CS-AD 220 – Spring 2016

Natural Language Processing

Session 16: 31-Mar-16

Prof. Nizar Habash

Some slides are adapted from Jurafsky and Martin's course slides on Speech and Language Processing

NYUAD Course CS-AD 220 – Spring 2016

Natural Language Processing

Assignment #3 : POS Tagging and Parsing

Assigned Mar 31, 2016 / Due Apr 17, 2016 (11:59pm)

I. Grading & Submission

This assignment is about the development of a dependency parser and a part-of-speech (POS) tagger for English. The assignment accounts for 15% of the full grade. It consists of five exercises. **There is also a bonus exercise that can count for up to 5% of the full grade**. The additional exercise consists of a parsing competition on an unseen test set. Participation earns 2%. The first, second and third ranked systems earn additional 3%, 2% and 1%, respectively.

Assignment #3 posted on NYU Classes

START EARLY!

DEADLINE PUSHED FORWARD TO APR 17

Moving Legislative Day Class

- Spring Break is March 18 – 25, 2016
- Sat March 26, 2016 is a Legislative *Thursday*
- Move to

Sat April 2, 2016 at 10am

Same Classroom C2-E049

Syntax

- By grammar, or syntax, we have in mind the kind of implicit knowledge of your native language that you had mastered by the time you were 3 years old without explicit instruction
- Not the kind of stuff you were later taught in “grammar” school
- Descriptive vs. Prescriptive Grammars

Syntax

- Why should you care?
- Grammars (and parsing) are key components in many applications
  - Grammar checkers
  - Dialogue management
  - Question answering
  - Information extraction
  - Machine translation

Syntax

- Key notions that we'll cover
  - Constituency
  - Dependency
- Key formalism
  - Context-free grammars
- Resources
  - Treebanks
- Parsing Algorithms
  - Constituency
  - Dependency

Constituency

- The basic idea here is that groups of words within utterances can be shown to act as single units.
- And in a given language, these units form coherent classes that can be be shown to behave in similar ways
  - With respect to their internal structure
  - And with respect to other units in the language

Multilingual Syntactic Variations

يقرأ الطالب المجتهد كتابا عن الصين في الصف

read the-student the-diligent a-book about china in the-classroom

---

the diligent student is reading a book about china in the classroom

---

这位勤奋的学生在教室读一本关于中国的书

this quant diligent *de* student in classroom read *one quant* about china *de* book

Multilingual Syntactic Variations

يقرأ الطالب المجتهد كتابا عن الصين في الصف

read the-student the-diligent a-book about china in the-classroom

---

the diligent student is reading a book about china in the classroom

---

这位勤奋的学生在教室读一本关于中国的书

this quant diligent de student in classroom read one quant about china de book

Multilingual Syntactic Variations

Multilingual Syntactic Variations

يقرأ الطالب المجتهد كتابا عن الصين في الصف

read the-student the-diligent a-book about china in the-classroom

---

the diligent student is reading a book about china in the classroom

---

这位勤奋的学生在教室读一本关于中国的书

this quant diligent de student in classroom read one quant about china de book

Multilingual Syntactic Variations

Multilingual Syntactic Variations

Multilingual Syntactic Variations

Multilingual Syntactic Variations

Multilingual Syntactic Variations

يقرأ الطالب المجتهد كتابا عن الصين في الصف

read the-student the-diligent a-book about china in the-classroom

---

the diligent student is reading a book about china in the classroom

---

这位勤奋的学生在教室读一本关于中国的书

this quant diligent de student in classroom read one quant about china de book

	Arabic		English		Chinese	
Subj-Verb	V Subj	Subj V	Subj V		Subj … V	
Verb-PP	V…PP		V…PP		V PP	PP V
Adjectives	N Adj		Adj N		Adj de N	
Possessives	N Poss		N of Poss	Poss 's N	Poss de N	
Relatives	N Rel		N Rel		Rel de N	

English Constituency Tree

```
(S
  (NP (DT the) (JJ diligent) (NN student))
  (VP (VBZ is)
    (VP (VBG reading)
      (NP
        (NP (DT a) (NN book))
        (PP (IN about)
            (NP (NN china))))
      (PP (IN in)
          (NP (DT the) (NN classroom))))))
```

