A Primer on Syntax and Context-free Grammars

Johnny Tian-Zheng Wei

September 15, 2018

1 Introduction

Language is often thought to be arbitrary. It is true that there is no limit to the ideas that we may express. The quote below

(1) The ships hung in the sky, much the way that bricks don't.

from Douglas Adams's book, the Hitchiker's Guide to the Galaxy, clearly conveys an imagined scenario that has yet been witnessed by anyone in the past, present, or future. However, a slight rearrangement of (1) below

(2) The ships hung the in sky, much the way that bricks don't.

now no longer sounds like it could have been a sentence Douglas Adams had written - a typo perhaps. Even a small vocabulary provides a combinatorial explosion of possible sequences, yet English is only a small fraction of those. As you and I both can quickly identify the membership of sequences of words within the restricted subset, there must be rules governing what can and cannot be English.

In this tutorial, the main formalism of interest is the context-free grammar, which is a model that recognizes which sentences should be part of English, and which sentences are not. In motivating context-free grammars, we will first give a primer on syntax. Then, We will provide a simple example of a context-free grammar and study its properties.

1.1 A Syntax Primer

Syntacticians, linguists who study syntax, define syntax roughly as

The laws that govern how words combine into sentences.

and presupposes what words and sentences are. [1] Speakers can agree on most cases of what these are, and we will begin our discussion with our shared notions of words and sentences. The three sentence below highlight the phenomenon that concerns syntax

(3) The cat in the hat is delighted.

(4) The cat understands that emotional logic floats.

(5) * The in cat the hat is delighted.

where there is a salient difference between (5) and (3, 4), as the former example is a sentence that is inconceivable to be spoken by any English speaker. In the latter two examples, we can agree both are speakable sentences. The study of syntax is built on the assumption that this difference is real, and we are mainly interested in systems that account for this difference. We define (3) and (4) to be *grammatical*, but (5) to be ungrammatical, and use a preceding "*" to denote the difference. Note that there is also a salient difference in (4) and (3), where (4) is non-sensical. Syntax is not interested in this distinction and we leave such details to semanticists.

In (3,4,5) I decided the grammaticality of sentences for illustration purposes, a methodology syntacticians term grammaticality judgments. There are two appealing properties of using grammaticality judgments like the ones above to study syntax. First, as an English speaker, I can reliably posit that since (5) is unspeakable to myself, other English speakers will share my intuition (which you will see throughout my tutorial!). Second, we may test sentences with phenomena that are the "edge cases" of our current knowledge of syntax. For instance, the judgments below

(6) The cat sings.

(7) The cat sings songs.

might lead us to hypothesize that when "songs" is appended to a grammatical sentence, it continues to be grammatical. A more capable syntactician could think of the counter-examples

(8) The cat sings songs.

(9) * The cat sings songs songs.

which shows, in fact, that our hypothesis need to be revised. This is the bitter sweet moment when we realize that an underlying mechanism is not what it seemed. Grammaticality judgements are the syntactician's main tool for refining our hypothesis of language syntax.

1.2 Observation of Compositionality

With grammaticality judgments, we can make our first set of general observations about the syntax of language. In §1.1 we defined syntax to be laws that dictate how words combine into sentences. I focus on combinations, or compositions, of words as the correct perspective towards syntax *a priori*. Hopefully, this section will convince you of the compositional perspective, and the curious reader may be interested in §3 for results in formal language theory for further justification.

Consider two short sentences

(10) Unicorns like cats.

(11) * Unicorns cats like.

where (11) is an ungrammatical permutation of the grammatical sentence (10). It is certainly not the case that the combination of "unicorns", "cats", and "like" is unmeaningful, as (10) is our grammatical counter-example that uses the same three words. Evidently, the order of the words in (11) did not allow for a *composition* as in (10).

We should begin our observations from words, the elementary units in our definition of syntax. Let's extend our notion of grammaticality contiguous subsequences of sentences, or fragments, as to whether they have compositional meaning. For the sentence

(12) The cat rides real unicorns.

