CS 4510 Automata and Complexity January 31st 2024

Lecture 6: Syntactic Structures

Lecturer: Abrahim Ladha *Scribe(s): Akshay Kulkarni*

1 Linguistics

1.1 Language

Linguistics is the study of language. This includes its structure, etymology, history, everything. Its systematic. Fundamentally, language is communication. Suppose we have a listener and a speaker, and the speaker says “horse”. The way we represent that idea syntactically, as a word or utterance, can have nothing to do with the idea itself. That’s language. Language is an agreement we all have about what things mean. “Horse” is nothing. “Horse” is an utterance I make by moving vocal cords and exhaling air at a frequency you pick up with your ears.

The concept, the semantic value of “horse” is vastly more complicated than the sound or word itself. I say or utter “horse” and in our mind, you create an image of a herbivore quadruped, maybe its brown. From an information theory perspective, this is a lossy channel, as the idea contains more information than the word. Yet, we are limited to express ideas only through the use of language.

1.2 Why Only Us

While animals have certainly demonstrated ability to understand and mimic language to a reasonable extent, humans are the only beings capable of implementing complex language. You can talk to animals, but they won’t talk back. Some argue that language is central to humanity’s evolutionary identity. As a species, or ability to communicate ideas is exactly what makes us human. It wasn’t the upright posture, larger prefrontal cortex, opposable thumbs, tool-making, no. It was the ability for us to come to consensus and work together on complex tasks. If I, as some paleolithic ape generate the idea “I go hunt mammoth”. This idea is totally worthless, I go hunt mammoth the outcome is I get trampled. Instead I communicate this idea and now its “we go hunt mammoth”, suddenly its more serious.

First we took down a mammoth, and second, we built a computer. There are not that many steps between those two events.

We also have the ability to communicate and pass down our wisdom. Every next generation stands on the shoulders of those before. The animals ability to pass down learned experiences is mostly through instinct, which comes through natural selection, and perhaps several generations of death. Animals may Pavlovianly condition their behavior, but like everything else, are mortal. Their offspring will relearn many things the hard way.

Since all ideas must be expressed as language, it was an old-world view that the study of language itself was the only way to study ideas. The study of language was the study of everything. One of Chomsky's accomplishments was to help separate these two. Syntax, the structure of language, and semantics, the meaning of language, are not interchangable.

1.3 The Computational Lens on the Sciences

The importance of Chomsky's short monograph was not that he solved language in general, but rather he came into someone elses house with more math than them. He came into a very empirical field, and brought in an as theoretical as possible perspective. Using relatively simple, intuitive, but undeniable arguments, he was able to make true generalizations about what is an incredibly complex system.

1.4 Chomskyan View of Language

Consider a baby. It it not born speaking any language. Googoo gaga and so on. Totally unintelligible. Although a baby is not born knowing any language, it somehow knows how to learn a language. Airdrop that baby into a group of people speaking a language and as it mentally develops, it will learn how to speak among them. This would be independent of anything about the structure of the language. It would not necessarily learn in school what nouns and verbs are in order to speak. Chomsky's view is that syntax is an innate aspect of language determined by a "Universal Grammar". There are biological conditions which shape the structure of language. The way our brains have the wires and pipes cause limitations in what possible structures language must take.

To study language, it is okay for us to limit ourselves to english. Languages share many universal features. For example, delimination with a space. Have you ever thought about why sentences come in lists and not trees or some other structure in which may not have a kind of topological sort? All grammatical operations appear to be binary as well. The set of grammatical sentences appears to be infinite, but appears to be constructed recursively from some set of finite atomic pieces, like a basis. In english, that would be our alphabet, Σ. Secondly, for any english specific artifacts we may encounter, there are almost certainly analoguous discussions in other languages.

A language learner can only learn through experience, by hearing only finitely many samples. Yet, they develop the abiliity to generate new sentences, ones that have never been uttered before. Some sort of generative device is trained behind the veil which through its ability to parse grammatically correct sentences, also can be tasked with producing grammatically correct sentences. As we develop a theoretical model for grammar, it is sufficient for us to focus on the ability of distinguishing the grammatical from the ungrammatical. Such a device or structure which can help us separate these two, can also help us generate

Figure 1: A Platonic Man

new grammatical sentences. Similarly how we may study decision problems, but understand their relationship to algorithms and search problems.

