Exercise 1.

Consider the sentence *She announced a program to promote safety in trucks and vans* and the following syntactic tree of one of its possible interpretations, in which the program promotes safety in trucks, and also promotes vans:

1. Draw the trees for at least three other interpretations for this sentence
2. Draw the trees for at least two interpretations for each of the following sentences
   - *The post office will hold out discounts and service concessions as incentives*
   - *They are hunting lions and tigers*
   - *Monty flies like mosquitoes*
**Exercise 2.**

Say we have the phrase *saw the dog with the telescope* and we are given the gold parse tree (left) and the predicted parse tree (right):

What are the precision and recall of this predicted parse tree?

**Exercise 3.**

Consider the following CFG:

\[
\begin{align*}
S & \rightarrow NP \ VP \\
NP & \rightarrow DT \ NN \\
NP & \rightarrow NP \ PP \\
PP & \rightarrow IN \ NP \\
VP & \rightarrow VB \ NP
\end{align*}
\]

1. How many parse trees are there under this grammar for the sentence *the man saw the dog in the park*?
2. How many parse trees are there under this grammar for the sentence *the man saw the dog in the park with the cat*?
3. Consider a sentence that is grammatical under the above context-free grammar, and has exactly \(k\) prepositions following the verb, and 0 prepositions before the verb (a preposition is any word with tag \(IN\)). How many parse trees will this sentence have? Reason why.

The \(n^{th}\) Catalan number is defined as

\[
C_n = \frac{(2n)!}{n!(n+1)!}
\]

It can be shown that \(C_n\) is the number possible different binary trees with \(n + 1\) leaves.

**Exercise 4.**

Consider the following CFG:

\[
\begin{align*}
S & \rightarrow NP \ VP \\
NP & \rightarrow DT \ NN \\
NP & \rightarrow NP \ PP \\
PP & \rightarrow IN \ NP \\
VP & \rightarrow VB \ NP
\end{align*}
\]

This grammar overgenerates incorrect English sentences, such as:

*the dog see the cat*
the dog in the park see the cat
the dog in the park see the cat in the park
the dogs sees the cat
the dogs in the park sees the cat
the dogs in the park sees the cat in the park

1. Modify the grammar so that all generated sentences respect third-person subject-verb agreement rules for English

Exercise 5.

Consider the following CFG:

```
S → NP VP
NP → DT NN NN
PP → IN NP
VP → VB NP
```

An infinite number of sentences can be generated by this grammar, for example:

the man saw the dog
the man saw the dog with the telescope
the man saw the dog with the telescope under the dog
the man saw the dog under the telescope with the dog under the telescope
etc.

The language $L(G)$ generated by a context-free grammar $G$ is defined as the set of sentences that can be derived with a sequence of grammar rule applications.

A hidden Markov model (HMM), defines a distribution $P(x_1 \ldots x_n, y_1 \ldots y_n)$ over sentences $x_1 \ldots x_n$ paired with PoS tag sequences $y_1 \ldots y_n$. The language generated by a HMM is defined as the set of sentences $x_1 \ldots x_n$ such that:

$$\max_{y_1 \ldots y_n} P(x_1 \ldots x_n, y_1 \ldots y_n) > 0,$$

that is, sentences with at least one possible PoS-tag sequence $y_1 \ldots y_n$ that gives a non-zero value for the probability $P(x_1 \ldots x_n, y_1 \ldots y_n)$.

1. Write a bigram HMM that generates the same language than the context-free grammar given above.
Probabilistic Context Free Grammars

Exercise 6.

Using the following PCFG in CNF:

\[
\begin{align*}
S & \rightarrow NP \ VP & 1.0 \\
NP & \rightarrow NP \ PP & 0.4 \\
PP & \rightarrow P \ NP & 1.0 \\
VP & \rightarrow V \ NP & 0.7 \\
VP & \rightarrow VP \ PP & 0.3 \\
P & \rightarrow with & 1.0 \\
P & \rightarrow saw & 1.0 \\
NP & \rightarrow astronomers & 0.1 \\
NP & \rightarrow ears & 0.18 \\
NP & \rightarrow saw & 0.04 \\
NP & \rightarrow stars & 0.18 \\
NP & \rightarrow telescopes & 0.1
\end{align*}
\]

Work with the sentence: astronomers saw stars with ears

• How many correct parses are there for this sentence?
• Write them, along with their probabilities.

Exercise 7.

