Natural Language Processing (CSE 447/547M): Phrase Structure Syntax and Parsing

Noah Smith

University of Washington
nasmith@cs.washington.edu

February 13, 2019

Finite-State Automata

A **finite-state automaton** (plural "automata") consists of:

- A finite set of states $\mathcal{S}$
  - Initial state $s_0 \in \mathcal{S}$
  - Final states $\mathcal{F} \subseteq \mathcal{S}$
- A finite alphabet Σ
- Transitions $\delta : \mathcal{S} \times \Sigma \to 2^{\mathcal{S}}$
  - Special case: **deterministic** FSA defines $\delta : \mathcal{S} \times \Sigma \to \mathcal{S}$

A string $\boldsymbol{x} \in \Sigma^n$ is recognizable by the FSA iff there is a sequence $\langle s_0, \ldots, s_n \rangle$ such that $s_n \in \mathcal{F}$ and

$$\mathbf{1}\{s_1 \in \delta(s_0, x_1)\} \wedge \cdots \wedge \mathbf{1}\{s_i \in \delta(s_{i-1}, x_i)\} \wedge \cdots \wedge \mathbf{1}\{s_n \in \delta(s_{n-1}, x_n)\}$$

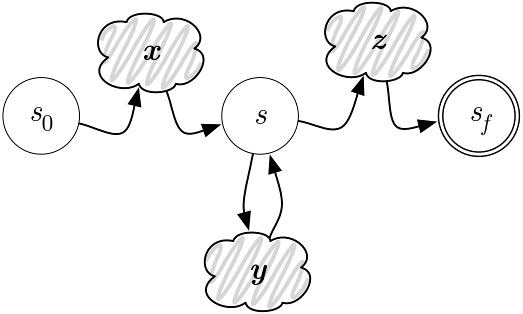
This is sometimes called a **path**.

Terminology from Theory of Computation

- A **regular expression** can be:
  - an empty string (usually denoted ϵ) or a symbol from Σ
  - a **concatentation** of regular expressions (e.g., *abc*)
  - an **alternation** of regular expressions (e.g., $ab|cd$)
  - a **Kleene star** of a regular expression (e.g., $(abc)^*$)
- A **language** is a set of strings.
- A **regular language** is a language expressible by a regular expression.
- Important theorem: every regular language can be recognized by a FSA, and every FSA's language is regular.

Proving a Language Isn't Regular

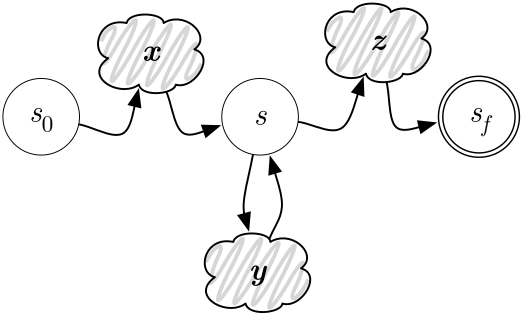
Pumping lemma (for regular languages): if L is an infinite regular language, then there exist strings $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$, with $\boldsymbol{y} \neq \epsilon$, such that $\boldsymbol{x}\boldsymbol{y}^n\boldsymbol{z} \in L$, for all $n \geq 0$.



If L is infinite and $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ do not exist, then L is not regular.

Proving a Language Isn't Regular

Pumping lemma (for regular languages): if L is an infinite regular language, then there exist strings $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$, with $\boldsymbol{y} \neq \epsilon$, such that $\boldsymbol{x}\boldsymbol{y}^n\boldsymbol{z} \in L$, for all $n \geq 0$.

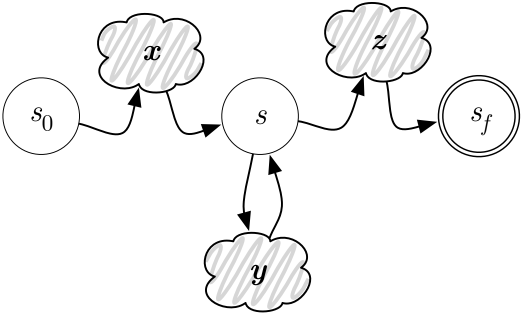


If L is infinite and $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ do not exist, then L is not regular.

