POS-tagged, syntactically bracketed
  - Phrase structure annotation ***(Xue, Xia, Chiou, Palmer 2005)***
- Additional layers of annotation
  - Propbank/Nombank ***(Xue and Palmer 2009)***
  - Discourse annotation ***(Zhou and Xue, 2012)***
  - Temporal annotation ***(Xue and Zhou, 2010)***

# Contributors over the years

- Martha Palmer, now at University of Colorado
- Fei Xia, now at University of Washington
- Fu-Dong Chiou, now in Taiwan
- Zixin Jiang, retired from Microsoft, now at San Francisco
- Meiyu Chang, worked at Penn and Colorado
- Andi Wu, worked at Microsoft, now at GrapeCity
- Xiuhong Zhang, research associate at Brandeis
- Pengfei Li, PhD student at Boston University
- Nianwen Xue, now at Brandeis University

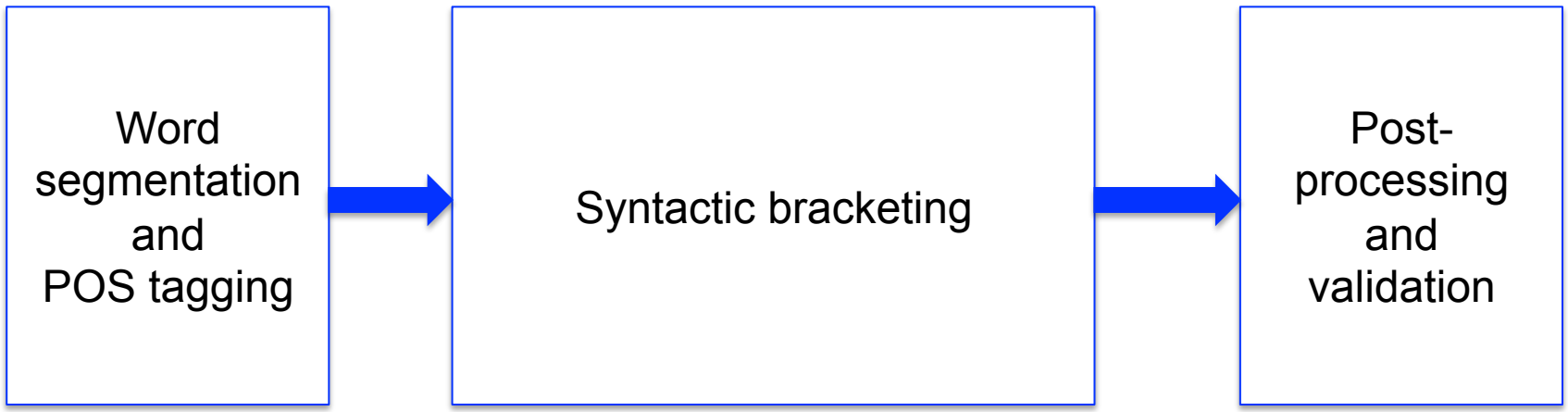
# Current treebanking team

- Xiuhong Zhang, PhD, Research Associate
  - In charge of syntactic bracketing
  - Pengfei Li, BU PhD student
- Andi Wu, PhD
  - In Charge of word segmentation and POS-tagging
  - Actual annotation performed by a company in China
- Nianwen Xue, PhD
  - Coordination

# CTB: Milestones

| Version | Year | Quantity (words) | Source | Propbank | 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,TC,WB | yes | no |
| CTB8.0* | 201?* | 1.6M | +DF | yes | no |

# Old Annotation Procedure



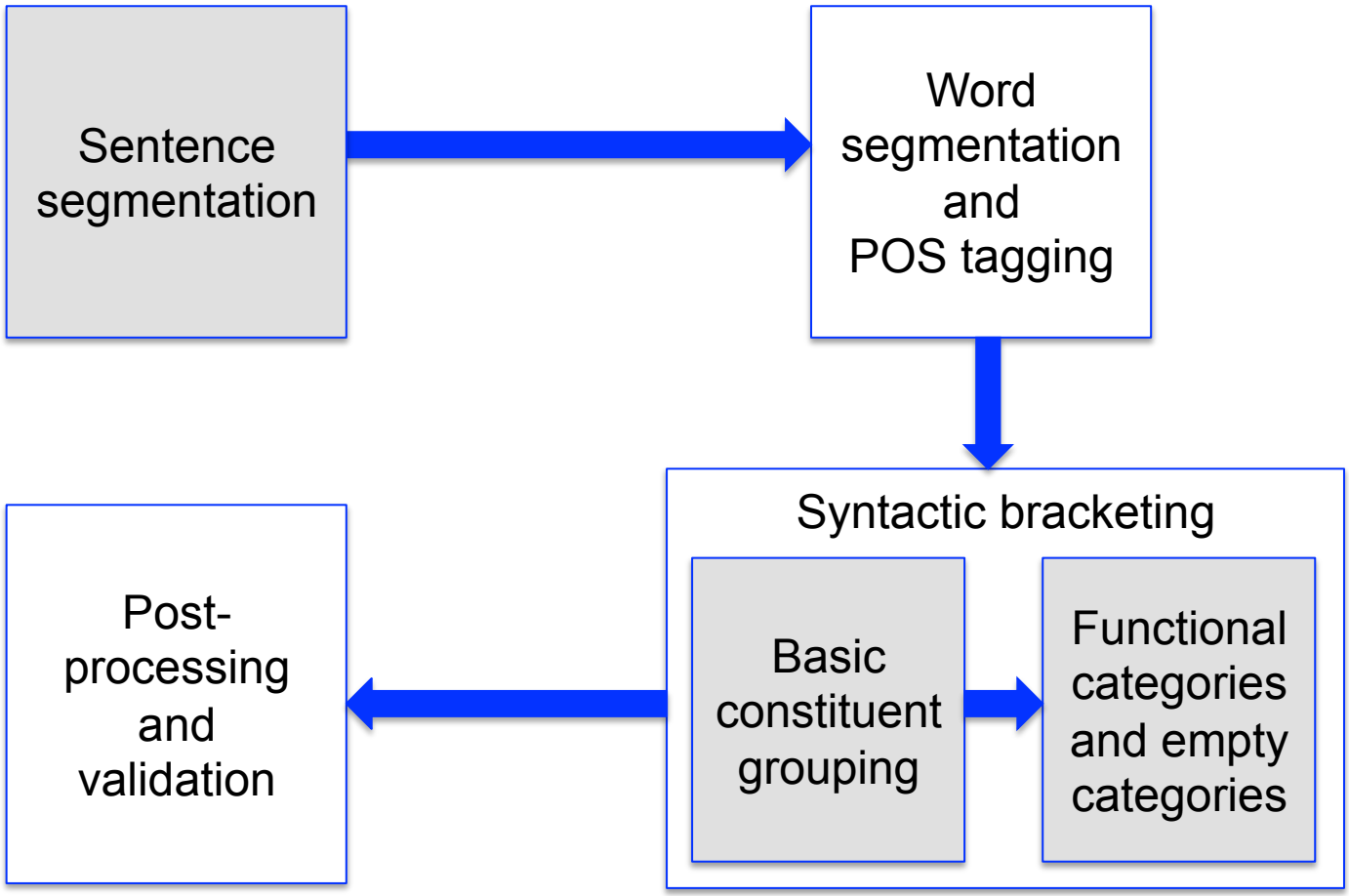
# Old Annotation Procedure

- Automatic preprocessing I:
  - Reformatting: converting source data to formats required by the tool
  - Automatic sentence boundary detection
    - Assumes that periods, question marks and exclamation marks are indicators of syntactic boundary
  - Automatic word segmentation and POS-tagging
- Manual correction of word segmentation and POS-tagging

