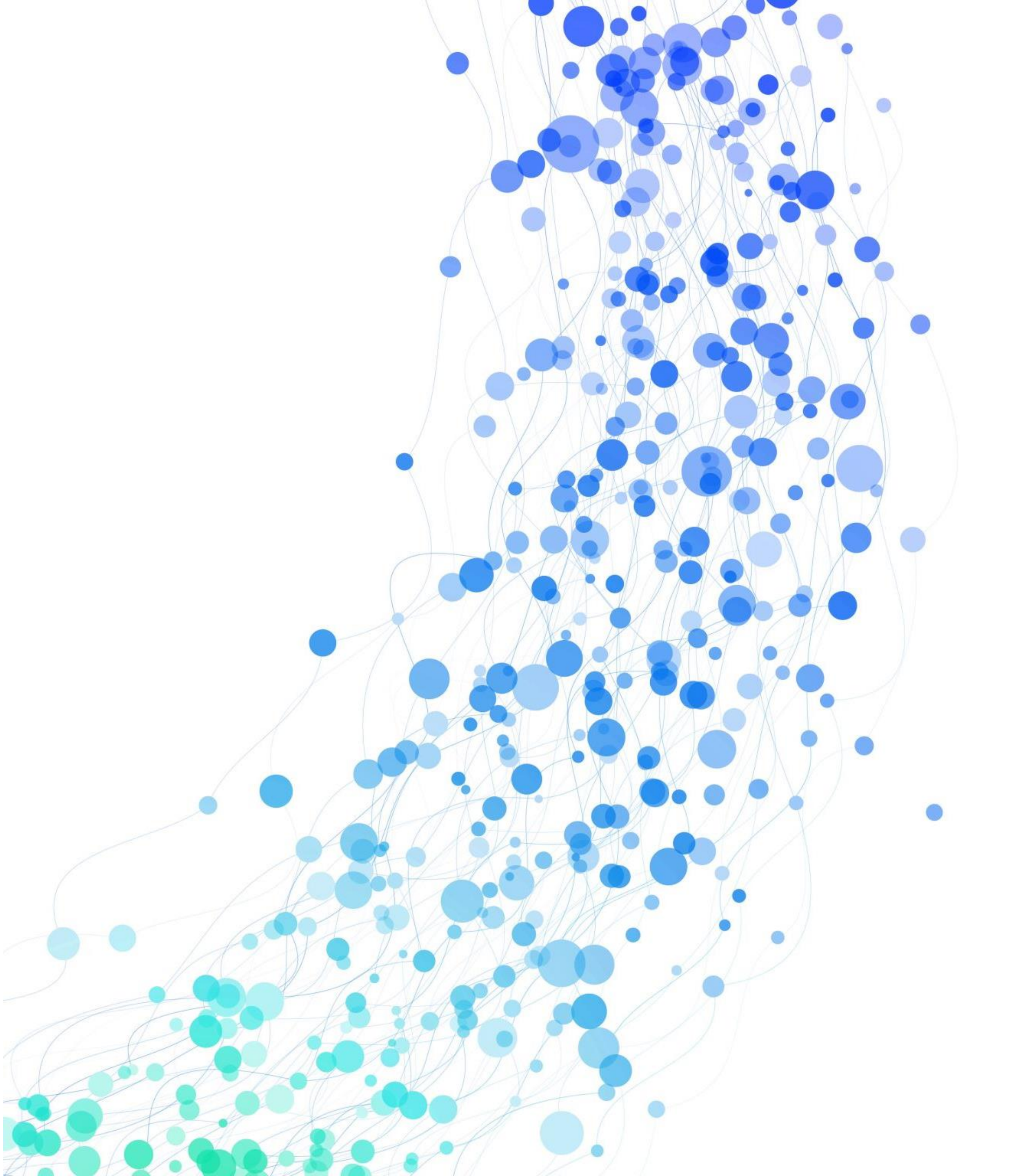
LECTURE 17: CFG PARSING



Mehmet Can Yavuz, PhD.