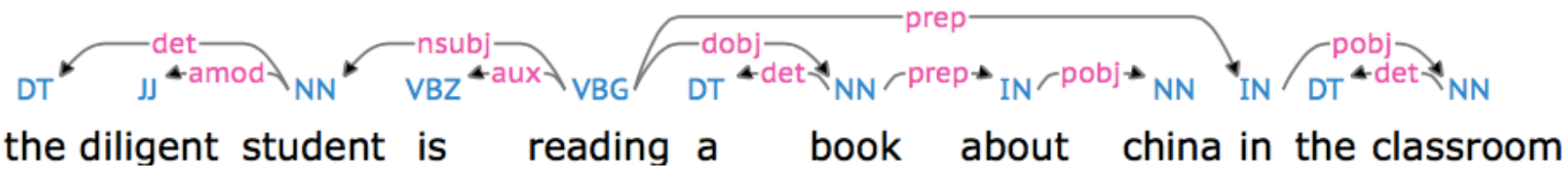
Arabic Constituency Tree

```
(S
  (VP (VBP yqrA/reads)
      (NP (DTNN AlTAlb/the-student) (DTJJ Almjthd/the-diligent))
      (NP
        (NP (NN ktAbA/a-book))
        (PP (IN En/about)
          (NP (DTNNP AlSyn/the-China))))
      (PP (IN fy/in)
          (NP (DTNN AlSf/the-classroom)))))
```

English Constituency Tree

```
(S
  (NP (DT the) (JJ diligent) (NN student))
  (VP (VBZ is)
    (VP (VBG reading)
      (NP
        (NP (DT a) (NN book))
        (PP (IN about)
          (NP (NN china))))
      (PP (IN in)
        (NP (DT the) (NN classroom))))))
```

English Dependency Tree



Dependency created with http://www.cs.cmu.edu/~ark/TurboParser/

Constituency

- Internal structure
  - We can describe an internal structure to the class (might have to use disjunctions of somewhat unlike sub-classes to do this).
- External behavior
  - For example, we can say that noun phrases can come before verbs

Constituency

- For example, it makes sense to the say that the following are all ***noun phrases*** in English...

Harry the Horse	a high-class spot such as Mindy's
the Broadway coppers	the reason he comes into the Hot Box
they	three parties from Brooklyn

- Why? One piece of evidence is that they can all precede verbs.
  - This is external evidence

Grammars and Constituency

- Of course, there's nothing easy or obvious about how we come up with right set of constituents and the rules that govern how they combine...
- That's why there are so many different theories of grammar and competing analyses of the same data.
- The approach to grammar, and the analyses, adopted here are very generic (and don't correspond to any modern linguistic theory of grammar).

Context-Free Grammars

- Context-free grammars (CFGs)
  - Also known as
    - Phrase structure grammars
    - Backus-Naur form
- Consist of
  - Rules
  - Terminals
  - Non-terminals

Context-Free Grammars

- Terminals
  - We'll take these to be words (for now)
- Non-Terminals
  - The constituents in a language
    - Like noun phrase, verb phrase and sentence
- Rules
  - Rules are equations that consist of a single non-terminal on the left and any number of terminals and non-terminals on the right.

Some NP Rules

- Here are some rules for our noun phrases

$$
\begin{aligned}
NP &\rightarrow Det\ Nominal \\
NP &\rightarrow ProperNoun \\
Nominal &\rightarrow Noun \mid Nominal\ Noun
\end{aligned}
$$

Some NP Rules

- Here are some rules for our noun phrases

$$
\begin{aligned}
NP &\rightarrow Det\ Nominal \\
NP &\rightarrow ProperNoun \\
Nominal &\rightarrow Noun \mid Nominal\ Noun
\end{aligned}
$$

- Together, these describe two kinds of NPs.
  - One that consists of a determiner followed by a nominal
  - And another that says that proper names are NPs.
  - The third rule illustrates two things
    - An explicit disjunction
      - Two kinds of nominals
    - A recursive definition
      - Same non-terminal on the right and left-side of the rule