2 Syntactic Structures

The monograph is only ninety pages, and I highly recommend you read it yourself. We paraphrase the first six sections.

2.1 Introduction

To explain why a rigorous foundational undertaking on distinguishing the grammatical from the ungrammatical is a hard problem, I will first tell you a story.

You may recall the parable of Diogenes of Sinope and Plato. Plato was an established and respected thinker, and Diogenes was just some guy who lived in a barrel. Plato announces he has devised the definition of “man”. All ancient Greek philosophy is this kind of simple stuff. Plato’s definition of man is that which is a “featherless biped”. Diogenes crashes the class, holds up a plucked chicken, and announces: “Behold! A man!”.

The moral here is that the definition provided by Plato was insufficient, and there was a readily available counter example. A man is a concept which is difficult to formalize, but everyone has an understanding on what a person is. It is an intuitive notion, and Plato attempted to give a formal one. A better definition would perhaps be like “A man to Plato is what Plato says is a man”.

There are many intuitive concepts we deal with which are difficult to formalize. Grammatical and ungrammatical is such a concept. We may understand a sentence to be grammatical or ungrammatical without appealing to a system of formal rules. This is just the power acquired by being a member of a language speaking community. Given a formal language like $L = \{a^n b^n \mid n \in mathbbN\}$ and a word w, you can easily, rigorously, and unambiguously determine when $w \in L$. But english does have such a nice simple description with set builder notation. The concept of grammatical and ungrammatical is an intuitive, possibly ambiguous one. We may attempt to connect it to a formal unambiguous concept. Like how Plato failed to connect the intuitive concept of humanity with the formal concept of a plucked chicken.

Even defining "language" formally can be problematic. Aristotle defined language as sound with meaning, which, in classic Greek fashion, is too vague to argue with.

2.2 The Independence of Grammar

Consider the following two sentences:

1. Colorless green ideas sleep furiously.

2. Furiously sleep ideas green colorless.

2.2.1 Syntax and Semantics

Let us study the first sentence. Certainly it is grammatical. Somehow in our brain exists a distinguisher, and we can read this sentence and come to the consensus that it is grammatically correct. Next of note is that the sentence is totally devoid of meaning. What would the subject the sentence be? Ideas? and they somehow can sleep? and do so furiously? Its colorless, yet green? It has no semantical value, and does not communicate any idea (besides perhaps confusion). This is a kind of counter example, and forces us to separate syntax, the structure of language, from semantics, the meaning of language. While this example is a grammatical sentence with no meaning, you should be able to come up with numerous examples of sentences full of meaning, which are ungrammatical.

2.2.2 Probabilistic Models

Second, note that the first sentence is grammatical, but the second one is not. The first, simply by intonation, cadence, and word pattern seems comfortable. It would be easier to remember and recite. The second is ungrammatical, and troubling. Yet, these two sentences, the frequency of sequential word choices is equivalent because one is a word reversal of the other. These two sentences have been uttered equally likely in all of english, that is, a negligible amount[1]. This is our second observation. The ability of a syntactic structure to distinguish the grammatical from the ungrammatical must be independent of the sentences proximity to english. The first being grammatical, and the second not. Any model based on probability may be unable to distinguish these two based on this kind of frequency alone, but we argue, must be able to. Most sentences are statistically infrequent, in that they have never been said before. Even the sentences you are reading now. Something like 15% of Google's daily searches have never been searched before.