If L_1 and L_2 are regular, then $L_1 \cap L_2$ is regular.

Proving a Language Isn't Regular

Pumping lemma (for regular languages): if L is an infinite regular language, then there exist strings $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$, with $\boldsymbol{y} \neq \epsilon$, such that $\boldsymbol{x}\boldsymbol{y}^n\boldsymbol{z} \in L$, for all $n \geq 0$.



If L is infinite and $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ do not exist, then L is not regular.
If L_1 and L_2 are regular, then $L_1 \cap L_2$ is regular.
If $L_1 \cap L_2$ is not regular, and L_1 is regular, then L_2 is not regular.

Claim: English is not regular.

$$\begin{aligned} L_1 &= (\text{the cat}|\text{mouse}|\text{dog})^*(\text{ate}|\text{bit}|\text{chased})^* \text{ likes tuna fish} \\ L_2 &= \text{English} \\ L_1 \cap L_2 &= (\text{the cat}|\text{mouse}|\text{dog})^n(\text{ate}|\text{bit}|\text{chased})^{n-1} \text{ likes tuna fish} \end{aligned}$$

$L_1 \cap L_2$ is not regular, but L_1 is $\Rightarrow L_2$ is not regular.

the cat likes tuna fish

the cat the dog chased likes tuna fish

the cat the dog the mouse scared chased likes tuna fish

the cat the dog the mouse the elephant squashed scared chased likes tuna fish

the cat the dog the mouse the elephant the flea bit squashed scared chased likes tuna fish

the cat the dog the mouse the elephant the flea the virus infected bit squashed scared chased likes tuna fish

Linguistic Debate

Linguistic Debate

Chomsky put forward an argument like the one we just saw.

Linguistic Debate

Chomsky put forward an argument like the one we just saw.

(Chomsky gets credit for formalizing a hierarchy of types of languages: regular, context-free, context-sensitive, recursively enumerable. This was an important contribution to CS!)

Linguistic Debate

Chomsky put forward an argument like the one we just saw.

(Chomsky gets credit for formalizing a hierarchy of types of languages: regular, context-free, context-sensitive, recursively enumerable. This was an important contribution to CS!)

Some are unconvinced, because after a few center embeddings, the examples become unintelligible.

Linguistic Debate

Chomsky put forward an argument like the one we just saw.

(Chomsky gets credit for formalizing a hierarchy of types of languages: regular, context-free, context-sensitive, recursively enumerable. This was an important contribution to CS!)

Some are unconvinced, because after a few center embeddings, the examples become unintelligible.

Nonetheless, most agree that natural language syntax isn't well captured by FSAs.

Noun Phrases

What, exactly makes a noun phrase? Examples (Jurafsky and Martin, forthcoming):

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

Constituents

More general than noun phrases: **constituents** are groups of words.

Linguists characterize constituents in a number of ways, including:

Constituents

More general than noun phrases: **constituents** are groups of words.

Linguists characterize constituents in a number of ways, including:

- where they occur (e.g., "NPs can occur before verbs")
- where they can *move* in variations of a sentence
  - On September 17th, I'd like to fly from Atlanta to Denver
  - I'd like to fly on September 17th from Atlanta to Denver
  - I'd like to fly from Atlanta to Denver on September 17th

Constituents

More general than noun phrases: **constituents** are groups of words.

Linguists characterize constituents in a number of ways, including:

- where they occur (e.g., "NPs can occur before verbs")
- where they can *move* in variations of a sentence
  - On September 17th, I'd like to fly from Atlanta to Denver
  - I'd like to fly on September 17th from Atlanta to Denver
  - I'd like to fly from Atlanta to Denver on September 17th
- what parts can move and what parts can't
  - *On September I'd like to fly 17th from Atlanta to Denver

Constituents

More general than noun phrases: **constituents** are groups of words.