# Old Annotation Procedure

- Automatic preprocessing II
  - Automatic parsing
  - Automatic repair and reformatting
- Manual syntactic bracketing
  - Making corrections to the parsing output
  - Manually adding functional tags and empty categories
- Post-processing and validation
  - Developing, adapting tools for automatic consistency checking

# New Annotation Procedure



# New Annotation Procedure

- Manual sentence boundary annotation (LDC)
  - Commas are ambiguous and can sometimes mark sentence boundaries *(Xue and Yang 2011, Yang and Xue 2012)*
- Decomposing the bracketing phase into smaller subtasks
  - Subtask 1: correcting the "geometry" of the parse trees
  - Subtask 2: adding the functional categories and empty categories
  - Redistribution of labor, a small step towards "crowdsourcing"

# Outline

- Overview
  - Background
  - Annotation procedure (old and new)
- Properties of Chinese and their impact on treebanking
  - Unreliable sentence boundaries
  - No word boundaries
  - Pervasive dropped elements
  - Fewer morpho-syntactic cues
- PTB and CTB
  - Similarities and differences

# Unreliable sentence boundary markers

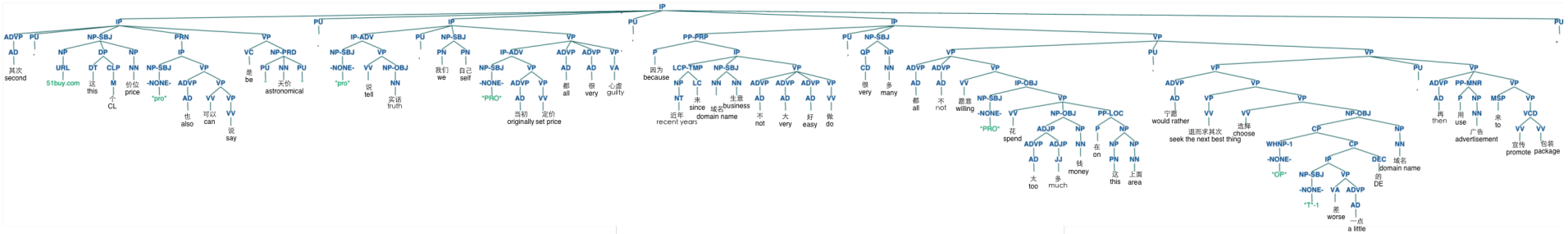
其次,这个价位也可以说是“天价”，说实话，我们自己当初定价都很心虚,因为近年来域名生意不大好做，很多客户都不愿意花太多钱在这上面，宁愿退而求其次选择差一点的域名,再用广告来宣传包装。

Second, this price can be said to be “astronomical”. To be honest, we felt unsure ourselves when we set the price, because the domain name business is not easy to do in recent years. Many customers don’t want to spend a lot of money on this. They would rather settle with the next best thing, choosing a domain name that is not as good, and then promote and package their products with advertisements.

# Unreliable sentence boundary markers

其次,这个价位也可以说是“天价”, 说实话, 我们自己当初定价都很心虚,因为近年来域名生意不大好做, 很多客户都不愿意花太多钱在这上面, 宁愿退而求其次选择差一点的域名,再用广告来宣传包装。

Second, this price can be said to be “astronomical”. To be honest, we felt unsure ourselves when we set the price, because the domain name business is not easy to do in recent years. Many customers don’t want to spend a lot of money on this. They would rather settle with the next best thing, choosing a domain name that is not as good, and then promote and package their products with advertisements.



# Dropped pronouns

这段时间✔一直在留意这款 nano 3, ✔还专门跑了几家电脑市场, 相比较而言, 卓越的价格算低的, 而且✔能保证✔是行货, 所以✔就下了单。

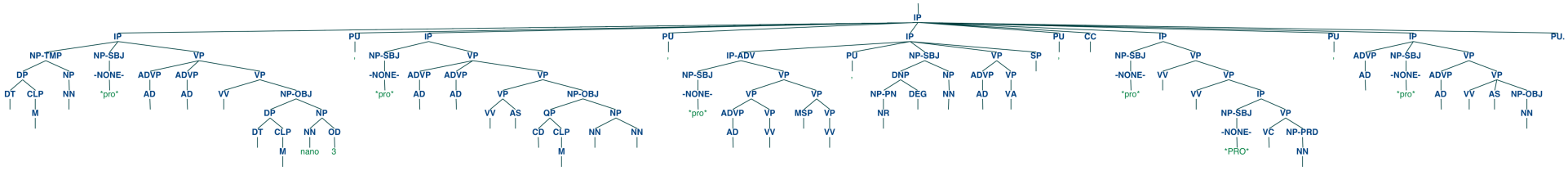
✗ pay attention to this Nano 3 recently, ✗ even visit a few computer stores in person, comparatively speaking, Zhuoyue's prices be relatively low, and ✗ can also guarantee that ✗ be genuine, therefore ✗ place the order.

**I** have been paying attention to this Nano 3 recently, and **I** even visited a few computer stores in person. Comparatively speaking, Zhuoyue's prices are relatively low, and they can also guarantee that **their products** are genuine. Therefore **I** placed the order.

% of empty elements: CTB7.0=8.5 vs PTB = 6.75%

# Dropped pronouns

这段时间✔一直在留意这款 nano 3, ✔还专门跑了几家电脑市场, 相比较而言, 卓越的价格算低的, 而且✔能保证✔是行货, 所以✔就下了单。




**I** have been paying attention to this Nano 3 recently, and **I** even visited a few computer stores in person. Comparatively speaking, Zhuoyue's prices are relatively low, and they can also guarantee that **their products** are genuine. Therefore **I** placed the order.

% of empty elements: CTB7.0=8.5 vs PTB = 6.75%

# 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
“???”

# The “one word” or “two words” dispute

- 鲜花 “fresh flower” is one word
- *鲜红花 fresh red flower
- 强队 “strong team” is one word
- *强篮球队 strong basket ball
- 擦净 “wipe clean” is one word
- 擦 干净 “wipe clean” are two words
- 走出 “walk out” is one word
- 走 出去 “walk out” are two words

# No morphological markings for tense

- 他 这 学期 **教** 中文 。
  He this semester teach Chinese .
  “He is teaching Chinese this semester”
- 他 上 学期 **教** 中文 。
  He last semester teach Chinese .
  “He taught Chinese last semester.”
- 他 下 学期 **教** 中文 。
  He next semester teach Chinese .
  “He will teach Chinese next semester.”