A Simple Grammar

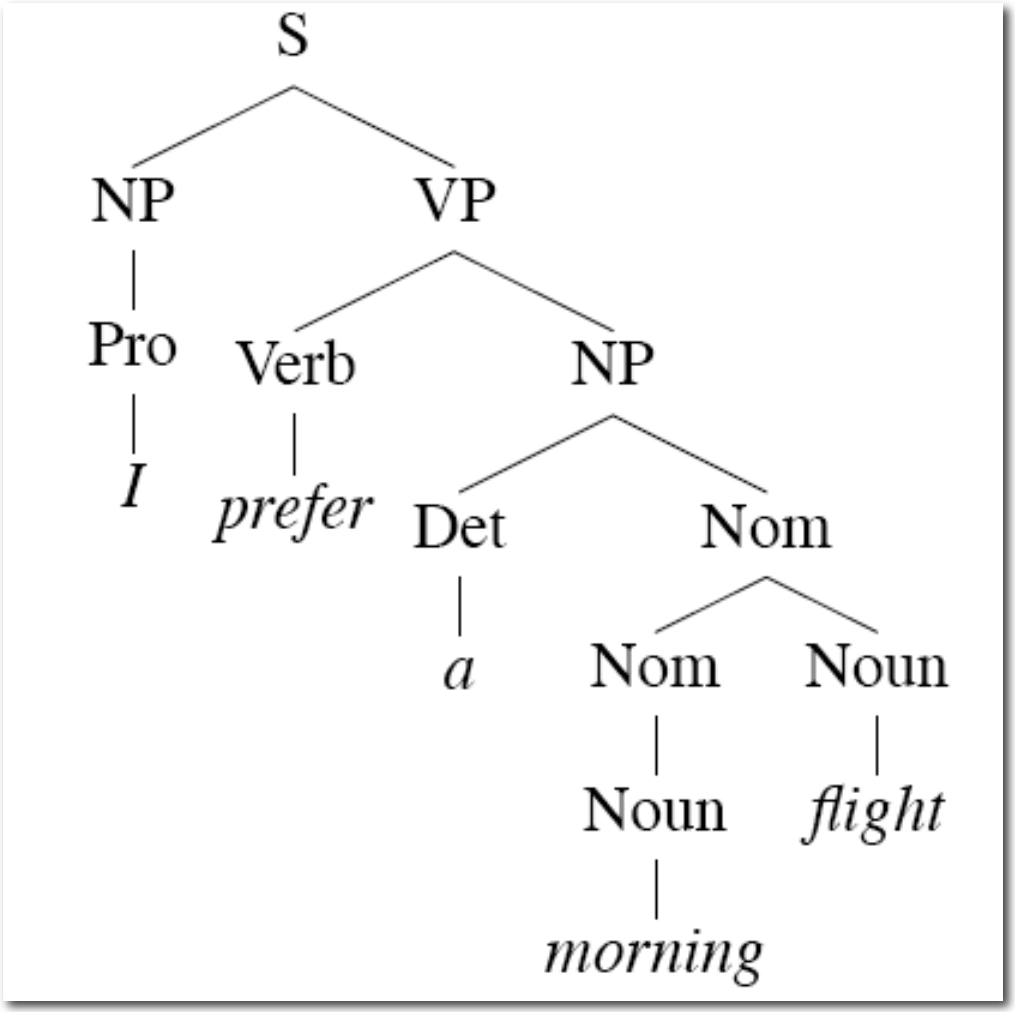
Grammar Rules			Examples
S	→	*NP VP*	I + want a morning flight
NP	→	*Pronoun*	I
	|	*Proper-Noun*	Los Angeles
	|	*Det Nominal*	a + flight
Nominal	→	*Nominal Noun*	morning + flight
	|	*Noun*	flights
VP	→	*Verb*	do
	|	*Verb NP*	want + a flight
	|	*Verb NP PP*	leave + Boston + in the morning
	|	*Verb PP*	leaving + on Thursday
PP	→	*Preposition NP*	from + Los Angeles