Linguists characterize constituents in a number of ways, including:

- where they occur (e.g., "NPs can occur before verbs")
- where they can *move* in variations of a sentence
  - On September 17th, I'd like to fly from Atlanta to Denver
  - I'd like to fly on September 17th from Atlanta to Denver
  - I'd like to fly from Atlanta to Denver on September 17th
- what parts can move and what parts can't
  - *On September I'd like to fly 17th from Atlanta to Denver
- what they can be conjoined with
  - I'd like to fly from Atlanta to Denver on September 17th and in the morning

Recursion and Constituents

this is the house

this is the house that Jack built

this is the cat that lives in the house that Jack built

this is the dog that chased the cat that lives in the house that Jack built

this is the flea that bit the dog that chased the cat that lives in the house the Jack built

this is the virus that infected the flea that bit the dog that chased the cat that lives in the house that Jack built

Not Constituents

(Pullum, 1991)

- *If on a Winter's Night a Traveler* (by Italo Calvino)
- *Nuclear and Radiochemistry* (by Gerhart Friedlander et al.)
- *The Fire Next Time* (by James Baldwin)
- *A Tad Overweight, but Violet Eyes to Die For* (by G.B. Trudeau)
- *Sometimes a Great Notion* (by Ken Kesey)
- [how can we know the] *Dancer from the Dance* (by Andrew Holleran)

Context-Free Grammar

A **context-free grammar** consists of:

- A finite set of nonterminal symbols $\mathcal{N}$
  - A start symbol $S \in \mathcal{N}$
- A finite alphabet Σ, called "terminal" symbols, distinct from $\mathcal{N}$
- Production rule set $\mathcal{R}$, each of the form "$N \rightarrow \boldsymbol{\alpha}$" where
  - The lefthand side N is a nonterminal from $\mathcal{N}$
  - The righthand side $\boldsymbol{\alpha}$ is a sequence of zero or more terminals and/or nonterminals: $\boldsymbol{\alpha} \in (\mathcal{N} \cup \Sigma)^*$
    - Special case: **Chomsky normal form** constrains $\boldsymbol{\alpha}$ to be either a single terminal symbol or two nonterminals

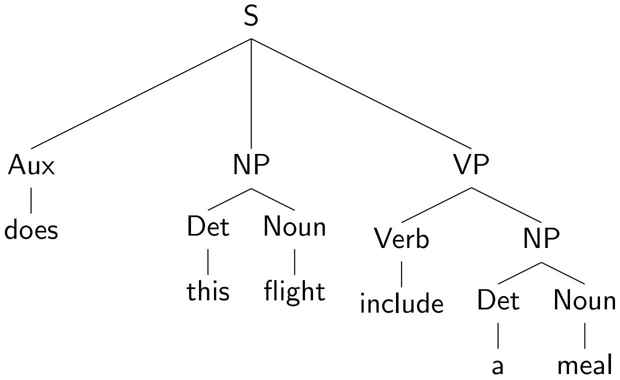
An Example CFG for a Tiny Bit of English

From Jurafsky and Martin (forthcoming)

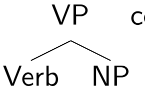
$S \rightarrow NP\ VP$
$S \rightarrow Aux\ NP\ VP$
$S \rightarrow VP$
$NP \rightarrow Pronoun$
$NP \rightarrow Proper\text{-}Noun$
$NP \rightarrow Det\ Nominal$
$Nominal \rightarrow Noun$
$Nominal \rightarrow Nominal\ Noun$
$Nominal \rightarrow Nominal\ PP$
$VP \rightarrow Verb$
$VP \rightarrow Verb\ NP$
$VP \rightarrow Verb\ NP\ PP$
$VP \rightarrow Verb\ PP$
$VP \rightarrow VP\ PP$
$PP \rightarrow Preposition\ NP$

$Det \rightarrow that \mid this \mid a$
$Noun \rightarrow book \mid flight \mid meal \mid money$
$Verb \rightarrow book \mid include \mid prefer$
$Pronoun \rightarrow I \mid she \mid me$
$Proper\text{-}Noun \rightarrow Houston \mid NWA$
$Aux \rightarrow does$
$Preposition \rightarrow from \mid to \mid on \mid near \mid through$

Example Phrase Structure Tree



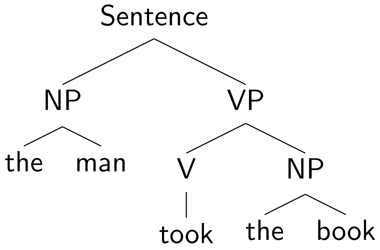
The phrase-structure tree represents both the syntactic structure of the sentence and the **derivation** of the sentence under the grammar. E.g.,



corresponds to the rule VP $\rightarrow$ Verb NP.