# No morphological markings for number

- 他 买 了 一 本 书 。
  He buy ASP one CL book .
  "He bought one book."
- 他 买 了 两 本 书 。
  He buy ASP one CL book .
  "He bought two books."
- 他 买 了 一 千 本 书 。
  He buy ASP one thousand CL book .
  "He bought one thousand books."

- No need for morphological analysis, but…

# The “verb or noun” dispute

美国 将 与 中国 讨论 贸易 赤字 。
U.S. will with China discuss 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 discussion .
“The U.S. will engage in a discussion on the trade deficit with china.”

# The “Verb/preposition” dispute

Google 用 了 33 亿 现金 收购 Double Click
Google use ASP 33 billion cash buy Double Click
“Google used 33 billion cash to buy Double Click.”

Google 用 33 亿 现金 收购 了 Double Click
Google use 33 billion cash buy ASP Double Click
“Google bought Double Click with/using 33 billion cash.”

# Comparison of POS tagsets

| PTB | | CTB | |
|---|---|---|---|
| VB | Base form | VA | predicate adjective |
| VBD | Past tense | VE | existential |
| VBG | gerund | VC | copula |
| VBN | Past participle | VV | other |
| VBP | Non-3rd per. sing. pres. | | |
| VBZ | 3rd per. sing. | | |
| NN | Sing. Or mass | NT | Temporal noun |
| NNS | Noun, plural | NR | Proper noun |
| NNP | Proper noun, singular | NN | Other noun |
| NNPS | Proper noun, plural | | |

No TO infinitive marker, determiner limited

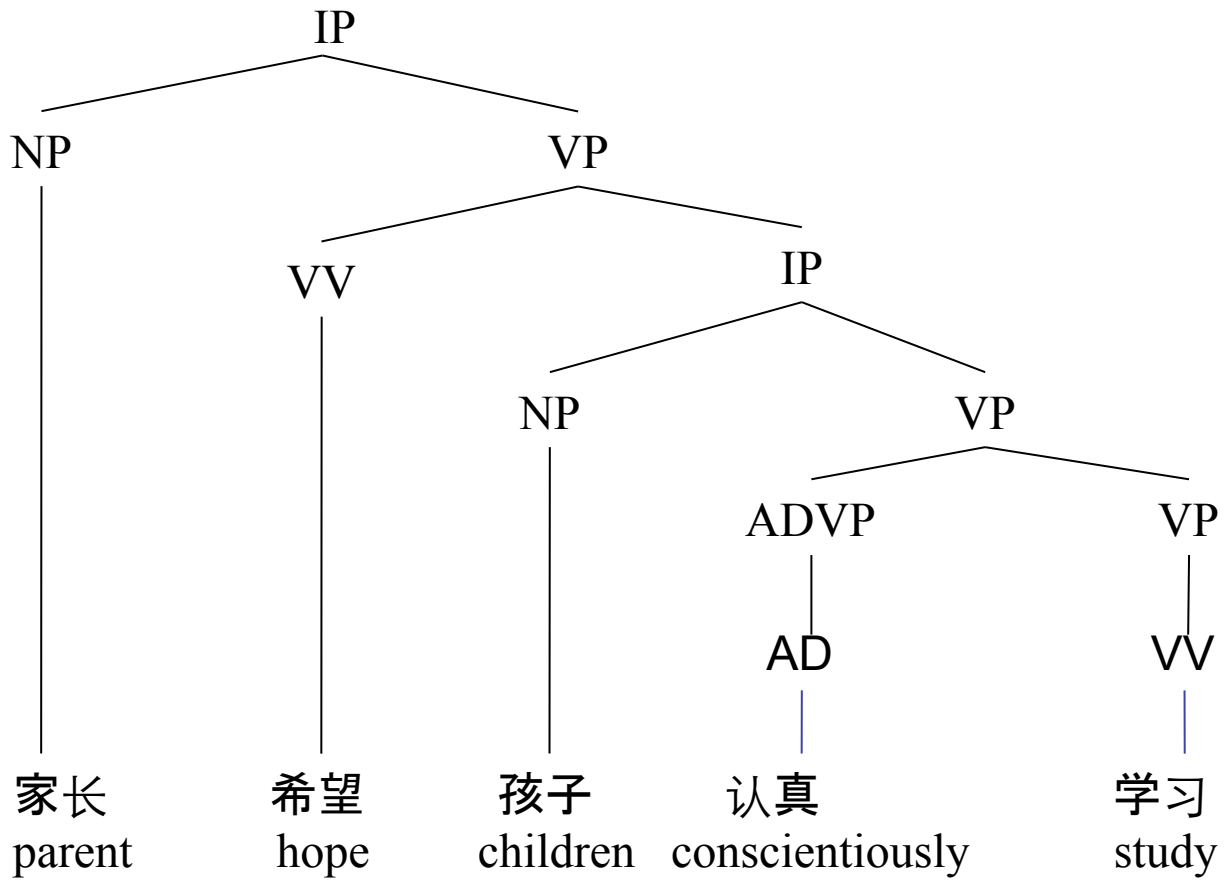
# Sentential complement or object control?

NP V NP ADV V

家长 希望 孩子 认真 学习
parent hope children conscientiously study
“Parents hope **that** their children will conscientiously study.”

