Lecture: Syntax Part 1

CS 585, Fall 2016

Introduction to Natural Language Processing

http://people.cs.umass.edu/~brenocon/inlp2016

Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

Midterm

- In-class next Thursday, 11/3. Review session on Tuesday 11/1.
- Closed book EXCEPT:
  One sheet of paper of any notes you want
  (front and back)
- Covers any material so far in the course; some subset of
  - Regular Expressions
  - Text normalization
  - Markov/N-gram Language Models
  - Naive Bayes
  - Classifiers
  - HMM
  - CRF
  - Perceptron
  - Parts of speech
  - Syntactic Parsing (this week)

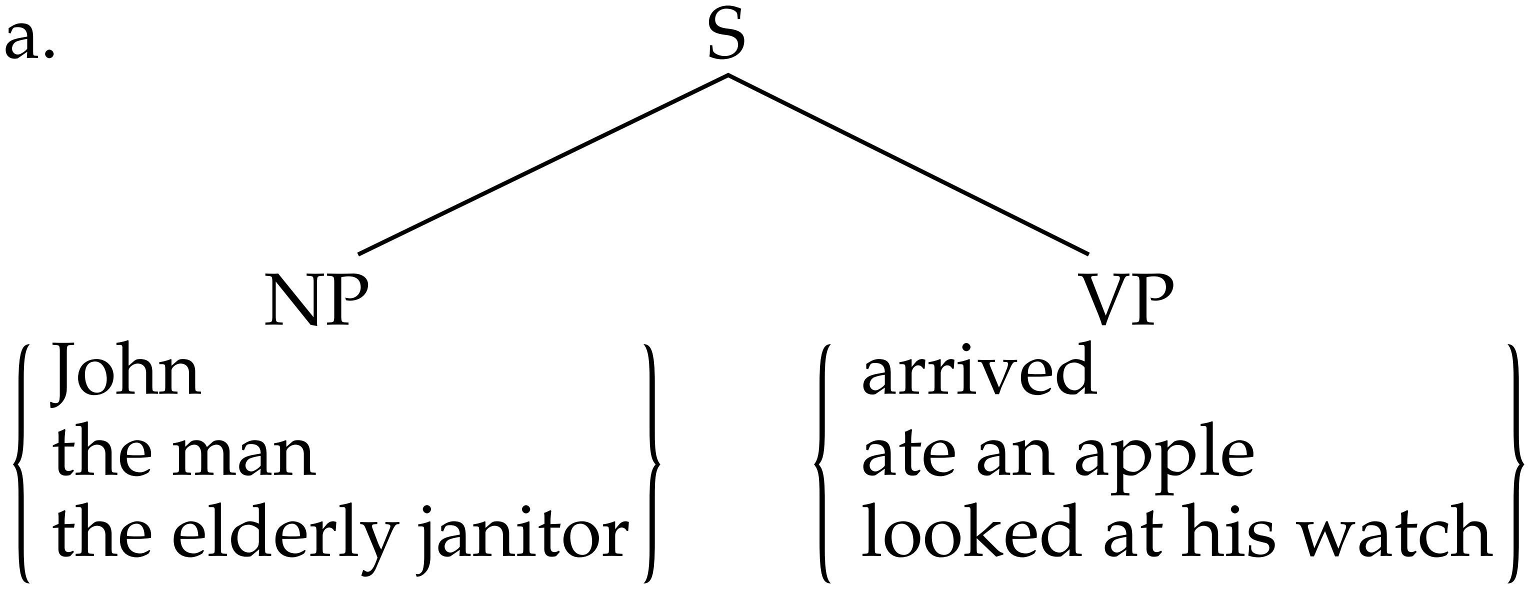
- Syntax: how do words structurally combine to form sentences and meaning?

- Order
  - dogs chase cats ..vs.. cats chase dogs
- Constituents
  - [the big dogs] chase cats
  - [colorless green clouds] chase cats
- Dependencies
  - The **dog chased** the cat.
  - My **dog**, a big old one, **chased** the cat.