The First Phrase-Structure Tree

(Chomsky, 1956)



Where do natural language CFGs come from?

As evidenced by the discussion in Jurafsky and Martin (forthcoming, chapter 10), building a CFG for a natural language by hand is really hard.

Where do natural language CFGs come from?

As evidenced by the discussion in Jurafsky and Martin (forthcoming, chapter 10), building a CFG for a natural language by hand is really hard.

- Need lots of categories to make sure all and only grammatical sentences are included.

Where do natural language CFGs come from?

As evidenced by the discussion in Jurafsky and Martin (forthcoming, chapter 10), building a CFG for a natural language by hand is really hard.

- Need lots of categories to make sure all and only grammatical sentences are included.
- Categories tend to start exploding combinatorially.

Where do natural language CFGs come from?

As evidenced by the discussion in Jurafsky and Martin (forthcoming, chapter 10), building a CFG for a natural language by hand is really hard.

- Need lots of categories to make sure all and only grammatical sentences are included.
- Categories tend to start exploding combinatorially.
- Alternative grammar formalisms are typically used for manual grammar construction; these are often based on constraints and a powerful algorithmic tool called *unification*.

Where do natural language CFGs come from?

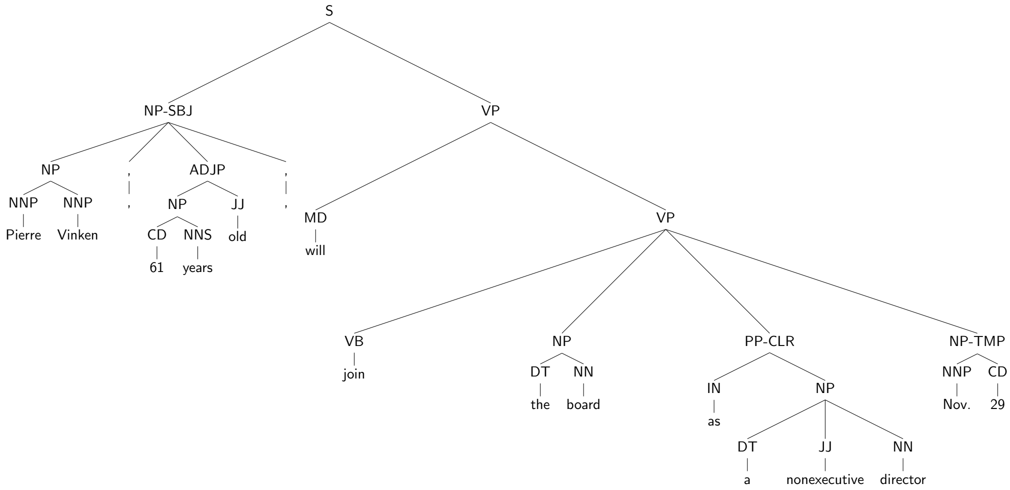
As evidenced by the discussion in Jurafsky and Martin (forthcoming, chapter 10), building a CFG for a natural language by hand is really hard.

- Need lots of categories to make sure all and only grammatical sentences are included.
- Categories tend to start exploding combinatorially.
- Alternative grammar formalisms are typically used for manual grammar construction; these are often based on constraints and a powerful algorithmic tool called *unification*.

One approach:

1. Build a corpus of annotated sentences, called a **treebank**. (Memorable example: the Penn Treebank, Marcus et al., 1993.)
2. Extract rules from the treebank.
3. Optionally, use statistical models to generalize the rules.