Adapted from Julia Hockenmaier, NLP S2023 - course material
https://courses.grainger.illinois.edu/cs447/sp2023/

PENN TREEBANK PARSING

THE PENN TREEBANK

The first publicly available syntactically annotated corpus Wall Street Journal (50,000 sentences, 1 million words)
also Switchboard, Brown corpus, ATIS

The annotation:

– POS-tagged (Ratnaparkhi's MXPOST)

– Manually annotated with phrase-structure trees

– Richer than standard CFG: *Traces* and other *null elements* used to represent non-local dependencies (designed to allow extraction of predicate-argument structure), although these are typically removed when we do parsing

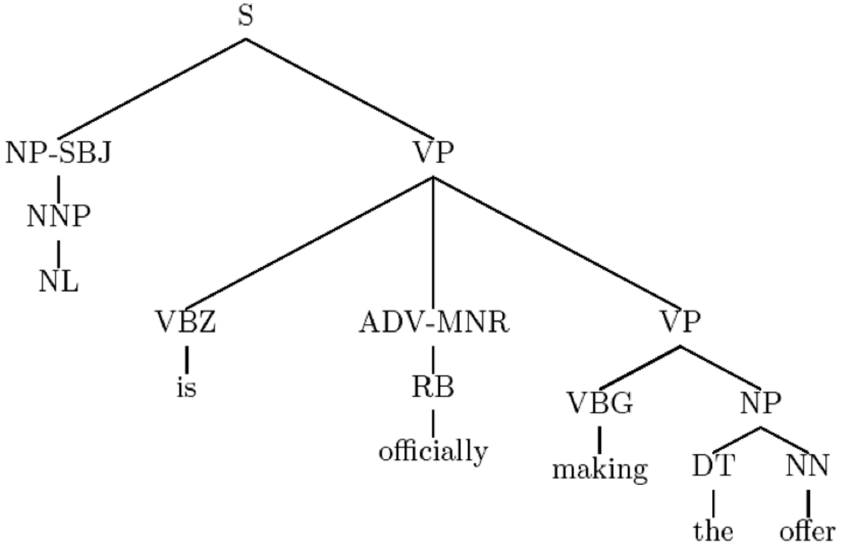
[more on non-local dependencies and traces later in the semester]

The standard data set for English phrase-structure parsers

THE TREEBANK LABEL SET

- 48 preterminals (tags):
- – 36 POS tags, 12 other symbols (punctuation etc.)
- – Simplified version of Brown tagset (87 tags)
  (cf. Lancaster-Oslo/Bergen (LOB) tag set: 126 tags)
- 14 nonterminals:
  Standard inventory (S, NP, VP, PP, ADJP, ADVP, SBAR,…)
- Many nonterminals have function tags indicating their syntactic roles (NP-SBJ: subject NP) or what role they play
- (e.g. PP-LOC: locative PP, i.e. indicating a location [“in NYC”] PP-DIR: directional PP, indicating a direction [“to NYC”],
- ADVP-MNR: manner adverb [“slowly”]).
- For historical reasons, these function tags are typically removed before parsing.

A SIMPLE EXAMPLE



Relatively flat structures:

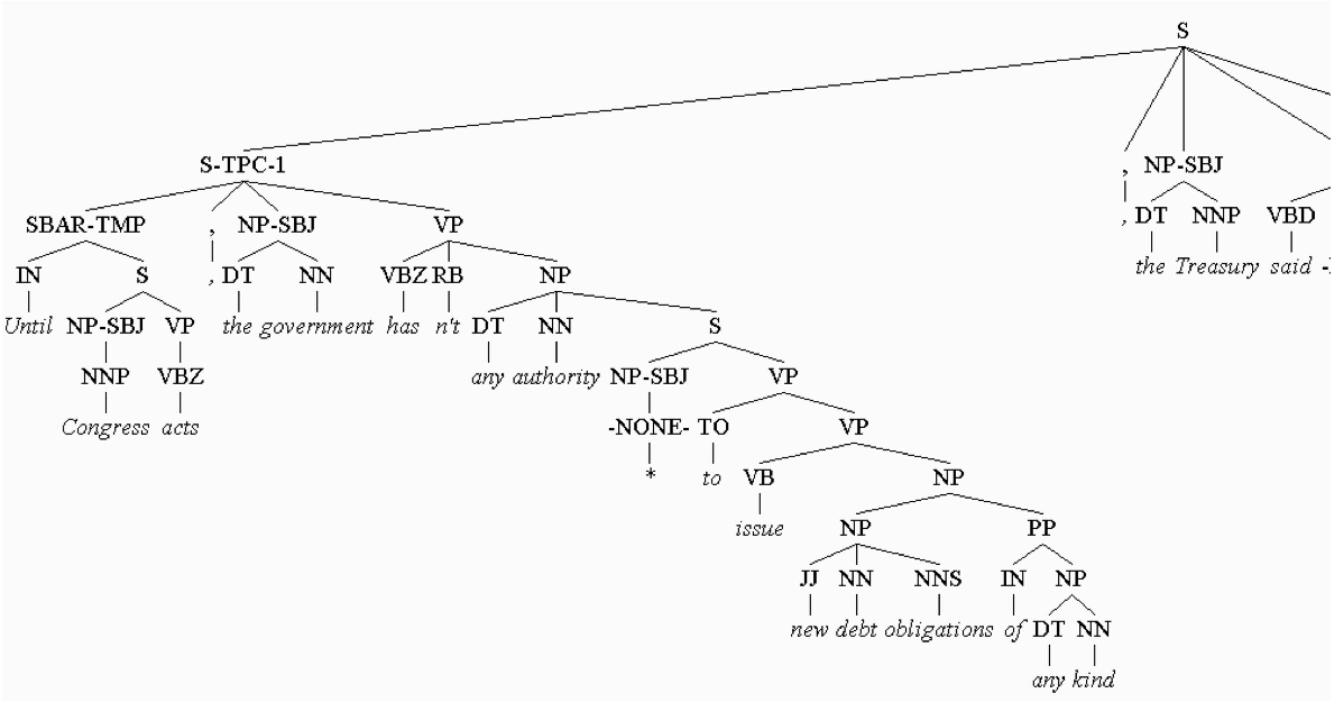
– There is no noun level

– VP arguments and adjuncts appear at the same level

Function tags, e.g. -SBJ (subject), -MNR (manner)

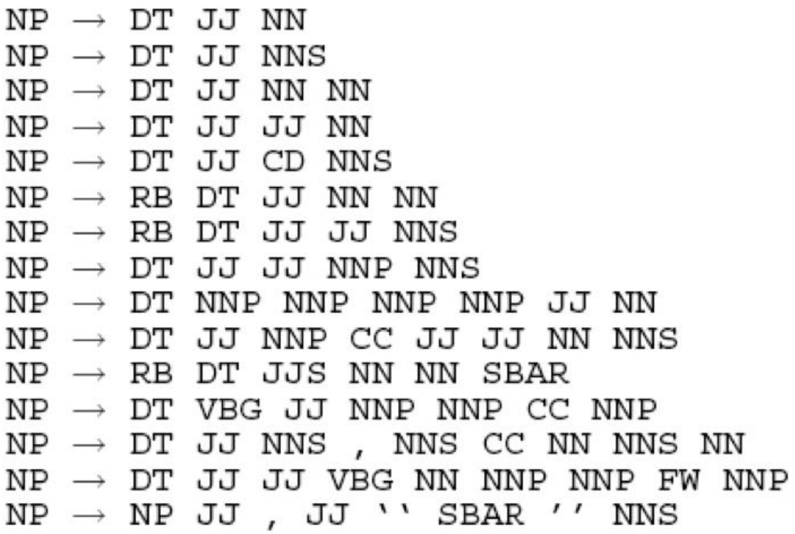
A MORE REALISTIC (PARTIAL) EXAMPLE

Until Congress acts, the government hasn't any authority to issue new debt obligations of any kind, the Treasury said .... .



