Syntax and grammars

NTNU

TDT4165 fall 2018

Formal grammars

A formal grammar, G, can be defined as a set G = N,T,R,S where N, T, R are finite sets and $S \in N$.

Terminal symbols T is a set that contains our alphabet. These are any characters that are legal in our language.
Example: T = {a,b,c,d,4,$}

Non-terminal symbols N contains of the non-terminal symbols of a language. They are also called syntactical variables. Non-terminal symbols are symbols that can be replaced with other symbols. It is common to denote non-terminal symbols using uppercase alphabetic characters or to surround them with angular brackets $\langle A \rangle$.
Example: $N = \{\langle Z \rangle, \langle Y \rangle, \langle S \rangle\}$

Starting point $\langle S \rangle$ is our starting point and is in our set of non-terminal symbols.

Production rules The set R contains our production rules. Production rules define how we can perform symbol substitution in G.
Example: $R = \{\langle Z \rangle \rightarrow a \mid a\langle Y \rangle b, \langle Y \rangle \rightarrow c \mid \epsilon, \langle S \rangle \rightarrow \langle Z \rangle \mid \langle Y \rangle\}$ To determine legal strings in our language, we recursively replace expressions on the left hand side, with expressions on the right hand side, starting at $\langle S \rangle$.

```
<S> ::= <Z> | <Y>
<Z> ::= a | a <Y> b
<Y> ::= c | ϵ
```

To get the string **ab**, we start at $\langle S \rangle$ which can be $\langle Z \rangle$ or $\langle Y \rangle$. Choosing $\langle Z \rangle$, we have the choices **a** and **a** $\langle$**Y**$\rangle$ **b**. For $\langle Y \rangle$, we also have two choices: **c** or the empty string. We have:

```
<S> → <Z> → a <Y> b → aϵb → ab
```

This is an example of a finite language. We can also create infinite languages, like the one below.

```
<S> ::= <Z>
<Z> ::= a <Z> b | c
```

This language contains all strings $a_1...a_ncb_1..b_n, n >= 0$:

```
n = 0: c
n = 1: acb
n = 2: aacbb
...
```

Chomsky hierarchy

The Chomsky hierarchy is a way to categorize formal grammars, by their properties. Each higher level category, inherits the properties of the lower categories.

Unrestricted grammar Production rules have the form $\alpha \to \beta$, where α and β are any combination of terminal and non-terminal symbols. The left hand side (α) of a production rule cannot be empty (ϵ) and contains at least one non-terminal.

Context-sensitive grammar Production rules have the form $\alpha\langle \mathrm{A}\rangle\beta \to \alpha\gamma\beta$, where α, β are any combination of terminal and non-terminal symbols or ϵ, γ is any combination of terminal and non-terminal symbols, but cannot be ϵ and A is a non-terminal.

Context-free grammar Every production rule is on the form $\langle \mathrm{A}\rangle \to \gamma$, where A is a non-terminal and γ is any combination of terminal and non-terminal symbols, but cannot be ϵ.

Regular languages Production rules have the form A → aB, or A → a, where A and B are non-terminals and a is a terminal. This example is a right regular grammar. Grammars with production rules on the form A → Ba, are left regular grammars.