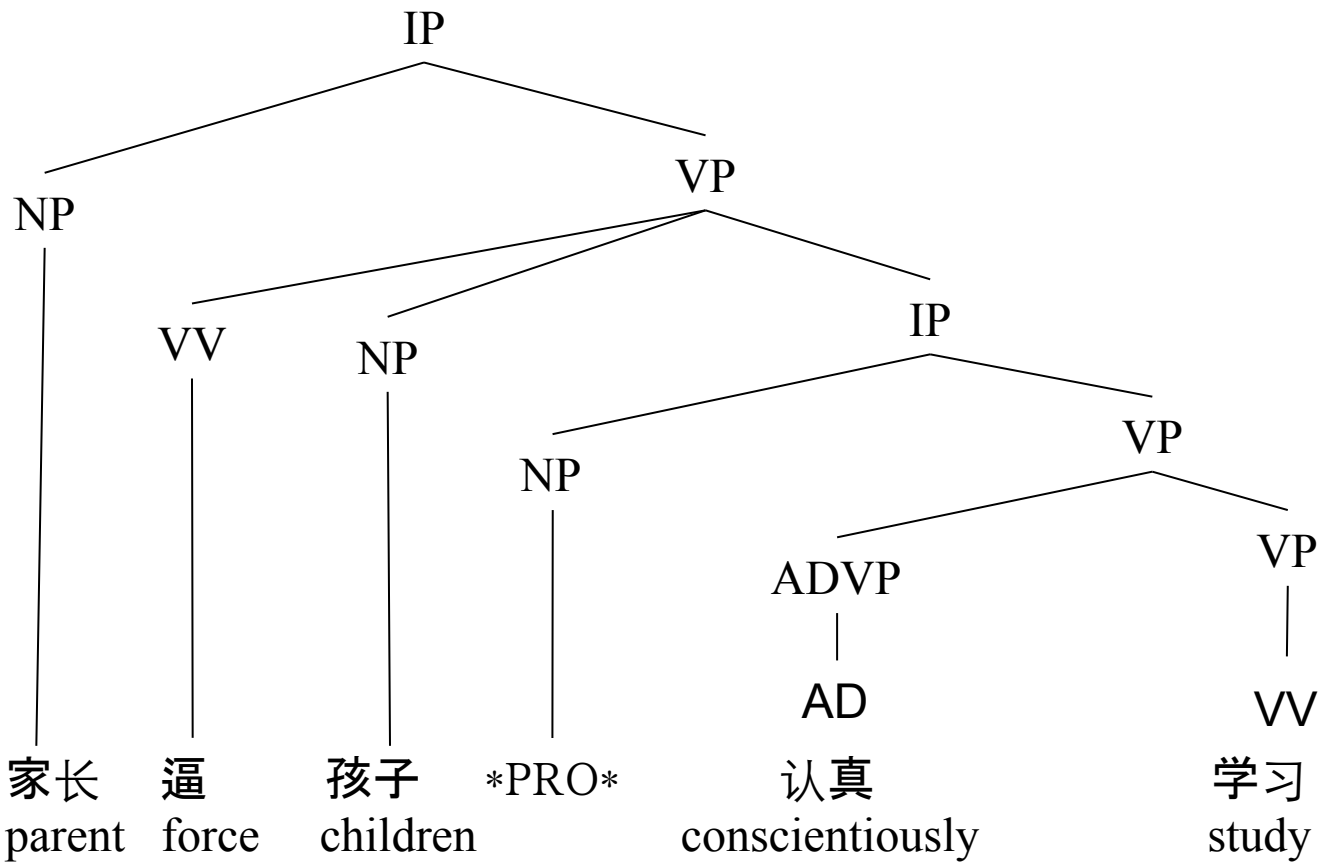
家长 逼 孩子 认真 学习
parent force children conscientiously study
“Parents force their children **to** study conscientiously.”

# Sentential complement




"Parents hope that their children will conscientiously study."

# Object control




“Parents force their children to study conscientiously.”

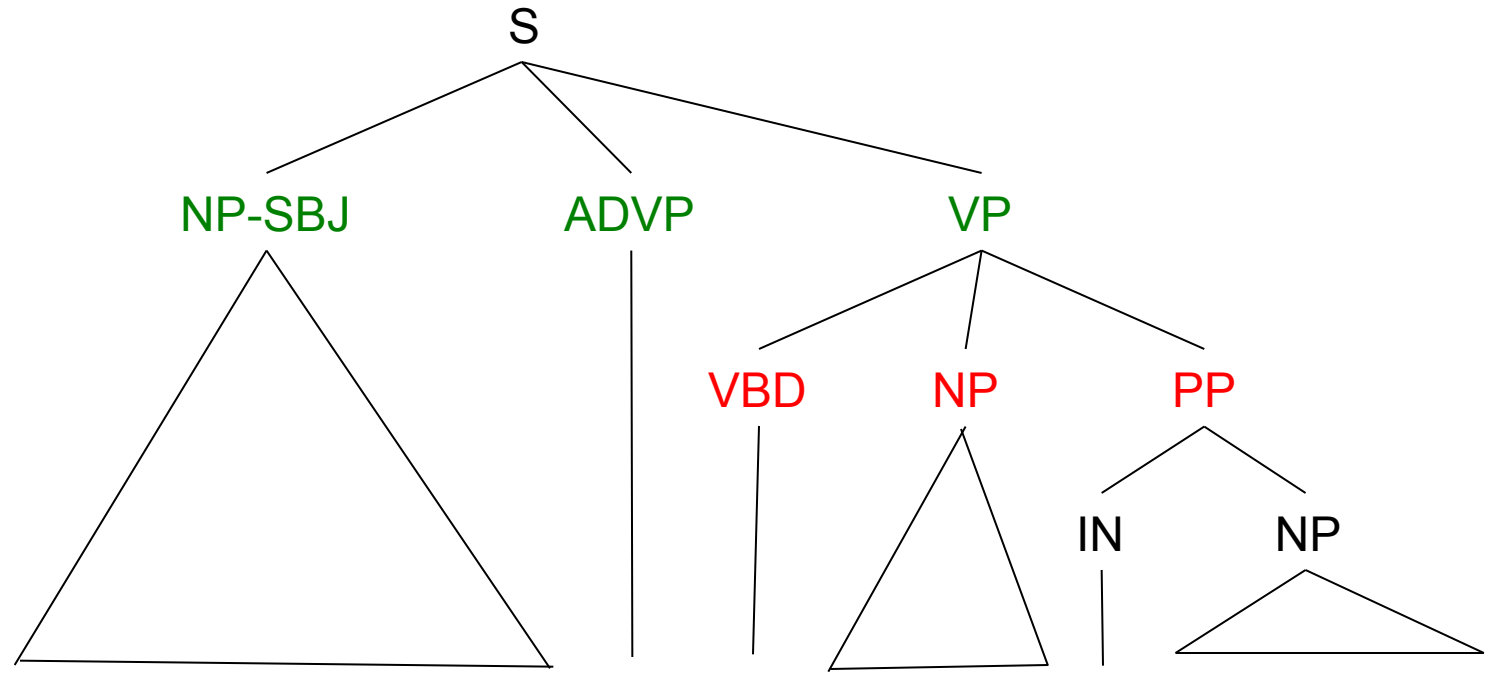
# Words of advice

- Next time when you're annoyed that you have to do morphological analysis or lemmatization, think about what happens if you don't have these little words or word pieces. You'll feel a lot better.