THE PENN TREEBANK CFG

The Penn Treebank uses very flat rules, e.g.:



```
NP → DT JJ NN
NP → DT JJ NNS
NP → DT JJ NN NN
NP → DT JJ JJ NN
NP → DT JJ CD NNS
NP → RB DT JJ NN NN
NP → RB DT JJ JJ NNS
NP → DT JJ JJ NNP NNS
NP → DT NNP NNP NNP NNP JJ NN
NP → DT JJ NNP CC JJ JJ NN NNS
NP → RB DT JJS NN NN SBAR
NP → DT VBG JJ NNP NNP CC NNP
NP → DT JJ NNS , NNS CC NN NNS NN
NP → DT JJ JJ VBG NN NNP NNP FW NNP
NP → NP JJ , JJ `` SBAR '' NNS
```

- Basic PCFGs don’t work well on the Penn Treebank
  - – Many of these rules appear only once.
  - – But many of these rules are very similar.

  Can we generalize by not treating each rule as atomic?

SUMMARY

The Penn Treebank has a large number of very flat rules.

Accurate parsing requires modifications to basic PCFG models:

- Generalizing across similar rules (“Markov PCFGs”)
- Modeling word-word dependencies
  (although this does not help as much as people used to think)
- Refining the nonterminals to capture more context
  How much of this transfers to other treebanks or languages?

APPENDIX: A CONTEXT-FREE GRAMMAR FOR A FRAGMENT OF ENGLISH

NOUN PHRASES (NPS)

- Simple NPs:
- [He] sleeps. (pronoun)
  [John] sleeps. (proper name)
  [A student] sleeps. (determiner + noun)
  [A tall student] sleeps. (det + adj + noun) [Snow] falls. (noun)
- Complex NPs:
- [The student in the back] sleeps. (NP + PP)
  [The student who likes MTV] sleeps. (NP + Relative Clause)

THE NP FRAGMENT

- NP → Pronoun
  NP → ProperName
- NP → Det Noun
  NP → Noun
  NP → NP PP
  NP → NP RelClause
- Noun → AdjP Noun
  Noun → N
  N → {class,... student, snow, ...}
- Det → {a, the, every,... } Pronoun → {he, she,...} ProperName → {John, Mary,...}

ADJECTIVE PHRASES (ADJP)

AND PREPOSITIONAL PHRASES (PP)

- AdjP → Adj
  AdjP → Adv AdjP
  Adj → {big, small, red,...} Adv → {very, really,...}
- PP → P NP
  P → {with, in, above,...}

THE VERB PHRASE (VP)

- He [eats].
  He [eats sushi].
  He [gives John sushi].
  He [gives sushi to John].
  He [eats sushi with chopsticks]. He [somtimes eats].

-
  ```
  VP → V
  VP → V NP
  VP → V NP NP
  VP → V NP PP
  VP → VP PP
  VP → AdvP VP
  V → {eats, sleeps gives,...}
  ```

CAPTURING SUBCATEGORIZATION

- He [eats]. ✓
  He [eats sushi]. ✓
  He [gives John sushi]. ✓
  He [eats sushi with chopsticks]. ✓ *He [eats John sushi]. ???
- `VP → Vintrans`
  `VP → Vtrans NP`
  `VP → Vditrans NP NP VP → VP PP`
- `Vintrans → {eats, sleeps}  Vtrans → {eats}`
  `Vditrans → {gives}`

SENTENCES

- [He eats sushi].
- [Sometimes, he eats sushi].
- [In Japan, he eats sushi].

- `S → NP VP`
- `S → AdvP S`
- `S → PP S`

CAPTURING AGREEMENT

- [He eats sushi]. ✓
- *[I eats sushi]. ???
- *[They eats sushi]. ???
- `S →NP3sg VP3sg`
- `S →NP1sg VP1sg`
- `S →NP3pl VP3pl`
- **We would need features to capture agreement:**

(number, person, case,...)