[1] It is ironic that Chomsky chose this sentence because it would be statistically infrequent, but by choosing it as an example, he has made it very famous. `https://en.wikipedia.org/wiki/Colorless_green_ideas_sleep_furiously`

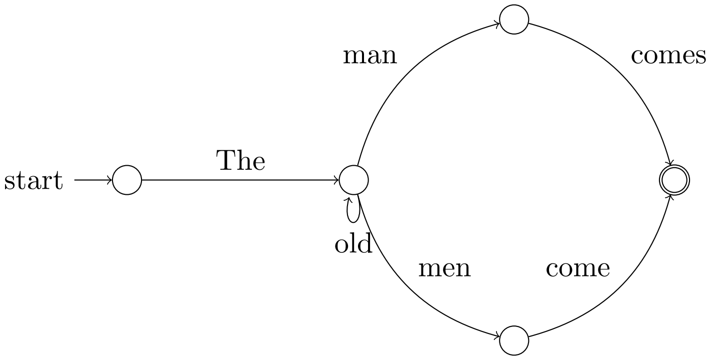
This is an argument of Chomsky which has not aged as well over the past 70 years[2]. You may be very comfortable with the fact there are now many probabilistic models capable of generating syntactically valid sentences. We may argue the factual correctness of the output of these systems, but the fact that the output is grammatically correct is undeniable. I do feel like it is cheating a bit. We measure the performance of algorithms on inputs much larger than the algorithms themselves. But we measure the accuracy and performance of these models on inputs and outputs much smaller than the terrabytes of weights they require.

Chomsky is arguing here against a very limited kind of probabilistic model. First, likely some sort of formula, which may take as input some number as a probability of the sentence being spoken, and outputs a number representing some certainty that the sentence is grammatical. Second, likely a Markov process. Imagine an NFA with probabilities on the transitions as well. The capabilities of randomized algorithms had not yet been known. Quicksort would not be invented for another two years.

2.3 An Elementary Linguistic Theory

2.3.1 English Contains some Regular Substructure

Natural languages are not formal formal languages, but we can apply similar arguments. Here we show a substructure of english has some similar structure to a regular language. For example, the following DFA[3] can be used to model the formation of a substructure of english.



We are not concerned with the study of finite languages, but of infinite ones decidable by finite structures. Here, this decides an infinite language because it has a point of recursion. It may be impolite to describe someone as “old old old old...”, but it is not ungrammatical, it is a hyperbole.

2.3.2 English is Not Regular

First, recall our three non-regular languages

- $\{a^n b^n \mid n \in \mathbb{N}\}$

[2] For a great summary of the arguments of his critics, see Norvig here `https://norvig.com/chomsky.html`

[3] Chomsky calls this a Finite-State Markov Process (FSMP)

- $\{ww^R \mid w \in \Sigma^*\}$
- $\{ww \mid w \in \Sigma^*\}$

We show by an analogy, that there does not exist a DFA for a substructure of english.

Proof. Let $S_1, S_2, ...$ be declarative sentences. Let S be the sentence defined as

$$S := \text{`` If } S_i \text{ Then } S_j\text{''}$$

S is a declarative sentence like any other. We may take S_i to be S then, and we may derive

$$\text{`` If If } S_i \text{ Then } S_j \text{ Then } S_j\text{''}$$

Observe that upon repetition of this n times, we get

$$\text{``}(\textsc{if})(\textsc{if})(\textsc{if})(\textsc{if}) \ldots S_i \ (\textsc{then})(\textsc{then})(\textsc{then})(\textsc{then}) \ldots S_j\text{''} = (\textsc{if})^n S_i(\textsc{then}\ S_j)^n$$

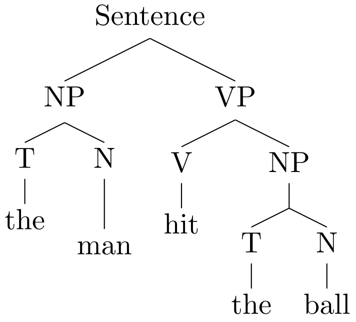
This is quite similar to our first known canonical non-regular language, $a^n b^n$. We proved that language was non-regular by pumping, and similarly here, there would not exist a DFA for this substructure of english. Another good example is the Dyck language, the set of balanced parenthesis. One would recognize if a sentence had unbalanced parenthesis and distinguish it as ungrammatical. An an example, consider "((hello there.)(((". □