- Idea of a *grammar*: global template for how sentences / utterances / phrases are formed
  - Linguistics
  - Generation
  - Parsing (structured prediction)

- “a Sentence made of Noun Phrase followed by a Verb Phrase”

a.



S

NP

VP

John
the man
the elderly janitor

arrived
ate an apple
looked at his watch

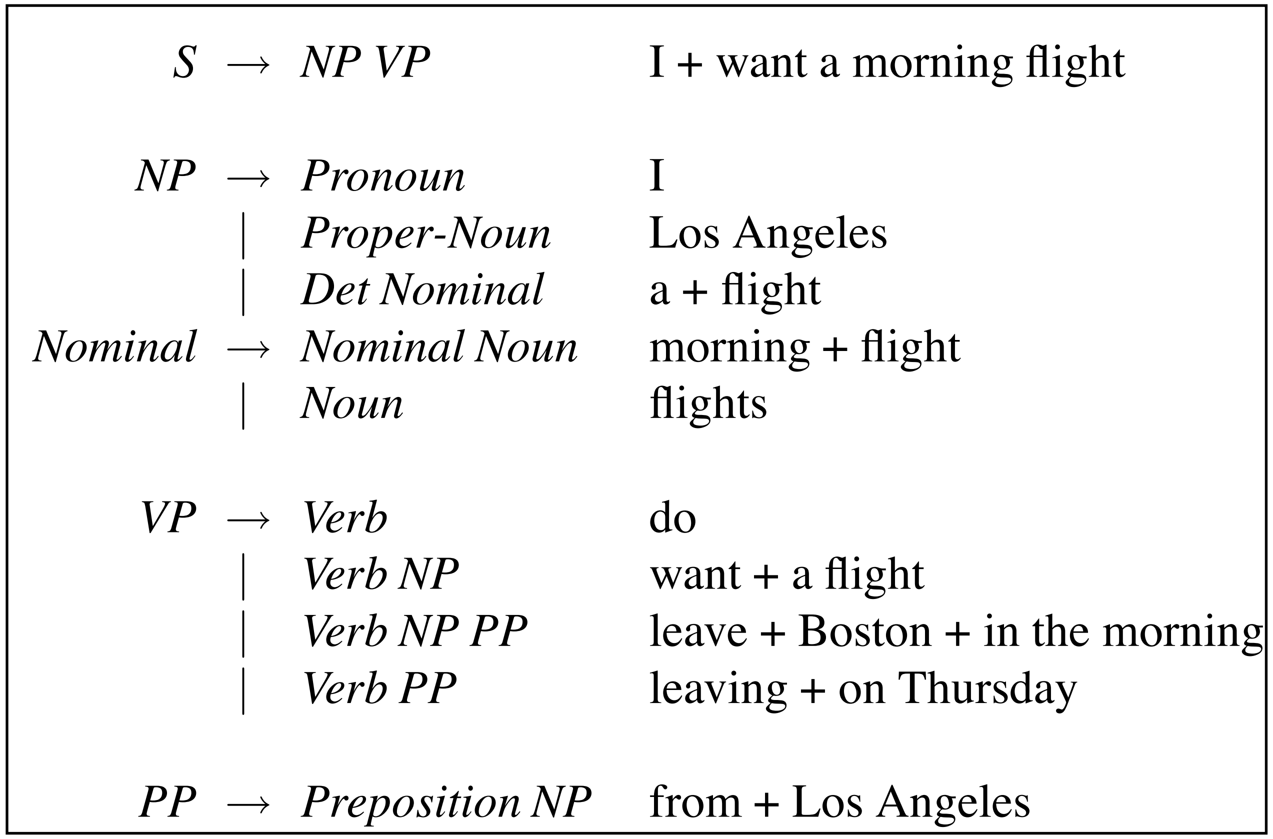
b. S → NP VP (1)