we list some of the fragments and judge their grammaticality below

(13) The cat

(14) * cat rides a

(15) rides real unicorns

(16) real unicorns

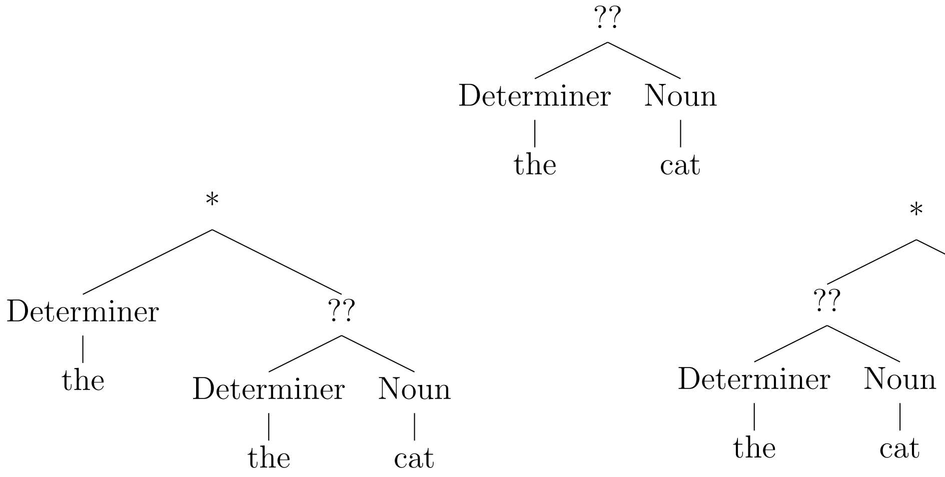
where my judgments reflect whether I think the fragment is grammatical. That is, (13, 15, 16) have some composition of meaning, even when it is non-sensical as in (16).

A game of madlibs is appropriate to deduce the properties of compositions that are grammatical. To start our investigation, we will hold our right-hand word as a constant "the". If I have the fill-in-the-blank question as

$$\underline{\text{the}}\ \underline{\qquad} \tag{1}$$

what words may go in the right-hand blank? In (13) we know that "cat" is one. My intuition further tells me "unicorn", "girl", or "fish" also allow for grammatical compositions. We see that this right-hand blank seems to permit composition when the proposed word is a noun (excluding proper nouns). Further, the left-hand blank can be replaced with determiners such as "a" or "one". In general, a determiner followed by a noun composes to form a grammatical fragment. The composed fragment is termed a *constituent* in linguistics.

We will now investigate the compositional properties of one such constituent "the cat" with the form above. Is this constituent a noun or a determiner? We can play the same madlib game to deduce the answer. Interestingly, "the cat" somehow has a category that is not accepted by (1), in either the right or left-hand blank. It seems like



where the constituent formed by a determiner and noun, labelled by "??", is evidently neither a noun or a determiner, or else "the the cat" and "the cat dog" would be grammatical. The trees above are intended to show the hierarchical composition of constituents, and "*" in this context denotes a non-existent composition rule, which explains the ungrammaticality of the fragments. Lets give the constituent "??" the name "Noun Phrase" as the meaning of the composed phrase is mainly carried by the noun.

We just deduced a (very simple) syntactic rule! This syntactic rule tells us that a sequence of a "Determiner" and a "Noun" can combine, but the combination of the two takes on a new category "Noun Phrase". We might decide to write down our rule in our theory of syntax as a three-tuple

$$\text{(Noun Phrase, Determiner, Noun)}$$

which denotes that a determiner and noun can form a noun phrase.

1.3 More Rules

The observation of compositionality gives us a starting point to write grammars, a system of rules that describe what is grammatical, and what isn't. To build a small grammar that explains the grammaticality of a subset of English sentences containing a transitive verb, I will quickly list two extra rules with examples.

In addition to the rule derived from §1.2, we have grammatical fragments

(17) kisses the cat

(18) kisses a unicorn

to deduce that (Verb Phrase, Verb-Transitive, Noun Phrase) is a rule of our grammar. Finally, we observe that full grammatical sentences can be formed as in

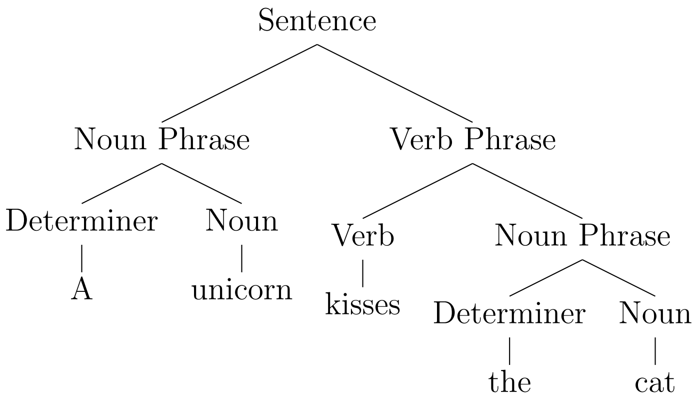
(19) A unicorn kisses the cat.

when two fragments apply the rule (Sentence, Noun Phrase, Verb Phrase) to compose into a sentence, and no further rule is applied.

