PROGRAMMING LANGUAGES AND COMPUTATION
Week 2: Context Free Grammars

We start with some revision of sets. If you are unfamiliar with any of the notation, please ask one of the TAs.

* 1. Write $\mathbb{N}$ for the set of natural numbers $0, 1, 2\ldots$, and write Σ for the alphabet $\{0, 1\}$. List the elements of the following sets in any order.

   (a) $\{1, 2, 3\} \cup \{2, 4, 6\}$

   (b) $\{1, 2, 3\} \cap \{2, 4, 6\}$

   (c) $\{1, 2, 3\} \times \{2, 4, 6\}$

   (d) $\{1, 2, 3\} \setminus \{2, 4, 6\}$

   (e) $\{2m \mid m \in \mathbb{N}, 0 \leq m \leq 5\}$

   (f) $\{uu \mid u \in \Sigma^*, |u| = 2\}$

   (g) $\{u0v \mid u \in \Sigma^*, v \in \Sigma^*, |uv| = 2\}$

   (h) $\{uvw \mid u \in \Sigma, v \in \Sigma, w \in \Sigma, w \text{ is the xor of } u \text{ and } v\}$

Recall the grammar from the notes for Boolean expressions.

* 2. Give derivations for the following strings which are in the language of Boolean expressions.

   (a) true && false

   (b) true || false && true

   (c) true && (false || true)

* 3. Give three distinct derivations for the string true || false.

* 4. Consider the following CFG G with start symbol R:

$$\begin{aligned} R &\longrightarrow XRX \mid S \\ S &\longrightarrow aTb \mid bTa \\ T &\longrightarrow XTX \mid X \mid \epsilon \\ X &\longrightarrow a \mid b \end{aligned}$$

(a) What are the non-terminals of G?

(b) What are the terminals of G?

(c) Give three strings in $L(G)$.

(d) Give three strings not in $L(G)$.

(e) True or false: $T \to aba$.

(f) True or false: $T \to^* aba$.

(g) True or false: $T \to T$.

(h) True or false: $T \to^* T$.

(i) True or false: $XXX \to^* aba$.

(j) True or false: $X \to^* aba$.

(k) True or false: $T \to^* XX$.

(l) True or false: $T \to^* XXX$.

(m) True or false: $S \to^* \epsilon$.

* 5. Recall the syntax of the While programming language from the reference notes. Which of the following are valid While programs (i.e. strings in the language of that grammar)? You do not have to give the derivations (but you should work through them in your head).

(a) $foo \leftarrow x + 2$

(b) $foo \leftarrow x + 2$; skip

(c) while x do skip

(d) if $x \leq 3 \mathbin{||} ! x = y$ then $x \leftarrow x + y$ else skip

(e) $x \leftarrow 2$; $y \leftarrow 0$;

(f) 6

(g) $\{bar \leftarrow foo\}$; $foo \leftarrow bar$

(h) while true do $\{y \leftarrow y + 1$; $z \leftarrow z + 1\}$; $x \leftarrow 1$

** 6. Design CFGs for the following programming language lexemes over the ASCII alphabet. You will find it convenient to use abbreviations like $\cdots$ to help present the expressions compactly.

2

(a) A *C program identifier* is any string of length at least 1 containing only letters ('a'–'z', lower and uppercase), digits ('0'–'9') and the underscore, and which begins with a letter or the underscore.

(b) In most programming languages, an *integer literals* is any string taking one of the following forms:

- a non-empty sequence of digits (decimal)
- a non-empty sequence of characters from '0'–'9','a'–'e' (upper or lowercase) that are preceded by "0x" (hexadecimal)
- a non-empty sequence of bits '0' and '1' that are preceded by "0b" (binary)

** 7. Give a CFG that describes the language of all subsequences of all permutations (i.e. no repetition) of the following keywords:

final static synchronized

** 8. Give CFGs for the following languages. The later parts are more difficult than 2-star.

(a) All odd length strings over $\{a, b\}$.

(b) All strings over $\{a, b\}$ that contain aab as a substring.

(c) The set of strings over $\{a, b\}$ with more a than b. Hint: every string w with at least as many a as b (possibly the same number of a as b) can be characterised inductively as follows. Either:

- w is just a
- or, w is of shape avb with v containing at least as many a as b
- or, w is of shape bva with v containing at least as many a as b
- or, w is of shape $v_1 v_2$ with v_1 and v_2 each separately containing at least as many a as b

(d) The complement of the language $\{a^n b^n \mid n \geq 0\}$ over $\{a, b\}$. Hint: express "not of shape $a^n b^n$ for some n" into one or more positive (i.e. without using *not* or similar) conditions.

(e) $\{v \# w \mid$ the reverse of v is a substring of $w\}$, over $\{a, b, \#\}$.

*** 9. An ϵ-production is a rule of shape $X \longrightarrow \epsilon$ (for some nonterminal X). A unit production is a rule of shape $X \longrightarrow Y$ (for some non-terminals X and Y). Consider the following grammar G with start symbol S:

$$
\begin{aligned}
S &\longrightarrow aSbb \mid T \\
T &\longrightarrow bTaa \mid S \mid \epsilon
\end{aligned}
$$

This grammar has two unit productions $S \longrightarrow T$ and $T \longrightarrow S$, and it has an epsilon production $T \longrightarrow \epsilon$.

Give a grammar with *no* unit productions and *no* ϵ-productions for the language $L(G) \setminus \{\epsilon\}$.

*** 10. Give an English description of the language of the grammar in Q4.