Lec16: Treebanks

HUL 242

17/3

Goals of X-bar Notation

capture intermediate elements
capture cross categorial similarity
find structural definitions for semantic notions

Not discussed..

Analyze particular linguistic phenomena: question formation, passivization, relative clauses ..
Cross-linguistic validity : can we apply it successfully across languages?

Goals of X-bar Notation

X': Elements larger than heads, but smaller than the entire XP

Shared constituent co-ordination

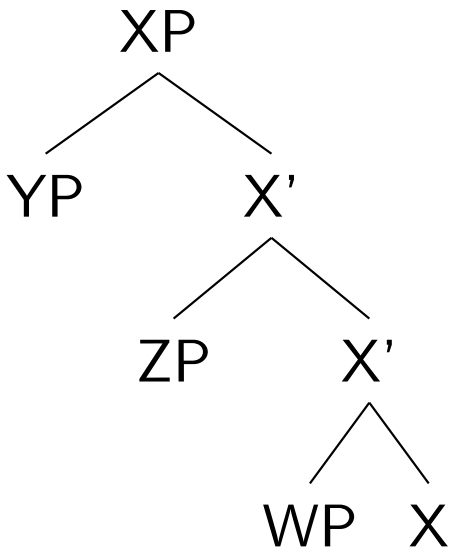
He [read the letters in the garden shed this afternoon $_{VP}$] and she *did so* too

He [read the letters in the garden shed $_{V'}$] this afternoon and she *did so* last night

He [read the letters$_{V'}$] in the garden shed this afternoon and she *did so* in the house last night

Old rule: VP $\rightarrow$ V NP (PP+)
New rule: VP $\rightarrow$ V' (PP)
V' $\rightarrow$ V (NP)

V' is smaller than a phrase, but larger than a head
represents successively larger expansions of a head



Note that X' notation imposes restrictions such that N' $\rightarrow$ V PP is *not* valid

Projections (X') should be of the same category as the head

Cross categorial similarity

Capture parallel structures across categories (VP, PP, NP, AdjP or CP)
All phrases are underlyingly similar with respect to their structure

Structural definitions for semantic notions

Semantic notion that complements are essential (when realized), adjuncts optional
Translate this semantic notion to a structural/'configurational' one

complement rule: sister to X
adjunct rule: sister to X'

Structural definitions

Extend the X bar notation to clausal types
Complement phrases, Tensed phrases
Complementizers are heads (C), sentences are complements
Modals are heads (T): verb phrases are complements

Constrain CFGs

Prevent overgeneration by linking to theta role/theta grid
Additional principles (endocentricity constraint)
Are CFGs adequate to capture natural languages?
Sometimes this depends on who you ask

Treebanks

Sentences annotated with syntactic structure
Early 1990s: English Treebank
Present: Arabic, Chinese, Dutch, Finnish, French, German, Hindi, Greek, Hebrew (many others)

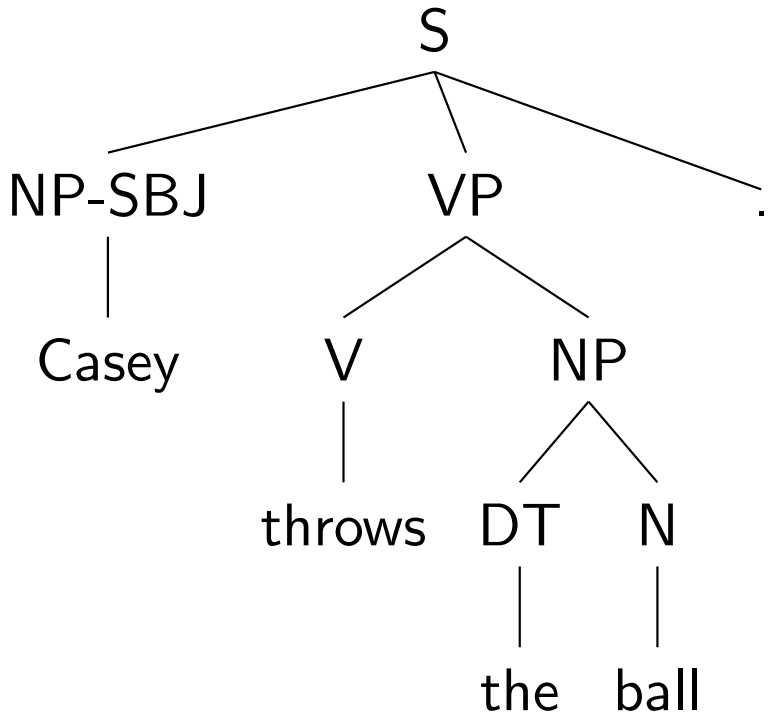
Annotation

Process of adding linguistic information to raw text: e.g. POS tags, parses, discourse relations etc.