COMPLEX VPS

In English, simple tenses have separate forms:

Present tense: the girl **eats** sushi

Simple past tense: the girl **ate** sushi

Complex tenses, progressive aspect and passive voice consist of auxiliaries and participles:

Past perfect tense: the girl **has eaten** sushi

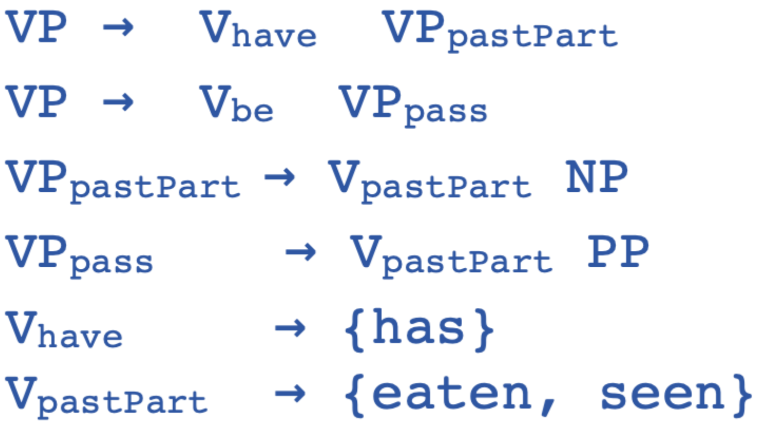
Future perfect tense: the girl **will have eaten** sushi

Passive voice: the sushi **is/was/will be/… eaten** by the girl

Progressive aspect: the girl **is/was/will be eating** sushi

VPS REDEFINED

- He [has [eaten sushi]].
  The sushi [was [eaten by him]].



$VP \rightarrow V_{have}\ VP_{pastPart}$

$VP \rightarrow V_{be}\ VP_{pass}$

$VP_{pastPart} \rightarrow V_{pastPart}\ NP$

$VP_{pass} \rightarrow V_{pastPart}\ PP$

$V_{have} \rightarrow \{has\}$

$V_{pastPart} \rightarrow \{eaten,\ seen\}$

We would need even more nonterminals (e.g. VPpastpart)!

- N.B.: We call VPpastPart, VPpass, etc. `untensed’ VPs

SUBORDINATION

- He says [he eats sushi].
  He says [that [he eats sushi]].

-
  ```
  VP → Vcomp S
  VP → Vcomp SBAR
  SBAR → COMP S
  Vcomp → {says, think, believes}
  COMP → {that}
  ```

COORDINATION

- [He eats sushi] but [she drinks tea]
- [John] and [Mary] eat sushi.
- He [eats sushi] and [drinks tea]
- He [sells and buys] shares
- He eats [at home or at a restaurant]
- `S →SconjS`
- `NP →NP conj NP`
- `VP →VP conj VP  V →VconjV`
- `PP →PP conj PP`

RELATIVE CLAUSES

Relative clauses modify noun phrases:
the girl [that eats sushi] (NP → NP RelClause)

Relative clauses lack an NP that is understood to be filled by the NP they modify:

- ‘the girl that eats sushi’ implies ‘the girl eats sushi’

Subject relative clauses lack a subject: ‘the girl that eats sushi’

- RelClause → RelPron VP [sentence w/o sbj = VP]

Object relative clauses lack an object: ‘the sushi that the girl eats’ Define “slash categories” S-NP,VP-NP that are missing object NPs

- RelClause → RelPron S-NP
- S-NP → NP VP-NP
- VP-NP → Vtrans
- VP-NP → VP-NP PP

YES/NO QUESTIONS

Yes/no questions consist of an auxiliary, a subject and an (untensed) verb phrase:

does she eat sushi?

have you eaten sushi?

YesNoQ → Aux NP VPinf

YesNoQ → Aux NP VPpastPart

WH-QUESTIONS

- Subject wh-questions consist of an wh-word,   an auxiliary and an (untensed) verb phrase:
- Who has eaten the sushi?
- `WhQ → WhPron Aux VPpastPart`
- Object wh-questions consist of an wh-word,
  an auxiliary, an NP and an (untensed) verb phrase that is missing an object.
- What does Mary eat?
- `WhQ → WhPron Aux NP VPinf-NP`