Generativity

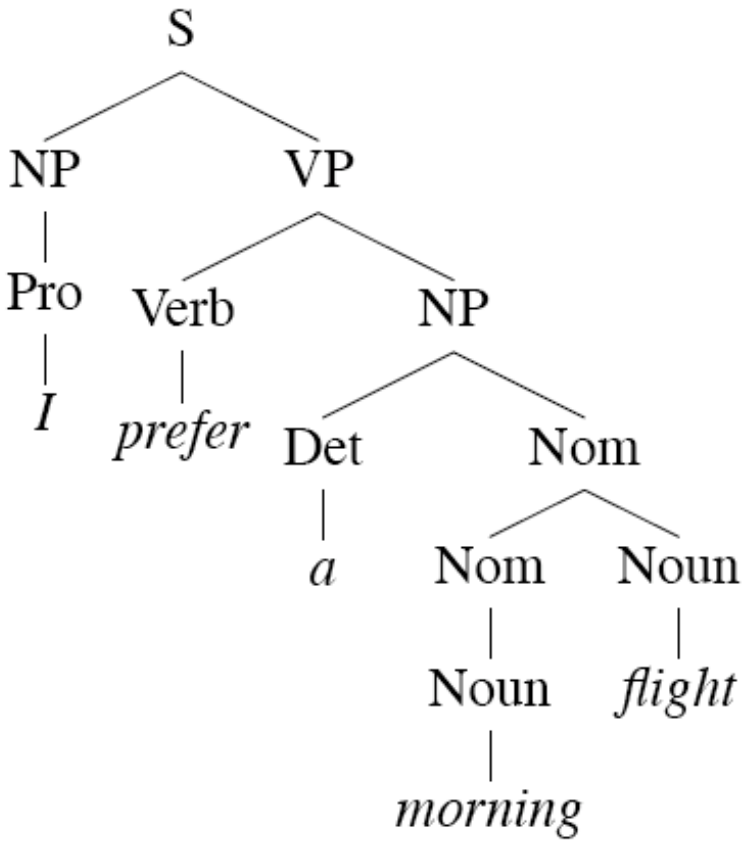
- As with FSAs and FSTs, you can view these rules as either analysis or synthesis machines
  - Generate strings in the language
  - Reject strings not in the language
  - Impose structures (trees) on strings in the language

Derivations

- A derivation is a sequence of rules applied to a string that *accounts* for that string
  - Covers all the elements in the string
  - Covers only the elements in the string



Equivalent Forms




```
(S (NP (Pro I))
   (VP (Verb prefer)
       (NP (Det a)
           (Nom (Nom (Noun morning))
                (Noun flight)))))
```

```
(S
  (NP
    I/Pro)
  (VP
    prefer/Verb
    (NP
       A/Det
       (Nom
            (Nom
               morning/Noun)
            flight/Noun))))
```

```
(S (NP (Pro I)) (VP (Verb prefer) (NP (Det a) (Nom (Nom (Noun morning)) (Noun flight)))))
```

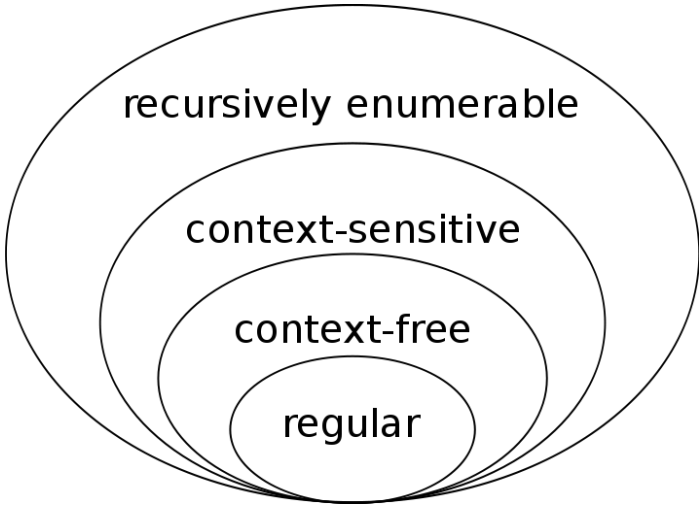
Definition

- More formally, a CFG consists of

N	a set of **non-terminal symbols** (or **variables**)
Σ	a set of **terminal symbols** (disjoint from N)
R	a set of **rules** or productions, each of the form $A \rightarrow \beta$, where A is a non-terminal, β is a string of symbols from the infinite set of strings $(\Sigma \cup N)*$
S	a designated **start symbol**

Parsing

- Parsing is the process of taking a string and a grammar and returning a (multiple?) parse tree(s) for that string
- It is completely analogous to running a finite-state transducer with a tape
  - It's just more powerful
  - *There are languages we can capture with CFGs that we can't capture with finite-states*
  - *Chomsky Language Hierarchy*
    - *E.g. $a^n b^n$ is a context-free language*