2.4 Phrase Structure

It had been known for centuries of the recursive nature of language. How sentences can be built from fragments, fragments from words, words from letters, and so on. Sentences have a hierarchical structure, and this structure is governed by the rules from grammar. Chomsky formalized these observations to justify what his next model of study was, and why it was ideal. He defines something called a phrase structure, which is a generalization of what we now call a context-free grammar. For now, lets suppose phrase structures are just CFGs. We can remark that this device is incredibly useful as generative model for language. Consider the following model:

$$\begin{aligned}
\textsc{Sentence} &\to \textsc{Noun Phrase} + \textsc{Verp phrase}\\
\textsc{Noun Phrase} &\to \textsc{Article} + \textsc{Noun}\\
\textsc{Verb Phrase} &\to \textsc{Verb} + \textsc{Noun phrase}\\
\textsc{Articles} &\to \{\text{a, the, etc.}\}\\
\textsc{Noun} &\to \{\text{man, men, ball, etc.}\}\\
\textsc{Verb} &\to \{\text{hit, took, etc.}\}
\end{aligned}$$

Here, we have a phrase structure for a declarative substructure of english. A production of a sentence from our phrase structure can be expressed as a parse tree.



Its notable here that a parse tree gives less information than a list of productions, as just from the tree, you don't know what order the rules were applied in. Both the parse tree and the order of productions also provide more information than the sentence itself. The generated structure is not available to those who simply view the sentence. This creates ambiguity during synthesis, as language is only communicated in a linear fashion. Ambiguity of a sentence, in this model, could be interpreted as multiple correct parsings of the same sentence.

2.5 Limitations of our Phrase Structure

Although we note that this generates grammatical sentences, it can also generate ungrammatical ones. This example is with respect to singular and plural words.

1. "The man hit the ball."
2. "A men hit the ball."

The second sentence is clearly not grammatically correct.

Chomsky: "We must be able to limit the application of a rule to a certain context"

A context-free grammar is quite literally, free of context. There is no restriction or limitation when you may apply a production. If you have a set of productions like $N \to \{$ nouns$\}$, then you can substitute in any noun. Like mad libs, it may not be grammatical. We want to consider applications of rules which are sensitive to context. A production can only be applied if conditions are met on the part of the working string before and after the substring we would insert. We define a context-sensitive grammar and compare it to previous models:

Model	Example rule
Regular grammars	$A \to bE$
Context-free grammars	$A \to bCdEf \dots$
Context-sentitive grammars	$xAz \to xyz$

Here, $x, z \in (V \cup \Sigma)^*$. You can only make the substitution $A \to y$ when in the current working string, A is preceded by x and followed by z. These types of rules are called context-sensitive, because they are quite literally, sensitive to context. They are a generalization of the kinds of productions allowed by CFGs, as the left-hand-side can now contain more than one non-terminal. They are strictly stronger than context-free grammars, and we will not spend any more time on them. For our small piece of english we are studying, we can modify the phrase structure with context sensitive rules to solve our issue with singular and plural words as follows.

$$\begin{aligned}
NP &\to NP_s + V \mid NP_p + V \\
NP_s + V &\to NP_s + V_s \\
NP_p + V &\to NP_p + V_p \\
NP_s &\to T_p + N_p \\
NP_p &\to T_s + N_s \\
T_s &\to \text{a} \\
T_p &\to \text{the} \\
N_s &\to \text{man} \\
N_p &\to \text{men} \\
V_s &\to \text{hit} \\
V_p &\to \text{hits}
\end{aligned}$$

Here, N_p, a non-terminal for plural nouns, cannot be preceded by T_s, singular articles. This makes the ungrammatical production of “a men” impossible.

I highly suggest you read Syntactic Structures in full. This is a high level overview of some of the simpler and early theorems made, and how they were argued.

2.6 On the Goals of Linguistic Theory

Chomsky argues that foundational theories about language should have the same desirable traits as those required by physics or chemistry. A scientific theory must not only explain all current observations and phenoma, but be able to make future predictions. Occam's razor should apply, and the theory should be able to explain the occurrence of events with as simple reasoning required.

2.7 Further Reading

- Syntactic Structures by Noam Chomsky
- Poverty of the Stimulus
- How To Know What Words Mean - Troublehacking with Drew Cleary

3 Chomsky Normal Form

Given a word w and a grammar G, is $w \in L(G)$? This is surprisingly non-trivial. Unlike an automata which reads the word as input and determines yes or no, a grammar must nondeterministically produce only the correct strings. Determining if a string is or isn't produced by a grammar isn't an obvious problem then.