Penn Treebank

One of the earliest English Treebanks
Wall Street Journal corpus of 1 million words

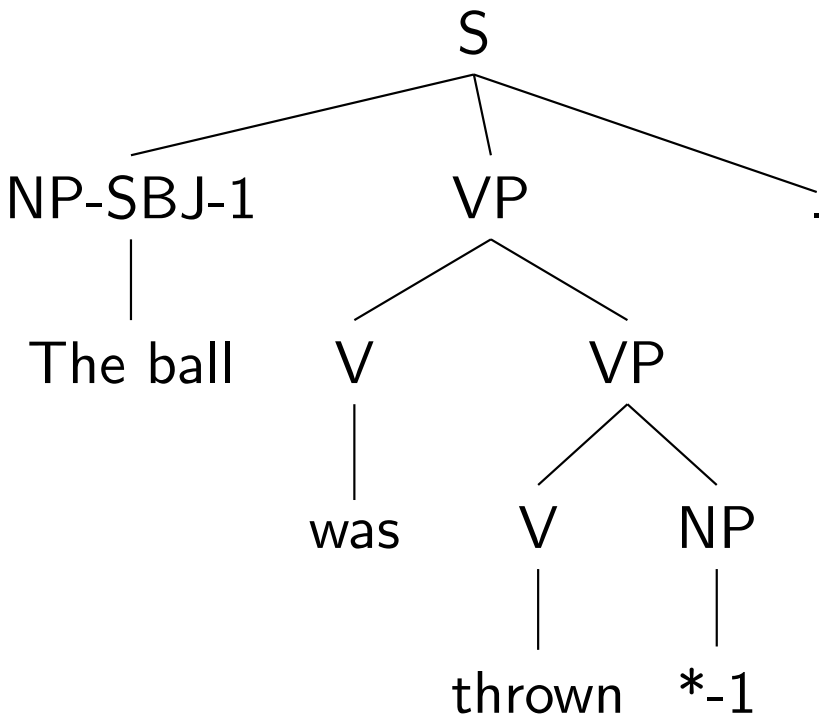
Syntactic labels: NP, VP
Function tags: (second version) -SBJ, -LOC
Empty categories for movement




```
S (NP-SBJ (NNP Casey))
   (VP (VBP throws)
          (NP (DT the) (NN ball)))
   (. .))
```

Trees not very deep or tall
Phrase labels are S, NP, VP
Function tags: (second version) -SBJ, -LOC Most sentences are long, approx 20 words

Penn Treebank



Empty categories for movement
Passivized subject originates as the object

Use of Treebanks

Naturally-occurring text as the basis for learning about language
Build parsers and evaluate them
Test linguistic grammars and discover patterns

Newspaper-ese

Copper finished [down 4.5 cents], [at $1.2345 a pound]
It was used to investigate wave behaviour, estimate the wave energy and forecast NP coastal changes

Introduced special function tags for these phrase types
Language used in newspapers can be very domain specific.