Context-Free Grammar

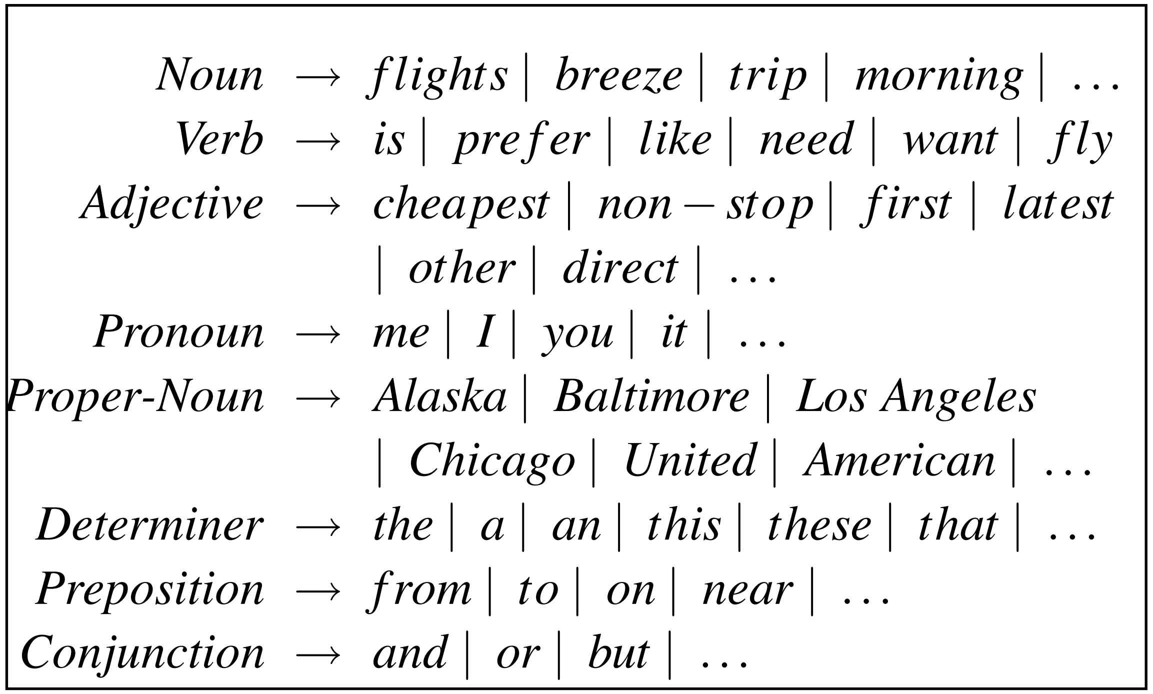
- CFG describes a generative process for an (infinite) set of strings
  - 1. Nonterminal symbols
    - “S”: START symbol / “Sentence” symbol
  - 2. Terminal symbols: word vocabulary
  - 3. Rules (a.k.a. Productions). Practically, two types:

<u>“Grammar”: *one* NT expands to >=1 NT</u>
always one NT on left side of rulep

<u>Lexicon: NT expands to a terminal</u>



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



Noun	→	*flights* | *breeze* | *trip* | *morning* | ...
Verb	→	*is* | *prefer* | *like* | *need* | *want* | *fly*
Adjective	→	*cheapest* | *non – stop* | *first* | *latest*
		| *other* | *direct* | ...
Pronoun	→	*me* | *I* | *you* | *it* | ...
Proper-Noun	→	*Alaska* | *Baltimore* | *Los Angeles*
		| *Chicago* | *United* | *American* | ...
Determiner	→	*the* | *a* | *an* | *this* | *these* | *that* | ...
Preposition	→	*from* | *to* | *on* | *near* | ...
Conjunction	→	*and* | *or* | *but* | ...