Given the following PCFG:

\[
\begin{align*}
S & \rightarrow NP \ VP & 1.0 \\
NP & \rightarrow N \ N & 0.25 \\
NP & \rightarrow D \ N & 0.4 \\
NP & \rightarrow N & 0.35 \\
VP & \rightarrow V \ NP & 0.6 \\
VP & \rightarrow V \ ADVP & 0.4 \\
ADVP & \rightarrow ADV \ NP & 0.4 \\
ADV & \rightarrow like & 1.0 \\
V & \rightarrow flies & 0.5 \\
V & \rightarrow like & 0.5 \\
D & \rightarrow an & 1.0
\end{align*}
\]

and the sentence time flies like an arrow

1. Write two parse trees that this grammar generates for this sentence
2. Compute the probability of each tree.
3. Convert the grammar to CNF and emulate the behaviour of the CKY algorithm on this sentence.
   Provide the final chart with all the information involved.

Exercise 8.

Consider that you have as a training corpus a treebank containing the following trees. Each tree was observed the number of times indicated below it.

\[
\begin{align*}
S & & & & & & & \\
A & & & A & & & & \\
| & & | & & | & & | \\
a & & a & & a & & f & \\
75 & & 10 & & 325 & & 8 & & 428
\end{align*}
\]

1. What PCFG would one get from this treebank (using MLE)?
2. Given the obtained grammar:
   • What is the most likely parse of the string a a?
   • Is this a reasonable result? Discuss why.
Exercise 9.

Consider the two following parse trees:

Discuss whether the following statements are true or false and why:

1. The two parse trees receive the same probability under any PCFG
2. The first parse tree receives higher probability if $P(N' \rightarrow NN N') > P(N' \rightarrow N' N')$
3. The first parse tree receives higher probability if $P(N' \rightarrow NN N') > P(N' \rightarrow N' N') + P(N' \rightarrow NN)$

Exercise 10.

Consider the following PCFG:

1. List all sentences generated by this grammar, along with their probability
2. Define a bigram language model that gives the same probability distribution $p(x)$ than the PCFG shown above. The vocabulary of the language model should be $\Sigma = \{a, the, president, support, hate\}$. Specify the value for each parameter of the language model.

A bigram language model consists of a finite vocabulary $\Sigma$, and a parameter $q(u, v)$ for each bigram $(u, v)$ such that $u \in \Sigma \cup \{\text{START}\}$ and $v \in \Sigma \cup \{\text{STOP}\}$. The value for $q(u, v)$ can be interpreted as the probability of seeing word $v$ immediately after word $u$, i.e. $P(v|u)$.

For any sentence $x_1, \ldots, x_n$ where $x_i \in \Sigma$, the probability of the sentence under the bigram language model is $p(x_1, \ldots, x_n) = \prod_{i=1}^{n+1} q(x_{i-1}, x_i)$, where we define $x_0 = \text{START}$ and $x_{n+1} = \text{STOP}$.

Exercise 11.

Consider the following PCFG

<table>
<thead>
<tr>
<th>Production</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S \rightarrow NP \ VP$</td>
<td>1.0</td>
</tr>
<tr>
<td>$NP \rightarrow DT\ NBAR$</td>
<td>1.0</td>
</tr>
<tr>
<td>$NBAR \rightarrow NN$</td>
<td>0.7</td>
</tr>
<tr>
<td>$NBAR \rightarrow NBAR\ NBAR$</td>
<td>0.3</td>
</tr>
<tr>
<td>$VP \rightarrow sleeps$</td>
<td>1.0</td>
</tr>
<tr>
<td>$DT \rightarrow the$</td>
<td>1.0</td>
</tr>
<tr>
<td>$NN \rightarrow mechanic$</td>
<td>0.1</td>
</tr>
<tr>
<td>$NN \rightarrow car$</td>
<td>0.2</td>
</tr>
<tr>
<td>$NN \rightarrow metal$</td>
<td>0.7</td>
</tr>
</tbody>
</table>
1. What is the parse tree with highest probability for the sentence *the metal car mechanic sleeps*?

2. Modify the grammar above so that the sentence *the human language technology rules* has two interpretations (one about *human language* and another about *human technology*). Draw the trees for both interpretations, and point out which is the most likely.

**Exercise 12.**

This exercise considers several forms of language models that compute the probability of sentences $P(x)$. In each of the following cases you need to write an expression that indicates how the particular language model computes the probability $P(x)$, making clear what parameters of the model are used to compute the probability for the example sentence.

1. **$n$-gram language models.** The model considers only the words of $x$, the rest of the linguistic structure is ignored. Write the expression for $n = 2$ and $n = 3$.

2. **Hidden Markov Models (HMM).** The model represents pos tags in the state sequence and words in the observation sequence. The syntactic tree is ignored. Write the expression of $P(x)$ for a bigram HMM, where states correspond to single PoS tags, and for a trigram HMM, where states correspond to two adjacent pos tags.

3. **Probabilistic Context-Free Grammars (PCFG).** The model considers the full syntactic tree.