3.1 Definition

We say a CFG is in Chomsky Normal Form (CNF) if it has productions only of the form:

$$A \to BC$$
$$A \to a$$

where the capital letters are any non-terminals, and the lower-case letters are any terminals. Additionally B, C cannot be the start state. and the rule $S \to \varepsilon$ is present if and only if $\varepsilon \in L(G)$.

Note that obviously $\mathscr{L}(CNF) \subseteq \mathscr{L}(CFG)$. Any context-free grammar in CNF is still a context-free grammar. We have a process to convert any CFG into CNF form, proving that $\mathscr{L}(CFG) = \mathscr{L}(CNF)$.

1. Add a new start State $S_0 \to S$. Now every rule will not have the start state anywhere on the right-hand-side.
2. Delete and patch all $A \to \varepsilon$ rules. For example if you have rules $R \to uAv, A \to \varepsilon$, you now have rules $R \to uAv \mid uv$.
3. Remove all unit rules $A \to B$ (i.e. $(A \to B, B \to C) = A \to C$
4. Convert rules of length greater than two into a chain of rules as follows.

   $$(A \to u_1 \ldots u_k) \to (A \to u_1 A_1, A_1 \to u_2 A_2, \ldots, A_{k-1} \to u_{k-1} u_k)$$

   where $u_1, ..., u_k$ can be terminals or non-terminals.
5. $\forall a \in \Sigma$, replace a in any right-hand-side of every production with new nonterminal A and add production $A \to a$.

Steps three and four may need to be repeated many times because applying one patch may introduce a need for another.

3.2 Advantages of CNF

Lets prove that if a word of length $n \geq 1$ is produced by a grammar in CNF, it takes exactly $2n - 1$ productions. Lets work backwards.

$$w_1 ... w_n \overset{*}{\Longleftarrow}_1 W_1 ... W_n \overset{*}{\Longleftarrow}_2 S$$

- The last productions (1) goes from n terminals to n non-terminals. At each production, exactly one non-terminal is replaced by exactly one terminal, so this takes n productions.

- For (2), to go from n non-terminals to one terminal, our start terminal, requires $n-1$ productions. Every rule of a grammar in CNF takes one non-terminal, and adds two. So for each production, if non-terminals are added, a production adds exacly one.

Combined, we see that a grammar in CNF form will take exactly $2n-1$ productions to produce a word of length n. This is the point of CNF, the limited structure allows us to have a guarantee for an algorithm. We can now determine for any context-free-grammar G and word w if $w \in L(G)$. Convert your grammar to CNF, compute all possible productions of exactly $2n-1$ steps, and your candidate word is in this list $\iff w \in L(G)$.

3.3 Example

Consider the following conversion of a general CFG to one in CNF. Each line represents a transformation of the grammar.

$$\begin{aligned}
&S \to aSb \mid \varepsilon \\
&S_0 \to S,\ S \to aSb \mid \varepsilon \\
&S_0 \to S \mid \varepsilon,\ S \to aSb \mid ab \\
&S_0 \to aSb \mid ab \mid \varepsilon,\ S \to aSb \mid ab \\
&S_0 \to aX \mid ab \mid \varepsilon,\ S \to aX \mid ab,\ X \to Sb \\
&S_0 \to AX \mid AB \mid \varepsilon,\ S \to AX \mid AB,\ X \to SB,\ A \to a,\ B \to b
\end{aligned}$$

This grammar should produce $aabb \in \{a^n b^n \mid n \in \mathbb{N}\}$. Lets verify it takes $2n-1 = 7$ productions.

$$S_0 \overset{1}{\Longrightarrow} AX \overset{2}{\Longrightarrow} ASB \overset{3}{\Longrightarrow} AABB \overset{4}{\Longrightarrow} aABB \overset{5}{\Longrightarrow} aaBB \overset{6}{\Longrightarrow} aabB \overset{7}{\Longrightarrow} aabb$$