[only one token. ignore “L A”]

Constituent Parse Trees

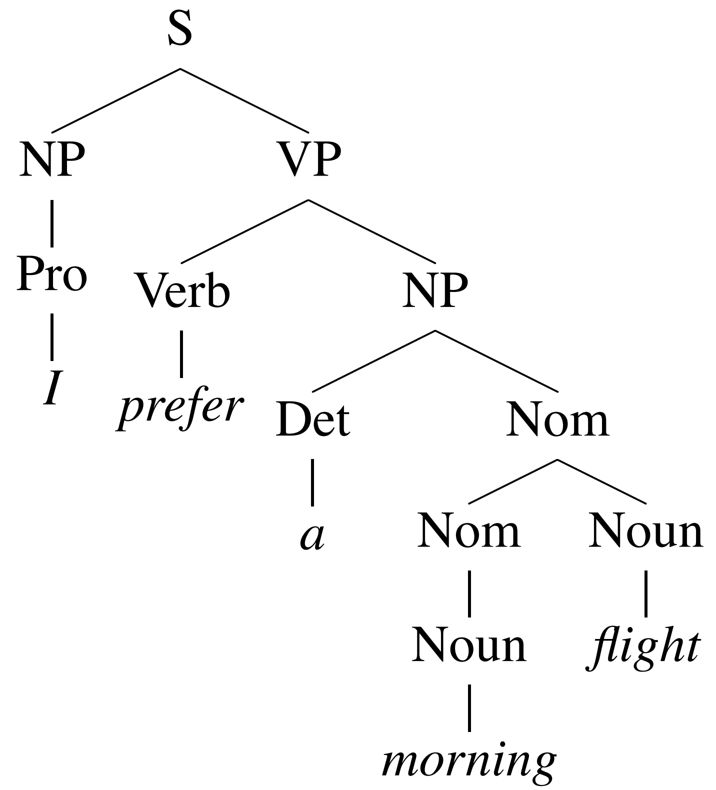


Figure 12.4 The parse tree for “I prefer a morning flight” according to grammar $\mathscr{L}_0$.

Representations:

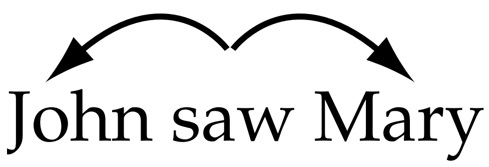
Bracket notation

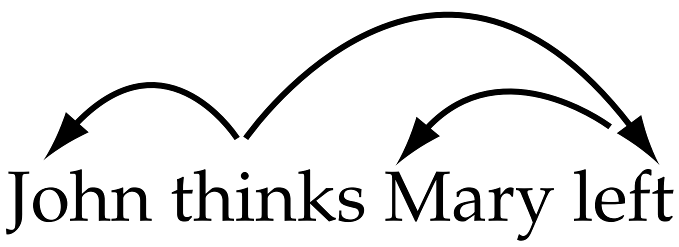
(12.2) $[_S\ [_{NP}\ [_{Pro}\ \text{I}]]\ [_{VP}\ [_V\ \text{prefer}]\ [_{NP}\ [_{Det}\ \text{a}]\ [_{Nom}\ [_N\ \text{morning}]\ [_{Nom}\ [_N\ \text{flight}]]]]]]$

Non-terminal positional spans

e.g. (NP, 0,1), (VP, 1, 5), (NP, 2, 5), etc.