Example from the Penn Treebank



LISP Encoding in the Penn Treebank

```
( (S
    (NP-SBJ-1
      (NP (NNP Rudolph) (NNP Agnew) )
      (, ,)
      (UCP
        (ADJP
          (NP (CD 55) (NNS years) )
          (JJ old) )
        (CC and)
        (NP
          (NP (JJ former) (NN chairman) )
          (PP (IN of)
            (NP (NNP Consolidated) (NNP Gold) (NNP Fields) (NNP PLC) ))))
      (, ,) )
    (VP (VBD was)
      (VP (VBN named)
        (S
          (NP-SBJ (-NONE- *-1) )
          (NP-PRD
            (NP (DT a) (JJ nonexecutive) (NN director) )
            (PP (IN of)
              (NP (DT this) (JJ British) (JJ industrial) (NN conglomerate) ))))))
    (. .) ))
```

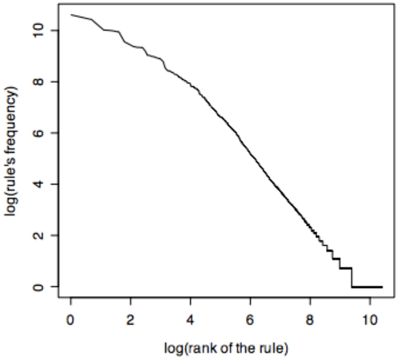
Some Penn Treebank Rules with Counts

40717 PP → IN NP
33803 S → NP-SBJ VP
22513 NP-SBJ → -NONE-
21877 NP → NP PP
20740 NP → DT NN
14153 S → NP-SBJ VP .
12922 VP → TO VP
11881 PP-LOC → IN NP
11467 NP-SBJ → PRP
11378 NP → -NONE-
11291 NP → NN
. . .
989 VP → VBG S
985 NP-SBJ → NN
983 PP-MNR → IN NP
983 NP-SBJ → DT
969 VP → VBN VP

100 VP → VBD PP-PRD
100 PRN → : NP :
100 NP → DT JJS
100 NP-CLR → NN
99 NP-SBJ-1 → DT NNP
98 VP → VBN NP PP-DIR
98 VP → VBD PP-TMP
98 PP-TMP → VBG NP
97 VP → VBD ADVP-TMP VP
. . .
10 WHNP-1 → WRB JJ
10 VP → VP CC VP PP-TMP
10 VP → VP CC VP ADVP-MNR
10 VP → VBZ S , SBAR-ADV
10 VP → VBZ S ADVP-TMP

Penn Treebank Rules: Statistics

32,728 rules in the training section (not including 52,257 lexicon rules)
4,021 rules in the development section
overlap: 3,128



(Phrase-Structure) Recognition and Parsing

Given a CFG $(\mathcal{N}, S, \Sigma, \mathcal{R})$ and a sentence $\boldsymbol{x}$, the **recognition** problem is:

Is $\boldsymbol{x}$ in the language of the CFG?

Related problem: **parsing**:

Show one or more derivations for $\boldsymbol{x}$, using $\mathcal{R}$.

(Phrase-Structure) Recognition and Parsing

Given a CFG $(\mathcal{N}, S, \Sigma, \mathcal{R})$ and a sentence $\boldsymbol{x}$, the **recognition** problem is:

Is $\boldsymbol{x}$ in the language of the CFG?

The proof is a derivation.

Related problem: **parsing**:

Show one or more derivations for $\boldsymbol{x}$, using $\mathcal{R}$.

(Phrase-Structure) Recognition and Parsing

Given a CFG $(\mathcal{N}, S, \Sigma, \mathcal{R})$ and a sentence $\boldsymbol{x}$, the **recognition** problem is:

Is $\boldsymbol{x}$ in the language of the CFG?

The proof is a derivation.

Related problem: **parsing**:

Show one or more derivations for $\boldsymbol{x}$, using $\mathcal{R}$.

With reasonable grammars, the number of parses is exponential in $|\boldsymbol{x}|$.

References I

Noam Chomsky. Three models for the description of language. *Information Theory, IEEE Transactions on*, 2(3): 113–124, 1956.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, third edition, forthcoming. URL `https://web.stanford.edu/~jurafsky/slp3/`.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Geoffrey K. Pullum. *The Great Eskimo Vocabulary Hoax and Other Irreverent Essays on the Study of Language*. University of Chicago Press, 1991.