# PTB and CTB

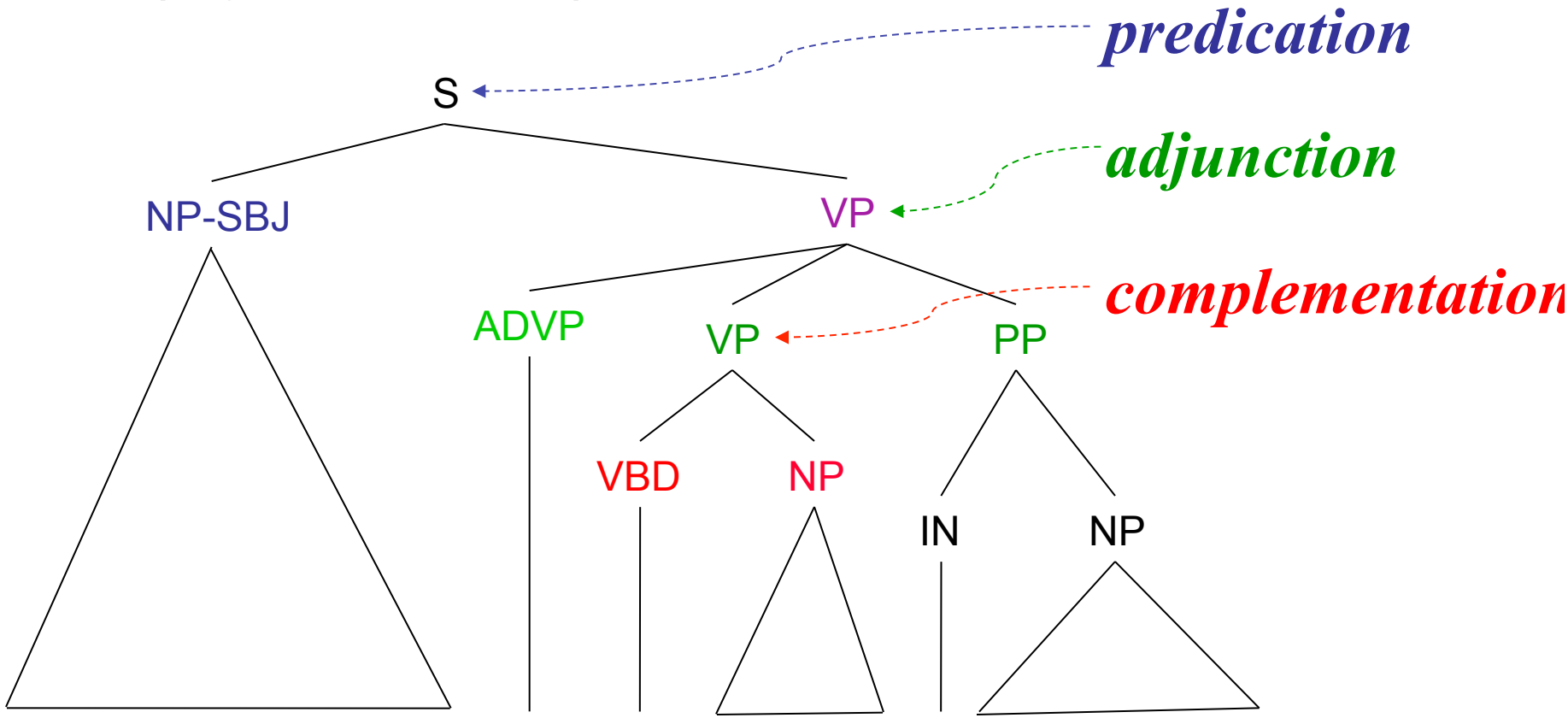
- Similarities
  - Same formal devices: labeled brackets, functional categories, empty categories and co-indexation
  - X-bar framework
- Differences
  - Trivially:
    - IP = S, CP = SBAR
  - Substantively…

# A Penn Treebank example




The Mortgage and equity last paid a dividend on August 1, 1988
real estate investment trust

# (Hypothetical) CTB annotation




The Mortgage and equity last paid a dividend on August 1, 1988 real estate investment trust

One grammatical relation per bracket

# NP-internal structures

*Co-ordination*

(NP (NN kidney)
(, ,)
(NN liver)
(, ,)
(NN heart)
(CC and)
(NN pancreas)
(NNS transplants))

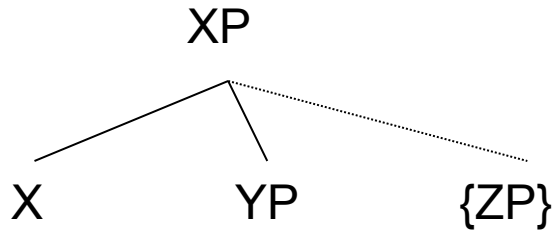
*adjunction*

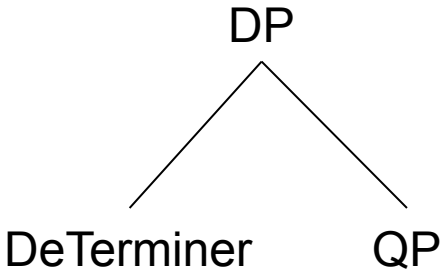
(**NP** (**NP** (NN kidney)
(, ,)
(NN liver)
(, ,)
(NN heart)
(CC and)
(NN pancreas))
(NP (NNS transplants)))

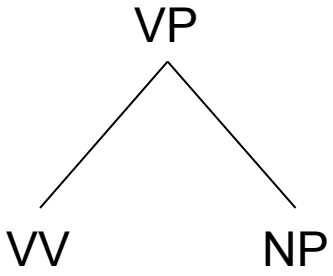
# Consequences of the change

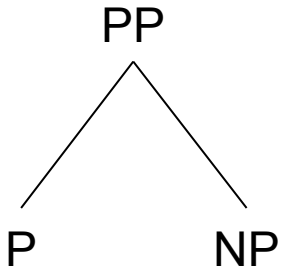
- Reduces the possible structures to a few primitive structures, and thus reduces the cognitive loads on the treebankers ***(ITA 94%, Xue et al 2005)***
- Supports tree transformation, e.g., phrase structure to dependency conversion by implicitly marking the head for each constituent ***(CoNLL 2009 Shared Task)***