Dependencies

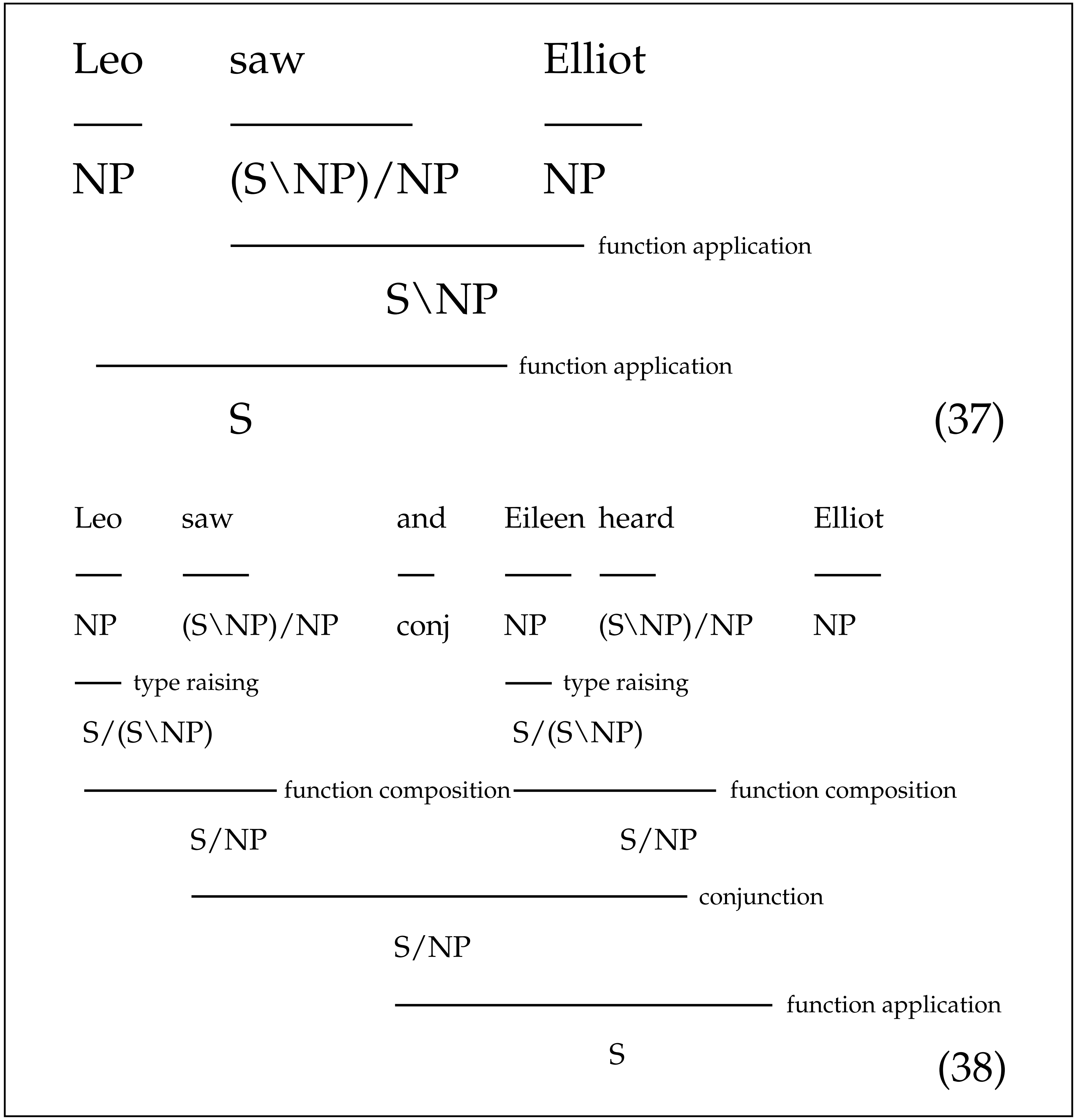




- Grammatical relationships between words in a sentence
- Most typically used type of parsing in modern NLP: e.g. "dependency bigrams" features
- e.g. http://nlp.stanford.edu:8080/corenlp/process
- Creeping toward semantics

CCG

- Combinatory Categorial Grammar
- Syntactic theory based on constituent types that can combine to left and right



Modern parsing

- Structured prediction: a tree structure
- Supervised training data: treebanks
  - Very labor intensive to create!
  - English: 1993 Penn Treebank / Wall Street Journal is still standard, more or less...
- Either constituents or dependencies
  - Dependencies are most common right now
  - Creeping towards semantics...

- Tomorrow: parsing with CFGs (most basic)