Disambiguate

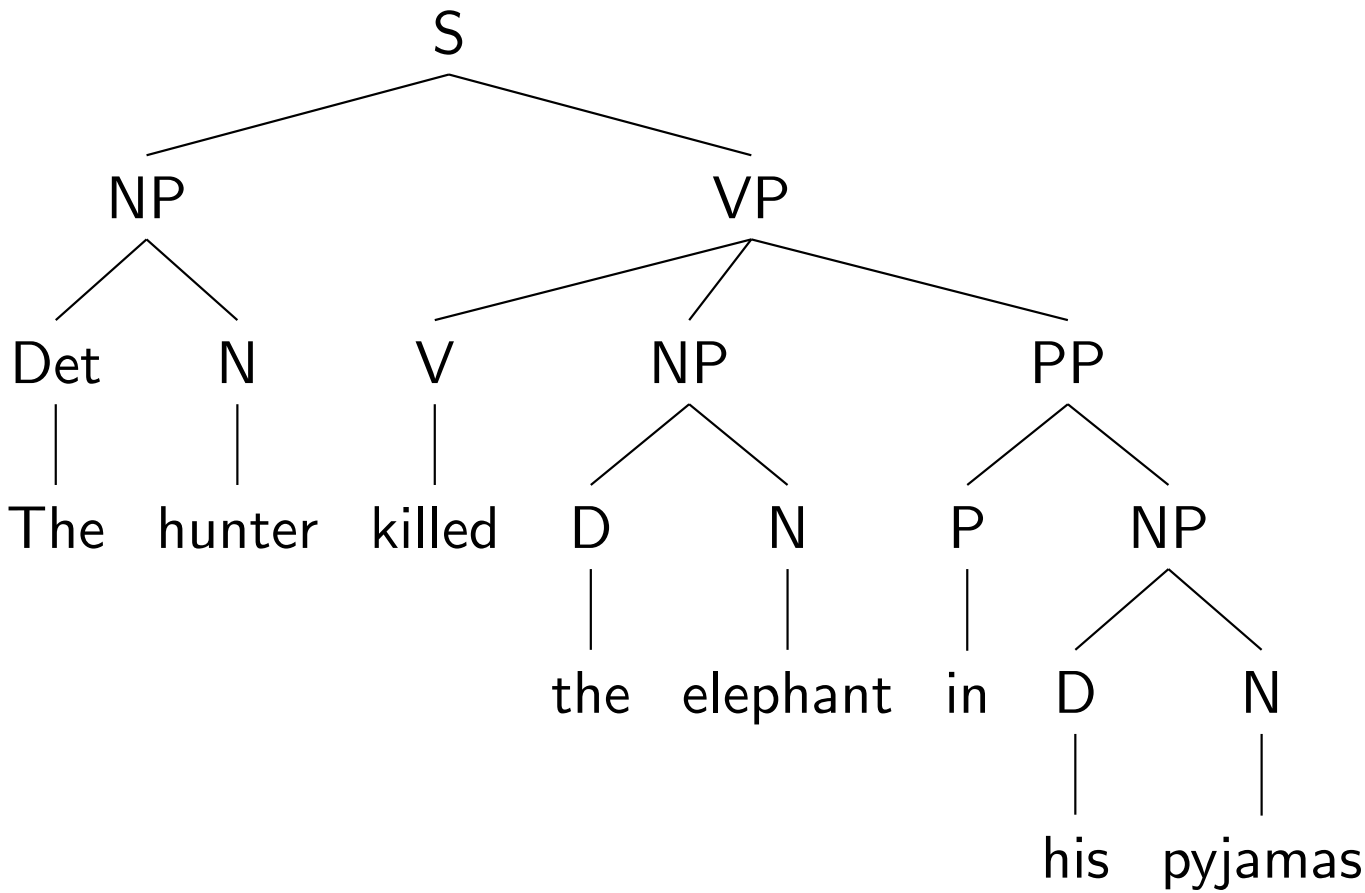
English

The hunter [killed [the elephant]] [in his pajamas]
The hunter [killed [the elephant [in his pajamas]]

