Formal Languages

Parsing
ISCL-BA-06

Çağrı Çöltekin
ccoltekin@sfs.uni-tuebingen.de

University of Tübingen
Seminar für Sprachwissenschaft

Winter Semester 2020/21

What is a language?

- English, German, Chinese
- Latin, Coptic, Sanskrit, Sumerian
- Proto-Germanic, Proto-Uralic, Proto-Dravidian
- Sign languages
- Esperanto
- Traffic signs, computer icons, emoticons
- Chemical formulas
- Arithmetic expressions
- Python, Java, C++
- XML, JSON, HTML, YAML
- HTTP, TCP, UDP,
- The set of strings $\{ba, baa, baaa, baaaa, \ldots\}$

According to Jurafsky and Martin (2009), the last set of strings form the 'sheep language'.

Natural, artificial, formal languages

- Some languages in our list are natural languages
- In contrast some are designed, they are artificial
- Formal languages are those that we can study formally
  - we can analyze them in principled ways
  - provably answers the questions about them
- All languages in our list can be studied as formal languages (to some extent)

Languages as sets of strings

We define a *formal language* as a set of finite-length string over an *alphabet*.

- The sheep language from the first slide was represented as a set: $\{ba, baa, baaa, baaaa, \ldots\}$

The alphabet of a language is the set of "symbols" in the language, conventionally denoted as Σ.

- For the sheep language, $\Sigma = \{a, b\}$
- What is the alphabet for English syntax?

Formal grammar

A formal *grammar* is a finite specification of a (formal) language.

- Since we consider languages as sets of strings, for a finite language, we can (conceivably) list all strings
- How to define an infinite language?
- Is the definition $\{ba, baa, baaa, baaaa, \ldots\}$ 'formal enough'?
- Using regular expressions, we can define it as `baa*`
- But we will introduce a more general method for defining languages soon
- Are natural languages infinite?

Formal languages

Some definitions

Alphabet	is the set of 'atomic' symbols in the language
String	is a sequence of symbols from the alphabet, For example, 101100 is a string over alphabet $\Sigma = \{0, 1\}$ • String can be concatenated: if $x = 10$ and $y = 11000101$, their concatenation $xy = 1011000101$ • We represent the empty string with ϵ (some books use λ) • The notation x^* indicates zero or more concatenation of string x with itself, e.g., ϵ, 01, 010101 (the operation is called Kleene star) • The notation x^+ is a shorthand for xx^*
Σ^*	is all possible strings that can be defined over alphabet Σ
Sentence	of a language is a string that is in the language (confusingly the term *word* is also common)

Operations on languages

Since we define languages as sets, all set operations are applicable to languages. If L_1 and L_2 are languages,

- Intersection: $L_1 \cap L_2$
- Union: $L_1 \cup L_2$
- Difference: $L_1 - L_2$
- Complement: $\Sigma^* - L_1$
- Concatenation: $L_1L_2 = \{xy | x \in L_1 \text{and} y \in L_2\}$

Three different view on formal languages

- In formal language theory, a language is studied for itself. Languages are simply set of strings, we do not attach 'meaning' to them. The questions of interests are abstract. For example, 'how to find the intersection of two languages for which we have grammars?'
- In computer science, we want to analyze the structure (of, e.g., a computer program) to get some information, or 'meaning'. The most common area is compiler construction, but almost any syntactic analysis task is supported by formal definitions of the respective languages.
- In (computational) linguistics, the aim is to analyze sentences (syntax), and associate them with their meanings (semantics). Formal languages provide a way to study a seemingly chaotic object, natural language, in a principled way.

Grammars: how to describe a language?

- In daily use, a 'grammar' is a book, it defines a language in detail
- But we are interested in more *formal* grammars
- The challenge is describing a possibly infinite set with a finite specification
- We already see that it was possible (e.g., regular expressions)
- Another possible way would be writing a computer program that determines if the given string is in the language
- However, we want more general descriptions: grammars that can describe any 'describable' language in a concise and easy to study formalism

Aside: can any languages be described by a finite description?

Phrase structure grammars

- A phrase structure grammar is a generative device
- If a given string is in the language is generated by the grammar, the string is in the language
- The grammar generates all and the only strings that are valid in the language
- A phrase structure grammar has the following components
  - Σ A set of *terminal* symbols
  - N A set of *non-terminal* symbols
  - $S \in N$ A special non-terminal, called the start symbol
  - R A set of 'rewrite rules' of the form: $\alpha \rightarrow \beta$, which means that the sequence α can be rewritten as β (both α and β are sequences of terminal and non-terminal symbols)

Phrase structure grammars
Some conventions

- We use uppercase letters (sometimes capitalized words) for non-terminal symbols: A, B, C, NP, End
- We use lowercase letters (sometimes lowercase words) for terminals: a, b, c, cat, dog
- We use Greek letters letters for *sentential forms*, (sequences of terminal and non-terminal symbols): α, β, γ
- For sequences of terminal symbols (strings) we use lowercase letters from the end of the alphabet: u, v, w, x, y, z

Generating sentences from a PSG

1. Start with the symbol `S` as the first sentential form
2. Pick a rule with matching the part of the current sentential form
3. Apply the rewrite (production) rule
4. Repeat 2 and 3, until there are no non-terminals left

- Exhaustively exploring all possible productions 'enumerates' all sentences of the language described by the grammar

Phrase structure grammars
A very simple example – the sheep language

A grammar

1. $S \rightarrow B\ A$
2. $B \rightarrow b$
3. $A \rightarrow a\ A$
4. $A \rightarrow a$

Quick exercise: try to define a different grammar for the same language.

