Ex. 1

Consider the following sentences.

(1) a. Alice eats cakes.
b. The caterpillar gives Alice cakes.
c. The cat with a grin disappears.
d. Alice paints white roses red.

Define a context-free grammar that could generate these sentences.

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . Answer . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

The minimal alphabet will contain all the "words" that occur in these sentences:
{Alice, eats, cakes, the, caterpillar, gives, cat, with, a, disappears, paints, white, roses, red}
The most obvious answer would be a grammar that produces exactly these four sentences:
$S \longrightarrow$ Alice eats cakes
$S \longrightarrow$ The caterpillar gives Alice cakes
$S \longrightarrow$ The cat with a grin disappears
$S \longrightarrow$ Alice paints white roses red
However such a grammar would miss the point of "grammar building" for natural language: we need a grammar much more general, so that for instance the sentence (2a) is also part of the grammar but we also need to ensure that the grammar we create is not "overgenerating" too much, avoiding for instance to produce (2b). The trade-off between those two requirements is exactly what syntacticians are working on (although with grammars and syntactic phenomena much more complex that our toy examples).

(2) a. The caterpillar eats cakes.
b. *Alice disappears cakes.

**(1a)** The first sentence can be produced by the following grammar, of which the first two rules are quite common, while the two others introduce specific lexical categories, one to account for bare plurals $N_p$, and less questionably one for transitive verbs $V_t$.

| | |
|---|---|
| $S \longrightarrow NP\ VP$ | |
| $NP \longrightarrow PN$ | $PN \longrightarrow$ Alice |
| $NP \longrightarrow N_p$ | $N_p \longrightarrow$ cakes |
| $VP \longrightarrow V_t\ NP$ | $V_t \longrightarrow$ eats |

This grammar produces, in addition to the target sentence, (3a) but also (3b).

(3) a. Alice eats Alice
b. Cakes eats Alice

**(1b)** To account for the second sentence, we need to add a quite classical analysis of *NP*s, and we choose to call $V_o$ verbs that allow for a double accusative construction.

| | | | | | |
|---|---|---|---|---|---|
| $NP$ | $\longrightarrow$ | $Det\ N$ | $Det$ | $\longrightarrow$ | the |
| $VP$ | $\longrightarrow$ | $V_o\ NP\ NP$ | $N$ | $\longrightarrow$ | caterpillar |
| | | | $V_o$ | $\longrightarrow$ | gives |

This grammar produces, in addition to the target sentence, (4a) (which is not so bad) and (4b), but also (4c) which is arguably syntactically well formed.

(4) a. Alice gives the caterpillar cakes
b. Alice gives Alice Alice
c. Alice gives cakes the caterpillar

**(1c)** The third sentence requires a treatment for prepositional phrases as noun modifiers. With basically the same lexicon, we could come up with several different options. Option [A] correspond to an adjunction of the PP at the level of the NP (may work here, but not very general); option [B] corresponds to a much more restricted view on PP modifiers (only one possible in an NP). The option [C] is probably the more general, claiming that PP modification occurs recursively at the intermediate level ($N'$). The rest of the grammar has to be adapted accordingly.

| | | | | | | |
|---|---|---|---|---|---|---|
| | $PP$ | $\longrightarrow$ | $P\ NP$ | $P$ | $\longrightarrow$ | with |
| A | $NP$ | $\longrightarrow$ | $NP\ PP$ | $N$ | $\longrightarrow$ | cat |
| B | $NP$ | $\longrightarrow$ | $Det\ N\ PP$ | $Det$ | $\longrightarrow$ | a |
| C | $NP$ | $\longrightarrow$ | $Det\ N'$ | $N$ | $\longrightarrow$ | grin |
| | $N'$ | $\longrightarrow$ | $N'\ PP$ | | | |
| | $N'$ | $\longrightarrow$ | $N$ | | | |

**(1d)** To account for the last sentence we have to introduce a way to deal with so-called resultative constructions. Here we assume that some verbs ($V_a$) allow for a resultative construction where an adjective is adjoined to a "direct" object. We also need to account for adjectival modification inside *NP*s. Assuming we chose option [C] earlier, we propose that adjectival modification is recursive at the $N'$ level.

| | | | | | |
|---|---|---|---|---|---|
| $VP$ | $\longrightarrow$ | $V_a\ NP\ Adj$ | $V_a$ | $\longrightarrow$ | paints |
| $N'$ | $\longrightarrow$ | $Adj\ N'$ | $Adj$ | $\longrightarrow$ | red |

Note: the proposed grammar(s) do not add a period at the end of a sentence. Since we can assume that every sentence ends with a period, and that there is no interference with the rest of the grammar, producing a period would simply require we take a new axiom $S'$ and add to the grammar the single rule $S' \longrightarrow S.$ .