With these three rules, we may sketch a rudimentary grammar. Let's say that our grammar decides a sentence is grammatical, if it is a composition of two grammatical noun phrase and verb phrase fragments. The definition of a grammatical noun or verb phrase then depends on the two other rules in our system. Under the rules of our grammar, (19) contains compositions for a grammatical sentence, as seen in

(20) [[A unicorn]$_{NP}$[kisses[the cat]$_{NP}$]$_{VP}$]$_{S}$

where the brackets denote the valid compositions, and the constituent labels were abbreviated with their initials. Alternatively, we may visualize a hierarchy of compositions as such



Knowing that there exists such a valid composition within the rules of our grammar, we say that (19) is accepted by our grammar. In other words, grammars define a language (formal language, a collection of strings), which contains all the sentences it accepts. The model deems a sentence grammatical if it is within its language. Syntacticians' goal is to define a grammar whose language is exactly English.

2 Context-free Grammars

We can use context-free grammars (CFGs) to formalize our notion of compositionality, and begin an elementary theory of syntax. The CFG formalism adopts a "top-down" view of our rules. Instead of applying the rules to combine words into fragments and fragments into sentences, we can generate a sentence, starting from the final constituent label of sentence or "S" as our start symbol. This start symbol then branches out to two non-terminals: verb phrase and noun phrase constituents which we abbreviate as "VP" and "NP". Therefore we represent the three-tuple (Sentence, Noun Phrase, Verb Phrase) alternatively as

$$\text{S} \rightarrow \text{NP VP}$$

and we recursively repeat this generation process to non-terminals. Eventually, we apply a rule that generates a word given a category. An example would be

$$\text{DT} \rightarrow \text{the}$$

and "the" is termed a terminal symbol. This rule explicitly states the syntactic category of a given word. More on how and why we adopt top-down generation in §2.2.

2.1 Definition

We adopt the notation used in Michael Collins' course notes. [2] A context-free grammar (CFG) is a 4-tuple $G = (N, \Sigma, R, S)$ where:

1. N is a finite set of non-terminal symbols.

2. Σ is a finite set of terminal symbols.

3. R is a finite set which includes the rules of the form $X \rightarrow Y_1 Y_2$ or $X \rightarrow T$, where $Y_i \in N$ for $i = 1 \ldots n$ and $T \in \Sigma$.

4. $S \in N$ is a distinguished start symbol.

We will give an example of a simple grammar of a transitive verb. Let us define G_{kiss} as:

$$N = \{\text{S, NP, VP, V, DT, N}\}, S = \text{S}, \Sigma = \{\text{the, a, cat, unicorn, kisses}\}$$

with non-terminal rules

$$\begin{aligned} \text{S} &\rightarrow \text{NP VP} \\ \text{NP} &\rightarrow \text{DT N} \\ \text{VP} &\rightarrow \text{V NP} \end{aligned}$$

and terminal rules

$$\begin{aligned} \text{DT} &\rightarrow \text{the} \\ \text{DT} &\rightarrow \text{a} \\ \text{N} &\rightarrow \text{cat} \\ \text{N} &\rightarrow \text{unicorn} \\ \text{V} &\rightarrow \text{kisses} \end{aligned}$$

We will refer to G_{kiss} in the sections that follow.

2.2 Left-most Derivations

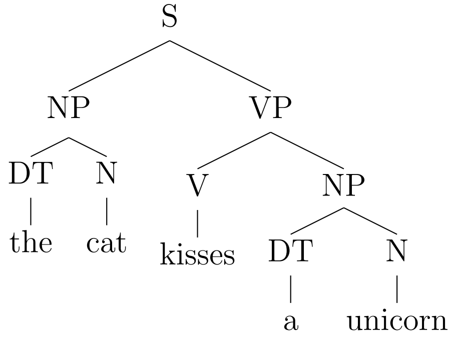
Refer to G_{kiss}. A left most derivation is a sequence of strings $s_1 \dots s_n$, where

1. We begin with $s_1 = \text{S}$, a single start element.
2. The final derivation $s_n \in \Sigma^*$, consisting of terminal symbols only.
3. Each intermediate derivation s_i, where $i = \{2, \dots, n\}$, is derived by picking the left-most non-terminal X in s_{i-1} and replacing it with a β where $X \rightarrow \beta \in R$.