An English Grammar Fragment

- Sentences
- Noun phrases
  - Agreement
- Verb phrases
  - Subcategorization

Sentence Types

- Declaratives: *A plane left.*

  S → NP VP

- Imperatives: *Leave!*

  S → VP

- Yes-No Questions: *Did the plane leave?*

  S → Aux NP VP

- WH Questions: *When did the plane leave?*

  S → WH-NP Aux NP VP

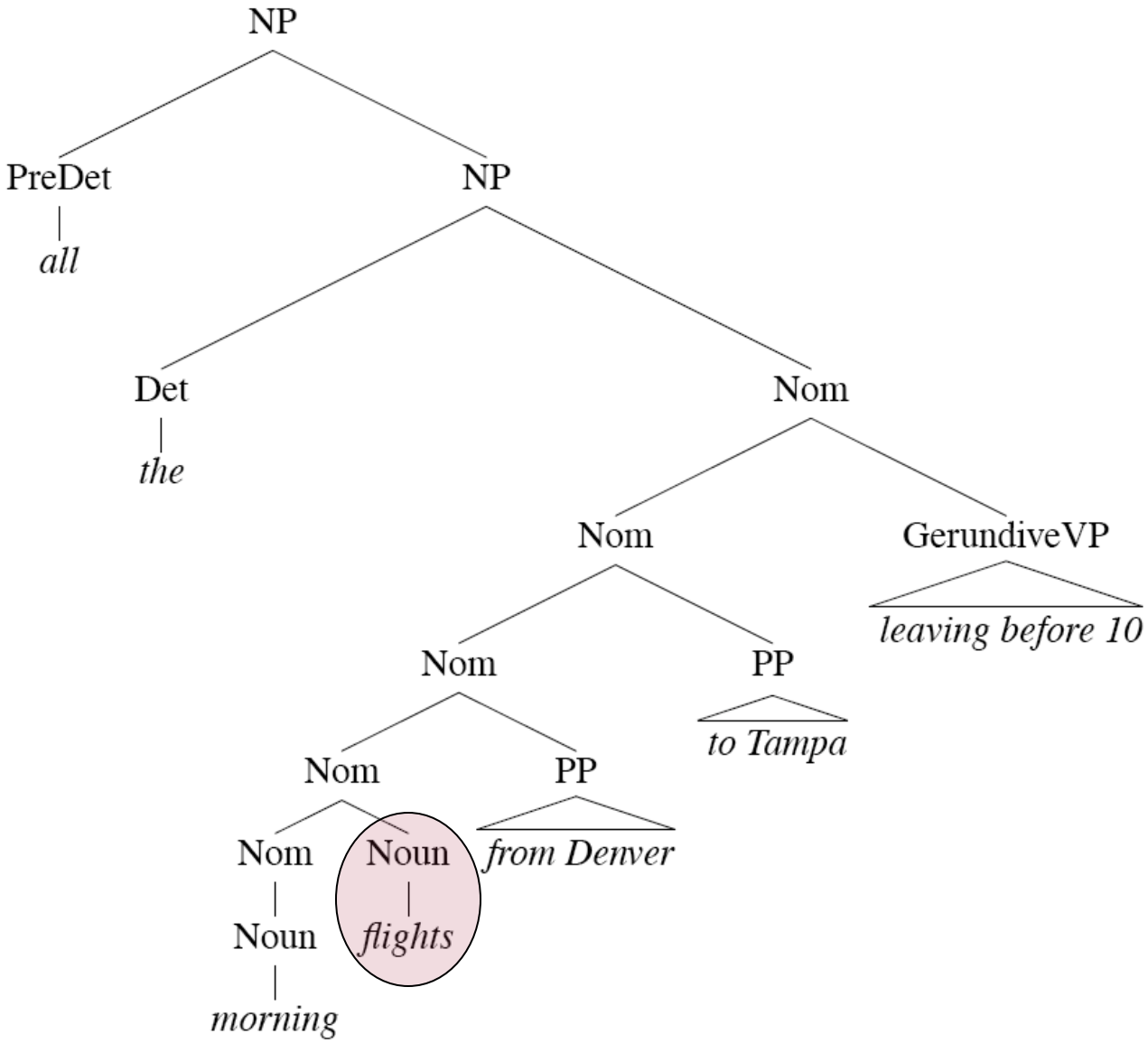
Noun Phrases

- Let's consider the following rule in more detail...