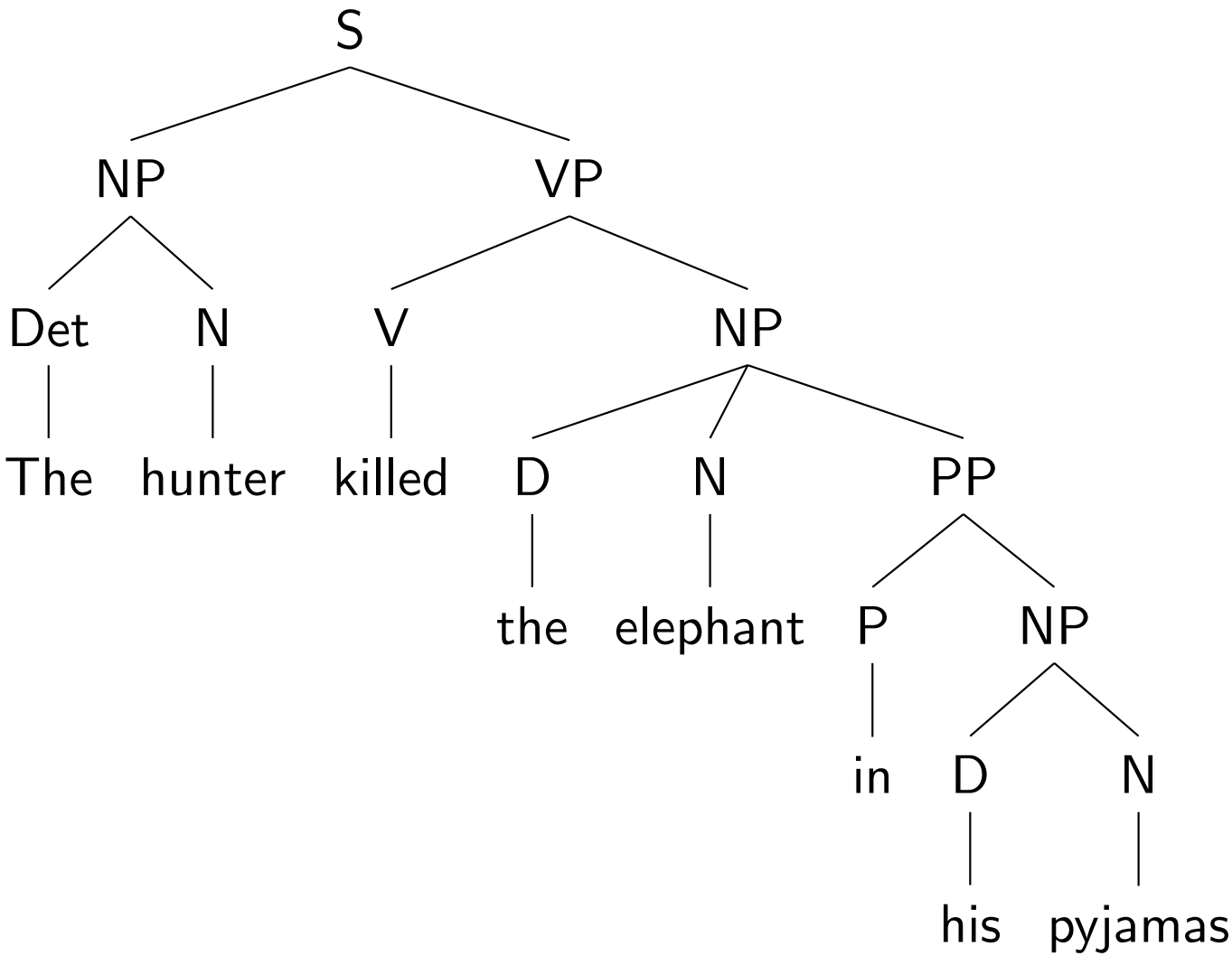
Disambiguate

English

The hunter [killed [the elephant]] [in his pajamas]
The hunter [killed [the elephant [in his pajamas]]



Disambiguate



Examples of knowledge gained from Treebanks

PP attaches to the verb: high
PP attaches to the noun: low

How will a parser disambiguate?

Important words

V	N1	P	N2	Attach
join	board	as	director	
named	director	of	conglomerate	
caused	percentage	of	deaths	
bring	attention	to	problem	
shot	elephant	in	pyjamas	
led	team	of	researchers	

Important words

V	N1	P	N2	Attach
join	board	as	director	V
named	director	of	conglomerate	N
caused	percentage	of	deaths	N
bring	attention	to	problem	V
shot	elephant	in	pyjamas	N/V?
led	team	of	researchers	N

Humans disambiguate 88% of the time given these 4 words only
If PPs are always attached low (N), we can get the correct answer about half the time

Treebank analysis

The preposition 'of' seems to occur most often with N (low)
Others like 'as' or 'to' occur more often with V (high)
Some like 'against' can occur almost equally with both

Usually, verbs like bring, buy, put may occur more often with high PP attachment

of is very frequent (as well as *to*) hence these will be most useful
against is rare, and can be ignored
Rules that cover most of the cases will end up being more effective

of is very frequent (as well as *to*) hence these will be most useful
against is rare, and can be ignored
Rules that cover most of the cases will end up being more effective

If the preposition is of, label the tuple as low.
If the preposition is to, label the tuple as high.
label the tuple as high

Human parsing/Sentence processing

How does the human parser tolerate ambiguity?

Garden path sentences

While Susan was dressing the baby played on the floor.
While Susan was dressing herself the baby played on the floor.

Evidence to show that human parsing is incremental
We don't wait to hear the whole sentence before parsing
But there is some processing cost at *the baby*

While Susan was dressing the baby played on the floor.
While Susan was dressing herself the baby played on the floor.

Build the wrong structure, find disambiguating information, discover the error
Build the right structure, revise

Late closure: not to build unnecessary structure, but work on the same clause
Minimal attachment: build structure with the fewest nodes

References

Slides on the Treebank are courtesy Fei Xia. PP attachment slides are courtesy Dragomir Radev's NLP course on Coursera. Processing slides courtesy Samar Husain