# Complementation

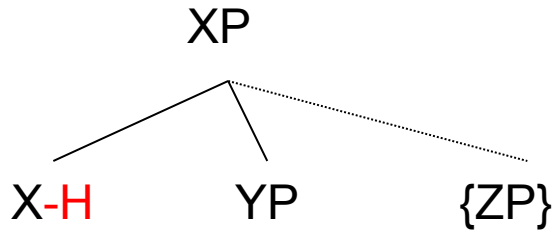


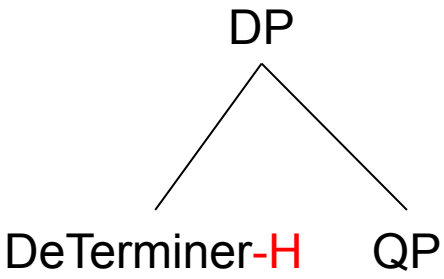



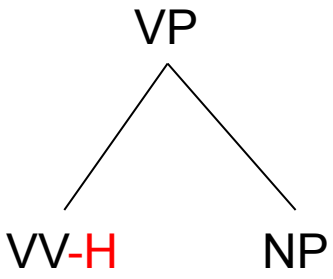



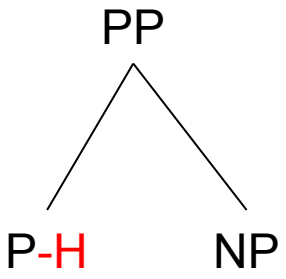



# Complementation

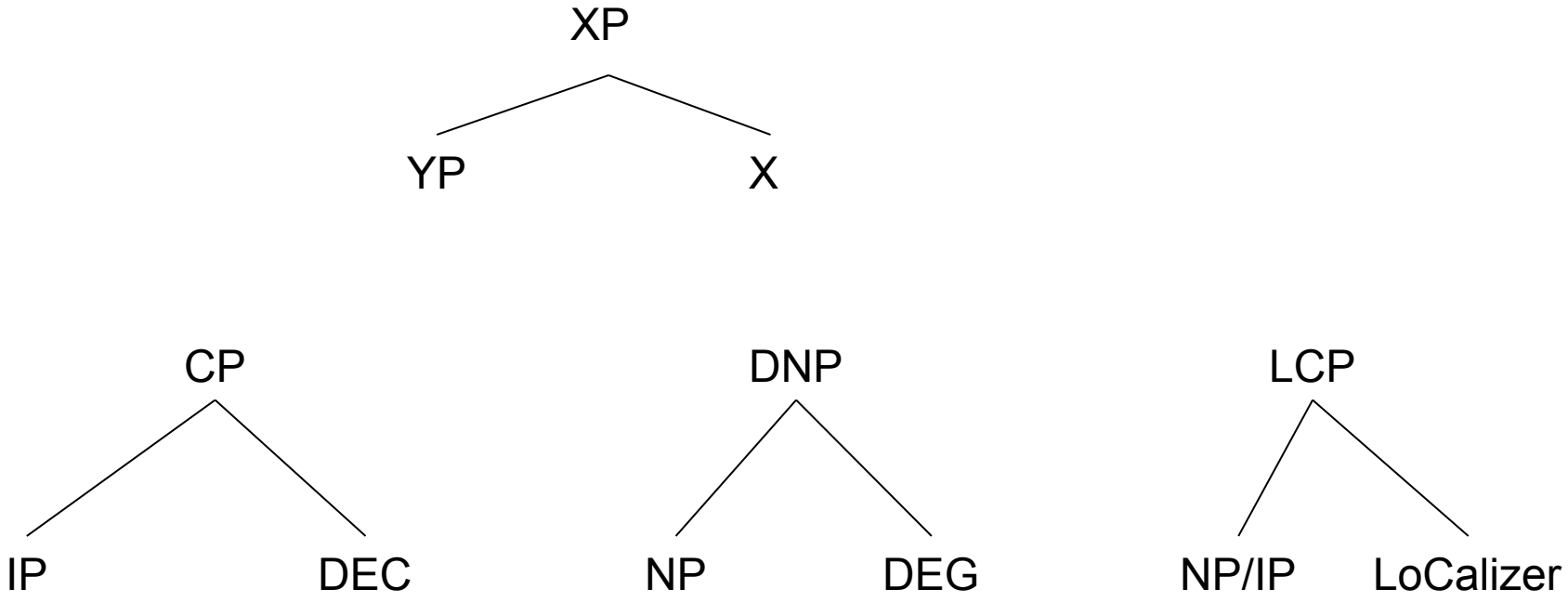


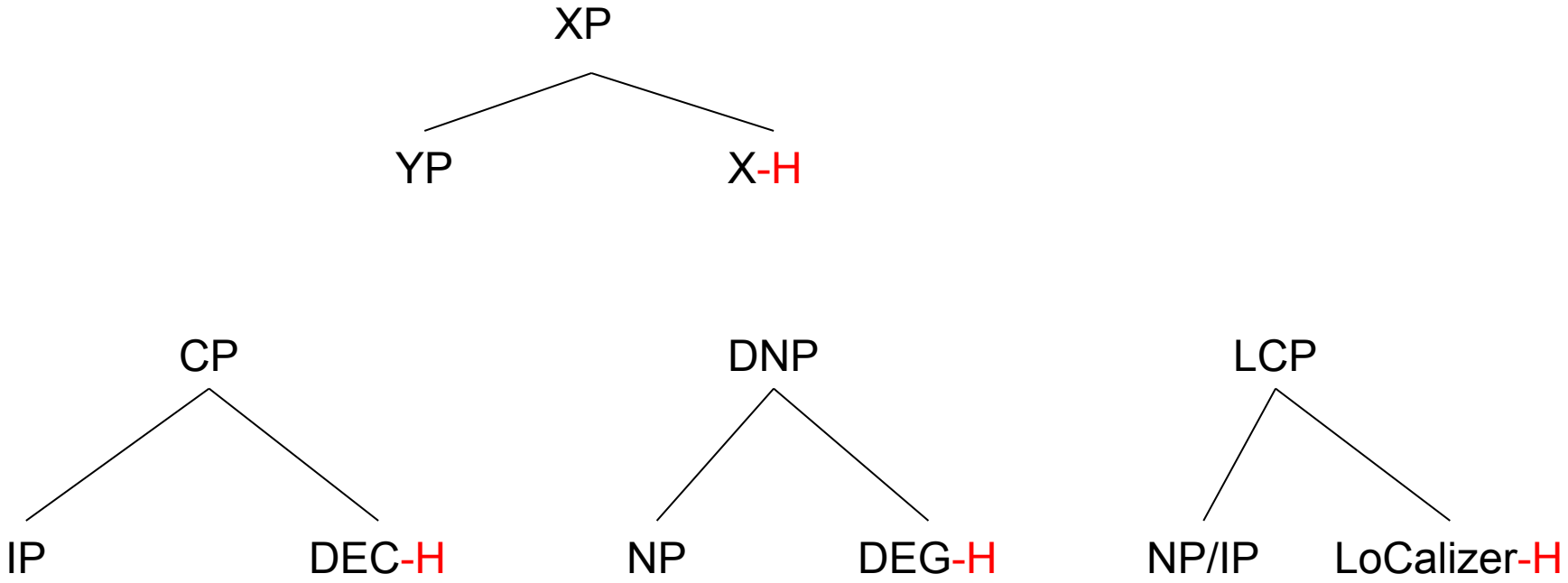



# Complementation (right-headed)

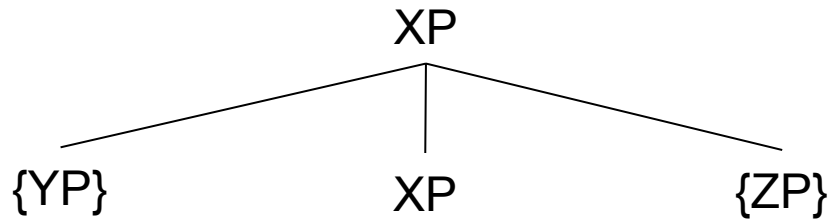


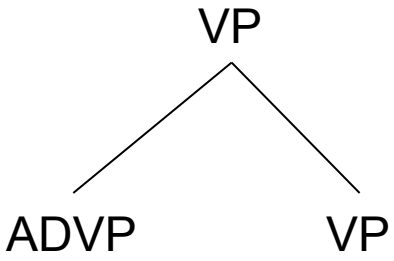
# Complementation (right-headed)

