Student's affirmation: I certify that I have neither taken help in completing this assignment nor helped anyone else with this assignment. I have never discussed this assignment with anyone other than the instructor and TA. I have not used ChatGPT, Llama, or other AI tools to create or influence my solutions.

(Signature) _____

Mandatory: Affirmation and signature on the first page; name on every page; submission as PDF.
If you make assumptions about any problem, state them, but be prepared to justify why they were necessary.

Problem	1	2	3	Total
Points:	30	30	40	100
Score:				

This assignment has 3 problems, for a total of 100 points.

Throughout, I prefer that you think afresh but if you come across a source on which you base your answer, please be sure to cite it. Please note that if you are citing a source to bolster your claim then the source should be authoritative (such as our textbook or a book by another credentialed author). Blog posts by strangers are not credible.

1. (30 points) Mark the following statements `true` or `false`. Provide a short explanation of about 10–20 words. You can and should provide a source where appropriate, including the specific page and line numbers.
   A. Since minimum edit distance is conceived for written language, it cannot be adapted to capture variations in spoken language
   B. Macroaveraging is a superior measure of performance (relative to microaveraging) for datasets in which some classes are substantially more numerous that others
   C. Laplace ("add one") smoothing is an effective method for dealing with words that have zero counts in a given corpus
   D. The TF-IDF model captures the intuition that words that occur rarely in a corpus are more informative of the documents in which they occur
   E. A (dense) distributional representation may be harder to produce than TF-IDF vectors but can be a lot easier to use
   F. Constituents are naturally characterized as spans of text that can replaced by a single word within the span, e.g., "baby oil" by "oil"
   G. Constituents are naturally characterized as spans of text that can naturally be moved as a group
   H. A major benefit of dependency grammars over constituency grammars is that they can produce the same tree for a sentence even when its word order is changed slightly
   I. Sentences that people report as easy to parse may be ambiguous but perhaps people don't notice the ambiguity
   J. Although people can reread a written garden path sentence there is no solution to dealing with a confusing sentence when spoken by someone in conversation
   K. People can't talk about the future or past if their language doesn't include tense markers
   L. A span of text that can be identified as a named entity cannot nest another named entity within it
   M. Since named entities rarely arise consecutively in English, a potentially useful feature in judging whether something is a named entity is whether it is adjacent to a named entity
   N. The relevant relations between entities capture established stable facts, not transient phenomena
   O. Sentiment analysis involves figuring out both what the sentiment is about and what the sentiment is

2. Consider the sentence "Sally saw a man loitering through the planetary observatory with a terrestrial telescope."

   Use Jurafsky's $\mathscr{L}_1$ miniature grammar (Figure 18.8 in the 2024-08-20 draft, reproduced below) as your starting point. I suggest reading the chapter. Assume we have expanded the lexicon to include the words in this sentence. (No need to show the lexicon.)

S $\longrightarrow$ NP VP	NP $\longrightarrow$ Determiner Nominal	VP $\longrightarrow$ Verb NP
S $\longrightarrow$ Aux NP VP	Nominal $\longrightarrow$ Noun	VP $\longrightarrow$ Verb NP PP
S $\longrightarrow$ VP	Nominal $\longrightarrow$ Nominal Noun	VP $\longrightarrow$ Verb PP
NP $\longrightarrow$ Pronoun	Nominal $\longrightarrow$ Nominal PP	VP $\longrightarrow$ VP PP
NP $\longrightarrow$ Proper-Noun	VP $\longrightarrow$ Verb	PP $\longrightarrow$ Preposition NP

   (a) (12 points) Insert one or more productions that would generate the above sentence. (It is OK to delete some productions if that helps simplify your answer.) Explain all the insertions and deletions you make.

   (b) (10 points) Show one possible parse tree of this sentence that corresponds to the interpretation that Sally was the one loitering.

   (c) (8 points) Show a lexicalized version of the same tree. Follow conventions compatible with the Universal Dependencies project, where applicable.

   Explain your answers in under 100 words.

3. Remember to insert brief (∼30–70 words each) explanations for each part.

   Consider the Amazonian Yodas, a language community that speaks OhEssVese. OhEssVese is a OSV language in which the default word order is object–subject–verb. The Wikipedia article on OSV languages https://en.wikipedia.org/wiki/Object-subject-verb_word_order provides interesting background, as do the related articles on other word orders. OSV languages are rare and include some Amazonian languages as well as Yoda's speech—hence the nickname of this language community. As the above article explains, OSV representations are used in non-OSV languages such as English to convey special meanings such as emphasizing a contrast.

   OhEssVese is surprisingly similar in spirit to the $\mathscr{L}_1$ grammar from the textbook (Figure 18.8 in the 2024-08-20 draft, reproduced below). However, OhEssVese has some quirks relative to English:

   - It places articles after nouns
   - It places modifiers for objects before the head noun and modifiers for subjects after the head noun
   - It places modifiers on verbs after the main verb

S $\longrightarrow$ NP VP	NP $\longrightarrow$ Determiner Nominal	VP $\longrightarrow$ Verb NP
S $\longrightarrow$ Aux NP VP	Nominal $\longrightarrow$ Noun	VP $\longrightarrow$ Verb NP PP
S $\longrightarrow$ VP	Nominal $\longrightarrow$ Nominal Noun	VP $\longrightarrow$ Verb PP
NP $\longrightarrow$ Pronoun	Nominal $\longrightarrow$ Nominal PP	VP $\longrightarrow$ VP PP
NP $\longrightarrow$ Proper-Noun	VP $\longrightarrow$ Verb	PP $\longrightarrow$ Preposition NP

   For example, the English | The cheeky monkey hungrily ate a ripe banana. |

   would map to | Ripe banana a monkey cheeky the ate hungrily. |

   (a) (4 points) Write the equivalent of the following English sentence in OhEssVese.
   | The village chieftain greedily retains the plantain tree. |

   (b) (16 points) Provide a constituency grammar for OhEssVese based on $\mathscr{L}_1$ and suitably extending it.

   (c) (6 points) Convert the above OhEssVese grammar to Chomsky Normal Form.

   (d) (14 points) Work out the steps of the CYK parser as shown in class for a parse of
   | Ripe banana a monkey cheeky the ate hungrily. |