Ex. 2

1. Let $G$ be the grammar $S \rightarrow aSbb \mid \varepsilon$. Describe informally the language generated by $G$.
2. Let $G'$ be the grammar $S' \rightarrow SSS$, $S \rightarrow aSbb \mid \varepsilon$, with $S'$ as the start symbol (axiom). Describe informally the language generated by $G'$.
3. Let $G_1$ and $G_2$ be context-free grammars; $L(G_1)$ and $L(G_2)$ the languages they generate. Show that there is a context-free grammar generating each of the following sets:
   (a) $L(G_1) \cup L(G_2)$
   (b) $L(G_1)L(G_2)$
   (c) $L(G_1)^*$

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . Answer . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

1. All the words of $L(G)$ are formed by a sequence of $a$'s followed by a sequence of twice as many $b$'s. The number of $a$'s is unconstrained ($\geqslant 0$).
   More formally, $L(G) = \{a^k b^{2k} \ / \ k \in \mathbb{N}\}$.

2. $L(G')$ is equivalent to $L(G)^3$, the set of words that can be decomposed into a sequence of 3 words from $L(G)$. All the words of $L(G')$ are formed by $k$ $a$'s followed by $2k$ $b$'s, then $k'$ $a$'s followed by $2k'$ $b$'s, then followed by $k''$ $a$'s followed by $2k''$ $b$, with $k, k', k'' \in \mathbb{N}$.

3. We can get some inspiration from question 2 which illustrates a method to build a grammar for a language $L^3$ given a grammar for $L$.

   Let's assume that $G_1 = \langle \Sigma, N_1, S_1, P_1 \rangle$ and $G_2 = \langle \Sigma, N_2, S_2, P_2 \rangle$, with $N_1 \cap N_2 = \emptyset$ (possibly after having renamed symbols, without loss of generality).
   The same general procedure will apply for the three cases:
   Let $S$ be a new non terminal symbol ($S \notin N_1 \cup N_2$). The new grammar can be thus defined: $G = \langle \Sigma, \{S\} \cup N_1 \cup N_2, S, P_1 \cup P_2 \cup P \rangle$, where $P$ is the set of additional production rules (see below). Since in each case these additional rules are context-free (while those of $P_1$ and $P_2$ are context-free by hypothesis), the grammars we provide are context-free:

   | | | |
   |---|---|---|
   | (a) | $L(G_1) \cup L(G_2)$ | $P = \{S \longrightarrow S_1 \ ; \ S \longrightarrow S_2\}$ |
   | (b) | $L(G_1)L(G_2)$ | $P = \{S \longrightarrow S_1 S_2\}$ |
   | (c) | $L(G_1)^*$ | $P = \{S \longrightarrow S_1 S \ ; \ S \longrightarrow \varepsilon\}$ |

   What remains to be done is to prove in each case that the new grammar engenders exactly the target language. A formal proof was not required, but it could be sketched along the following lines (case (b)):

   Let $L(G_3)$ be the language engendered by the new grammar.

   - $L(G_1)L(G_2) \subset L(G_3)$: every word $w \in L(G_1)L(G_2)$ can be decomposed into $uv$ with $u \in L(G_1)$ and $v \in L(G_2)$. Since $u \in L(G_1)$, there is a derivation from $S_1$ to $u$ in $G_1$: $S_1 \xrightarrow{*} u$. Similarly $S_2 \xrightarrow{*} v$. So the derivation $S \longrightarrow S_1 S_2 \xrightarrow{*} uv = w$ exists in $G_3$, which means that $w \in L(G_3)$. ■
   - $L(G_3) \subset L(G_1)L(G_2)$: any word produced by $G_3$ was necessarily produced through the derivation $S \longrightarrow S_1 S_2$ as this rule is by construction the only rule having the axiom as a left-handside member. Therefore, given the projectivity of context-free grammars, any word $w$ engendered by $G_3$ will be decomposed into $uv$, where $u$ is engendered by $S_1$ and $v$ is engendered by $S_2$. By definition $u \in L(G_1)$ and $v \in L(G_2)$. Therefore $w = uv \in L(G_1)L(G_2)$. ■