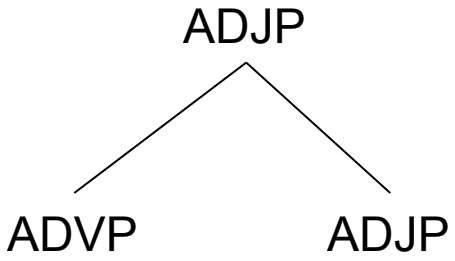



Chinese is mix-headed

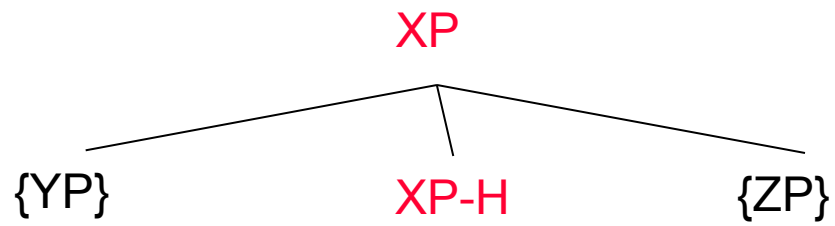
# Adjunction

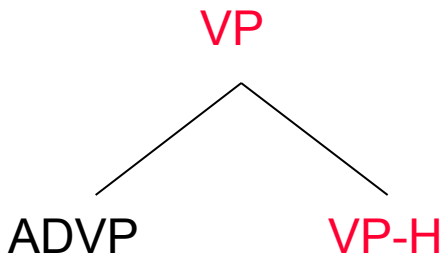


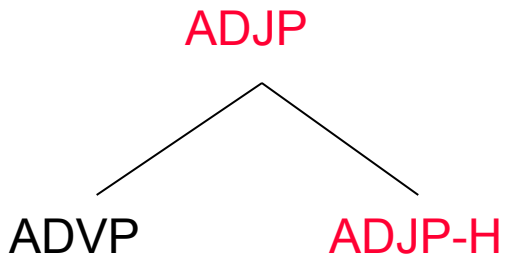



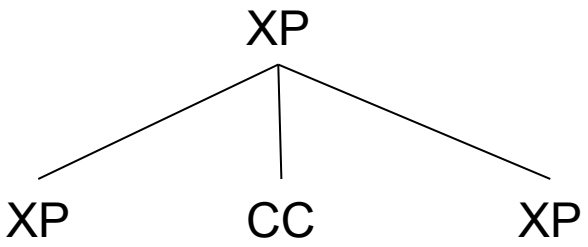
# Adjunction

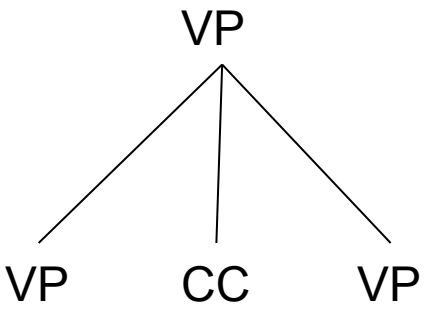


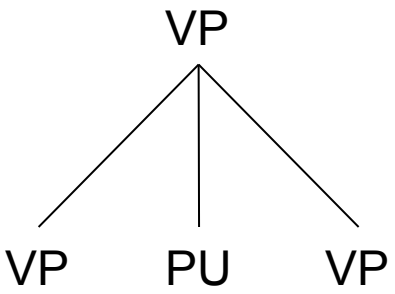



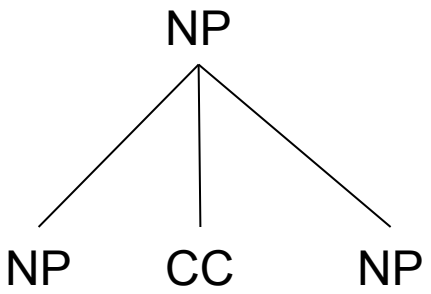



# Coordination

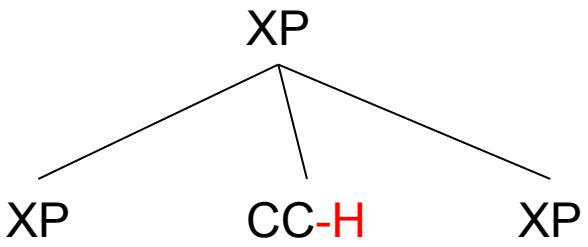


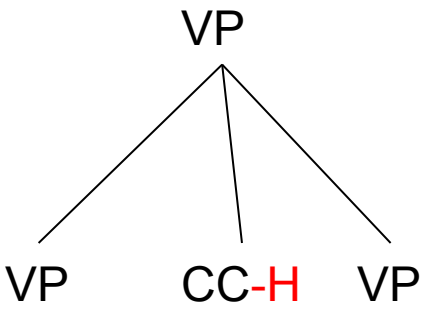



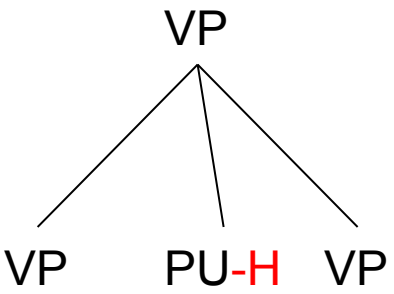



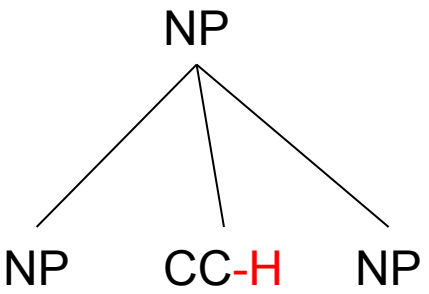



# Coordination



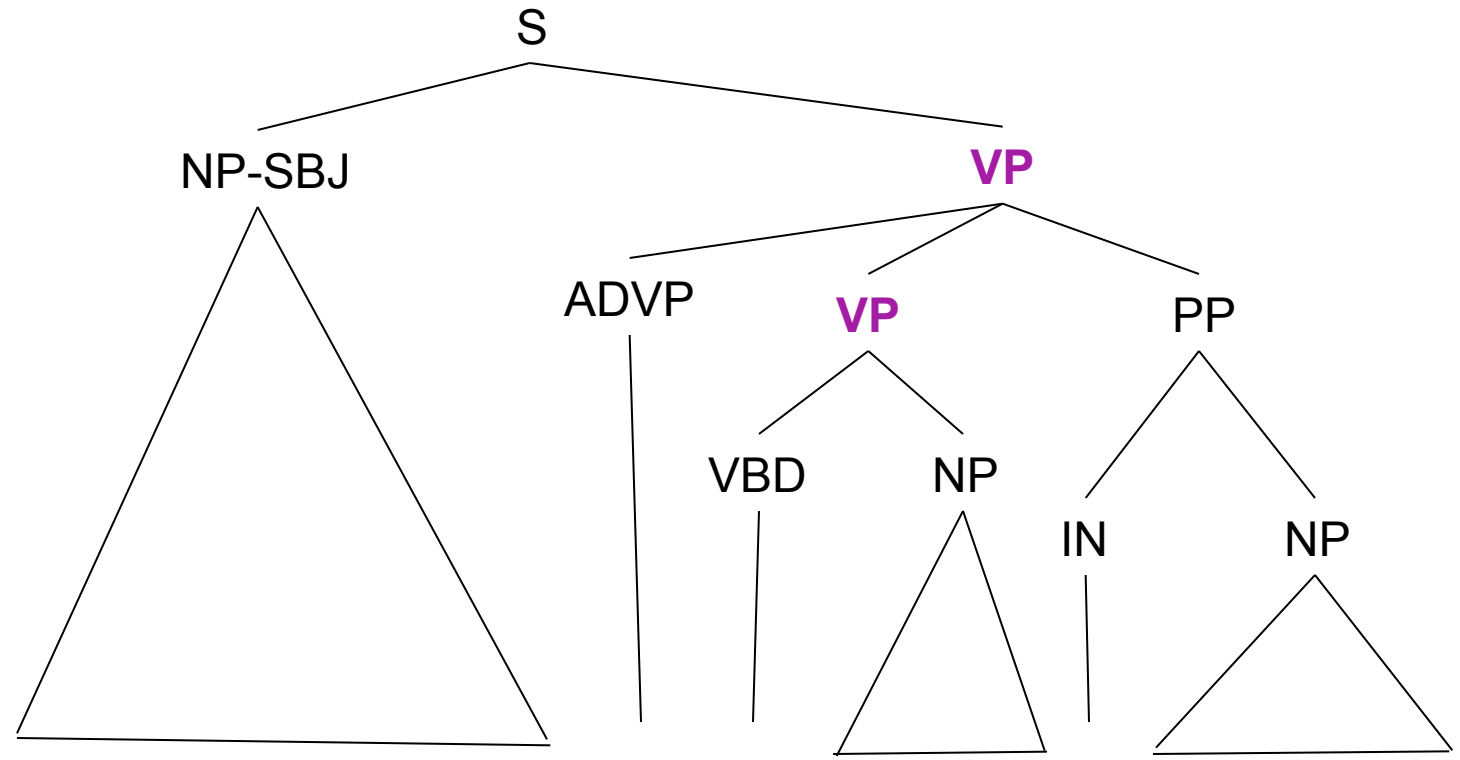



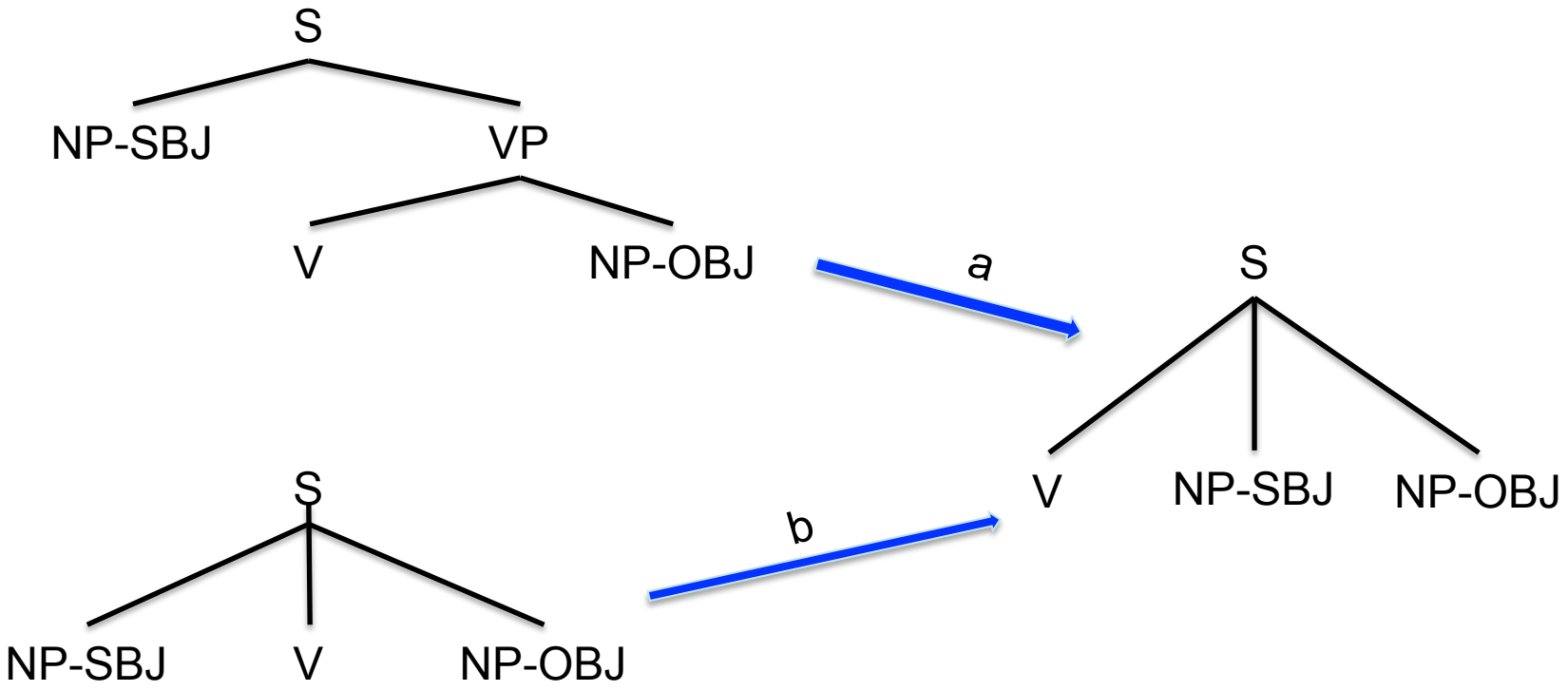
# Everybody will be happy, right?

- Levy and Manning. 2003. “Is it hard to parse Chinese, or the Chinese TreeBank?”. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics.* Sapporo, Japan

# Created recursive structures. Made it difficult to parse?



# "Deeper" not always better for all applications



**Wish list of machine learning types:**

Lots and lots of data quickly annotated:
Better data is more data

Intuitive annotation: no complicated
Linguistic concepts please!

Consistently annotated data: the kind
that I can report good scores on!

**Wish list of linguistics types:**

Data carefully annotated in a way that is conceptually elegant and theoretically satisfying

"Deep" annotation, reflecting abstractions and generalizations

Linguistically interesting data: the kind that I can write about!

**Everybody will be happy when:**

Lots and lots of data quickly annotated in a way that is theoretically satisfying

Intuitive “deep” annotation that comes with adapters

Linguists learn to use machine learning and machine learning types learn some linguistics

谢谢!

Thank you!