  NP → Det Nominal

- Most of the complexity of English noun phrases is hidden in this rule.
- Consider the derivation for the following example
  - *All the morning flights from Denver to Tampa leaving before 10*

Noun Phrases



NP Structure

- Clearly this NP is really about *flights.* That's the central criticial noun in this NP. Let's call that the *head*.
- We can dissect this kind of NP into the stuff that can come before the head, and the stuff that can come after it.

Determiners

- Noun phrases can start with determiners...
- Determiners can be
  - Simple lexical items: *the, this, a, an*, etc.
    - A car
  - Or simple possessives
    - John's car
  - Or complex recursive versions of that
    - John's sister's husband's son's car

Nominals

- Contains the head and any pre- and post-modifiers of the head.
  - Pre-
    - Quantifiers, cardinals, ordinals...
      - Three cars
    - Adjectives and APs
      - large cars
    - Ordering constraints
      - Three large cars
      - ?large three cars

Postmodifiers

- Three kinds
  - Prepositional phrases
    - From Seattle
  - Non-finite clauses
    - Arriving before noon
  - Relative clauses
    - That serve breakfast
- Same general (recursive) rule to handle these
  - *Nominal → Nominal PP*
  - *Nominal → Nominal GerundVP*
  - *Nominal → Nominal RelClause*

Agreement

- By ***agreement***, we have in mind constraints that hold among various constituents that take part in a rule or set of rules
- For example, in English, determiners and the head nouns in NPs have to agree in their number.

This flight

Those flights

*This flights

*Those flight

Problem

- Our earlier NP rules are clearly deficient since they don't capture this constraint
  - *NP → Det Nominal*
    - Accepts, and assigns correct structures, to grammatical examples (*this flight*)
    - But its also happy with incorrect examples (*these flight)
  - Such a rule is said to *overgenerate.*
  - We'll come back to this in a bit

Verb Phrases

- English *VP*s consist of a head verb along with 0 or more following constituents which we'll call *arguments*.

$VP \rightarrow Verb$ disappear
$VP \rightarrow Verb\ NP$ prefer a morning flight
$VP \rightarrow Verb\ NP\ PP$ leave Boston in the morning
$VP \rightarrow Verb\ PP$ leaving on Thursday

Subcategorization

- But, even though there are many valid VP rules in English, not all verbs are allowed to participate in all those VP rules.
- We can subcategorize the verbs in a language according to the sets of VP rules that they participate in.
- This is a modern take on the traditional notion of transitive/intransitive.
- Modern grammars may have 100s or such classes.

Subcategorization

- Sneeze: John sneezed
- Find: Please find [a flight to NY]$_{NP}$
- Give: Give [me]$_{NP}$[a cheaper fare]$_{NP}$
- Help: Can you help [me]$_{NP}$[with a flight]$_{PP}$
- Prefer: I prefer [to leave earlier]$_{TO\text{-}VP}$
- Told: I was told [United has a flight]$_{S}$
- …

Subcategorization

- *John sneezed the book
- *I prefer United has a flight
- *Give with a flight

- As with agreement phenomena, we need a way to formally express the constraints

Why?

- Right now, the various rules for VPs *overgenerate*.
  - They permit the presence of strings containing verbs and arguments that don't go together
  - For example
  - VP -> V NP therefore

    Sneezed the book is a VP since "sneeze" is a verb and "the book" is a valid NP

Possible CFG Solution

- Possible solution for agreement.
- Can use the same trick for all the verb/ VP classes.

- SgS -> SgNP SgVP
- PlS -> PlNp PlVP
- SgNP -> SgDet SgNom
- PlNP -> PlDet PlNom
- PlVP -> PlV NP
- SgVP ->SgV Np
- …

CFG Solution for Agreement

- It works and stays within the power of CFGs
- But it's ugly
- And it doesn't scale all that well because of the interaction among the various constraints explodes the number of rules in our grammar.

The Point

- CFGs appear to be just about what we need to account for a lot of basic syntactic structure in English.
- But there are problems
  - That can be dealt with adequately, although not elegantly, by staying within the CFG framework.
- There are simpler, more elegant, solutions that take us out of the CFG framework (beyond its formal power)
  - Lexical Functional Grammar (LFG), Head-Driven Phrase Structure Grammar (HPSG), Construction grammar, XTAG, etc.

Next Time

Sat April 2, 2016 at 10am

Same Classroom C2-E049

No reading assigned