A example derivation in G_{kiss} is given below.

$$\begin{aligned}
s_1 &= \text{S} \\
s_2 &= \text{NP VP} \\
s_3 &= \text{DT N VP} \\
s_4 &= \text{the N VP} \\
s_5 &= \text{the cat VP} \\
s_6 &= \text{the cat V NP} \\
s_7 &= \text{the cat kisses NP} \\
s_8 &= \text{the cat kisses DT N} \\
s_9 &= \text{the cat kisses a N} \\
s_{10} &= \text{the cat kisses a unicorn}
\end{aligned}$$

Derivations may be visually represented as trees. The corresponding trees for the derivation above is



where s_{10}, the final derivation, is the yield of the tree. We use left-most derivations to define the language of a context-free grammar. For a given sentence, if there exists a derivation whose yield is the sentence, then it is within the grammar's language. Alternatively, all sentences from a valid derivation are within the language of the grammar.

For G_{kiss}, it defines all the sentences that are of the form "a/the cat/unicorn kisses a/the cat/unicorn", which means there are 16 acceptable sentences. Certainly not a great model for English, as many actual grammatical sentences are excluded. Interestingly, there are no valid derivations for many ungrammatical sentences of English as well, which means it would correctly reject a lot of ungrammatical sentences. The sentence

(21) * the the the a

(22) * cat unicorn kisses

have no left-most derivations in our grammar and are correctly rejected. It is now up to the reader to develop a better theory of English grammar than I have here.

3 The Chomsky Hierarchy (Optional)

The section explains, in most basic terms, why context-free grammars are generally better models of language. The reader wishing to understand these results will need some prerequisite knowledge to formal language theory. The Chomsky Hierarchy describes the hierarchy of expressiveness of grammar formalisms. At the lowest rung of the hierarchy, regular grammars are the least expressive. Context-free grammars can describe all those languages regular grammars can, and more. There are a few more hierarchies that encompass the context-free languages and more, such as context-sensitive grammars. At the highest rung, Turing machines can describe any language.

Regular grammars, those that can be described by a finite state automaton, cannot describe recursion such as those found in center embedding constructions of relative clauses[1]

(23) A man [that a woman [that a child [...] knows] loves]

where we use brackets to denote nesting structure. Therefore, we believe the grammar of English cannot be regular. In contrast, context-free grammars have the capability to describe these strings. Now if our language were just context-free, syntax would be solved. However, "respectively" constructions in english such as

(24) The square roots of 16, 9 and 4 are 4, 3 and 2, respectively.

exhibit context-sensitivity.[2] You cannot write a context-free grammar that accepts only the set of strings $\{ABC, AABBCC, ...\}$ (convince yourself!). Now should we conclude that language is context-sensitive? These constructions are rare, and context-sensitive grammars describe languages beyond the capacity of human language performance. Therefore, syntacticians believe natural language to be context-free, with some context-sensitive rules.

4 Conclusion

In this tutorial, we reviewed elementary syntactic theory and wrote a context-free grammar to explain a small subset of English sentences. Grammars defined a language, and sentences that were in the language were deemed grammatical in our grammar. In syntax, we are interested in deducing grammars that define a language that is exactly English.

What's next? We did not cover those context-free grammars that contained sentences which had more than one derivation. This is a prominent feature of English, where sentences like "the cat saw the unicorn with a telescope" has an ambiguous sentence structure. I hoped this tutorial kindles your curiosity in syntax to investigate on your own, and helped you appreciate the complexities of natural language. Cheers!

[1] https://en.wikipedia.org/wiki/Center_embedding

[2] martin.kleppmann.com/2008/08/26/context-sensitive-constructions-in-english.html

References

[1] K. Johnson, "University of Massachusetts, Amherst, LINGUIST 401: Introduction to Syntax, Lecture Notes: Phrase Structure Rules," 2016. URL: `http://courses.umass.edu/kbj/ling401/content/401hand04_psrules.pdf`. Last visited on 2018/04/29.

[2] M. Collins, "Columbia Department of Computer Science, COMS W4705: Natural Language Processing, Lecture Notes: Elliptic Curves," 2017. URL: `http://www.cs.columbia.edu/~mcollins/courses/nlp2011/notes/pcfgs.pdf`. Last visited on 2018/04/05.