An example derivation

Sentential form	rule	notes
`S`		start symbol
`BA`	$S \rightarrow B\ A$	rule 1
`bA`	$B \rightarrow b$	rule 2
`baA`	$A \rightarrow a\ A$	rule 3
`baaA`	$A \rightarrow a\ A$	rule 3
`baaa`	$A \rightarrow a$	rule 4

Generation to parsing

- The above procedure (generating all sentences from a generative grammar) gives us a possible way to do parsing:
  - Enumerate all sentences from the grammar
  - If the string we are interested comes out, it is in the language: parsing is successful
  - If it does not come out, it is not in the language: parsing failed (we'll get back to this point soon)

Phrase structure grammars
Another example: goat language (a dialect of sheep language)[1]

The grammar

1. $S \rightarrow$ Begin B A End
2. $B \rightarrow b$
3. $A \rightarrow a$ End
4. $A \rightarrow a\ A$ End
5. a A End $\rightarrow$ a ' a
6. Begin b a $\rightarrow$ b b a
7. Begin b b $\rightarrow$ b b

A few exercises:

- Derive the string `bbaaa'a`
- Is the string `baa'a` in the language?
- Can you write a simpler grammar for this language?

[1]Some claim the grammar is just the same, but goats use the word `m` instead of the word `b`.

Phrase structure grammars
A few notes

- The phrase structure grammars are not the only way to define languages (sets)
- However, all known methods are either equivalent to, or less powerful than phrase structure grammars
- The formalism we sketched is general: any set (language) that can be generated by a computer program can be defined by a phrase structure grammar

The Chomsky hierarchy of grammars

Type 0	Unrestricted phrase structure grammars
Type 1	Context-sensitive or monotonic grammars
Type 2	Context-free grammars
Type 2.5	Linear grammars
Type 2.1	Mildly-context sensitive grammars
Type 3	Regular grammars
Type 4	Finite (choice) grammars

Type 0: unrestricted PSG

- As indicated - unrestricted, any form of the rewrite rules are allowed
- If a language can be generated at all, it can be defined/generated by a unrestricted PSG
- No general parsing algorithm exists, and in fact cannot exist
- In general, type 0 grammars are not interesting for practical applications
- The class of languages described by type 0 grammars is called *recursively enumerable* languages

Type 1: monotonic

- We introduce one restriction to PSG: the right hand side (RHS) of a rule cannot be shorter than the left hand side (LHS)
- The rule applications cannot 'shrink' the sentential forms
- For example, our 'goat language grammar' is not monotonic, because of the rule Begin b b $\rightarrow$ b b
- This also means no ϵ-rules
- Sometimes the language with only the empty string is allowed as an exception

Type 1: context sensitive

- A context-sensitive grammar rewrites only one of its non-terminal on the LHS.
- Our 'goat language grammar' is not context-sensitive, because of the rule
- a A End $\rightarrow$ a ' a
- Context-sensitive and monotonic grammars are equivalent
- Parsing is possible with Type 1 grammars, but inefficient
- In general, not much practical use

Type 2: context free

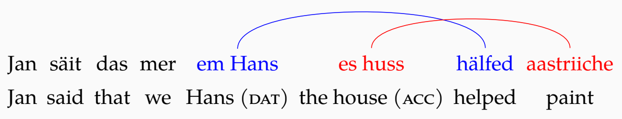
- A context free language requires its LHS to have only a single non-terminal symbol
- This means the rewrite rules cannot be conditioned on context, they are independent of their environment
- Our sheep language example is context free
- Context-free languages have efficient parsers, and used in practical applications
- All programming languages are (subclasses) of context free languages
- Most of natural language parsing is based on context-free parsing (more on this soon)

Type 3: regular

- Regular grammars come in two flavors: *right-regular* and *left-regular*
- A right-regular grammar allows only two types of rules: A $\rightarrow$ a and A $\rightarrow$ a B
- A left-regular grammar allows: A $\rightarrow$ a and A $\rightarrow$ B a
- Generally, ϵ-rules are also allowed A $\rightarrow$ ϵ
- Right-regular grammars are more common in practical use
- Almost all operations on regular languages are efficient, lots of practical use
- Regular grammars are equivalent to regular expressions

Where do natural language syntax fit?

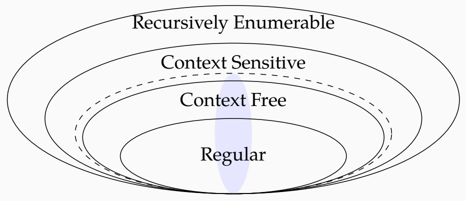
Cross-serial dependencies



Jan	säit	das	mer	em Hans	es huss	hälfed	aastriiche
Jan	said	that	we	Hans (DAT)	the house (ACC)	helped	paint

- The above structure is not possible to parse using context-free languages
- Otherwise, experience so far indicates that a CF-based grammar can describe natural language syntax

Chomsky hierarchy: the picture



- Chomsky hierarchy of languages form a hierarchy (with some care about empty language)
- It is often claimed that mildly context sensitive grammars (dashed ellipse) are adequate for representing natural languages
- Note, however, not even every regular language is a potential natural language (e.g., a^*bbc^*). The possible natural languages probably cross-cut this hierarchy (shaded region)

Summary

- Phrase structure grammars are generative grammars that are finite specifications of (infinite) languages
- They form the basis of theory of parsing
- More expressive grammar classes (type 0 and type 1) are computationally not attractive
- We will look into more practical grammar classes, context-free and regular languages, more closely (next lecture)