Student's affirmation: I certify that I have neither taken help in completing this exam nor helped anyone else with this exam. I have never discussed this exam with anyone other than the instructor and TA.

(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	32	38	100
Score:				

This exam 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. A contraction such as "prolly" is an example of a clitic

   B. The vector-space model yields dense vectors that are easy to incorporate as features in machine learning

   C. Accommodation (in the sense of meaning) is the idea that you adjust your interpretation of what someone is saying to make sense, and respond accordingly (possibly echoing their vocabulary)

   D. Skipgram training relies on a way to generate negative samples from pairs of words that don't occur together

   E. Stemming conceptually applies only to known words and identifies their stems

   F. Ideally, evaluation heuristics for a computational model should account for the costs of different errors

   G. Splitting off a test dataset for an NLP approach based on machine learning ensures that there won't be any risk of overfitting

   H. Backoff is the idea that you can simply skip a word of which there is not enough occurrences

   I. Newly introduced (to English) words such as Latinx are an example of a closed class of words changing for English

   J. Prescriptive grammar is an approach to language based on rules specified by an authoritative source on what is correct use of language

   K. The ri–di–cu–lous example from XKCD suggests that there is a constituency structure within words

   L. We can be certain that two native speakers of the same dialect of Swahili will agree with each other on the meanings of sentences in their dialect of Swahili

   M. We can be certain that if one native speaker of Hausa says something, another native speaker of the same dialect of Hausa will agree with them

   N. If a sentence has a correct parse tree and we substitute one word in the sentence with another word of the same part of speech, the same parse tree with one leaf changed will still be correct

   O. Whereas positive coordinating conjunctions (such as *and*) convey that the coordinated phrases are of the same type, negative coordinating conjunctions (such as *but*) convey that the coordinated phrases may or may not be of the same type

2. Consider a challenge for a distributional representation such as the word embeddings we discussed in class.

   Suppose our embeddings are trained for individual words and the baseline approach to compute the embedding for a phrase is to take the average of the embeddings of the words in that phrase.

   Consider the following requirements (the example is inspired by Bender and Wasow):

   - In the sense of attire, $\overrightarrow{\text{dress}} \approx \overrightarrow{\text{suit}}$
   - In the sense of parenthood, $\overrightarrow{\text{maternity}} \approx \overrightarrow{\text{paternity}}$, albeit with some differences (e.g., "paternity ward" is not an established concept)
   - But $\overrightarrow{\text{maternity dress}} \not\approx \overrightarrow{\text{paternity suit}}$ in that they occur in completely different conceptual spaces

   (a) (8 points) Give another such example of two pairs of words, each pair of which indicates a meaningful relationship, though the meanings of the resulting phrases (though well defined) are remote from each other.

   I prefer the example to be in English (for ease of grading). However, it is fine to give an example from a language other than English but in that case please document the individual words (original plus phonetic writing plus meanings and links to some online dictionary) and ideally the name of some local faculty member or PhD student who speaks that language. I may check your example with someone who knows that language.

   Explain your answer in ∼20 words.

   (b) (24 points) Describe the problem in technical terms and describe your proposed approach. You could describe it in pseudocode or as equations that describe the model. You should follow the original CBOW or Skipgram models or as those approaches are described by Jurafsky and Martin or by Goldberg and Levy. An answer based on a powerful method such as BERT or GPT is not acceptable.

   Explain your answer in ∼60–90 words. Be sure to highlight where your approach differs from the original approaches and how your approach addresses the above challenge.

3. Consider Verbelese, a dialect of English. Verbelese has a quirk that reveals the sentiment being ascribed to a sentence. Verbelese is surprisingly similar to the $\mathscr{L}_1$ from the textbook (Figure 17.8 in the 2023-01-07 draft). (Native speakers of Verbelese switch between a and an as in English, but we can ignore that complexity.)

   - For positive sentiment, they place hah after the main verb and ahh before its direct object.

     The flight from Houston includes a meal

     becomes (when positively described)

     The flight from Houston includes hah an ahh meal
   - For negative sentiment, they place oyez before the main verb and um before its subject (and just an um if there is no explicit subject).

     The flight from Houston includes a meal

     becomes (when negatively described)

     Um the flight from Houston oyez includes a meal

   (a) (10 points) Write the equivalent of the following English sentence in Verbelese in each sentiment.

   The cat in the hat sells tofu and noodles.

   (b) (12 points) State a context-free grammar that modifies $\mathscr{L}_1$ to support Verbelese. Highlight the productions you would delete and those you would insert to convert $\mathscr{L}_1$ to support Verbelese dialect. Explain each addition or deletion briefly (∼5–8 words each.)

   (c) (8 points) Convert the grammar you produced above to Chomsky Normal Form.

   (d) (4 points) Provide a constituency parse for the sentence produced with a positive reading and equivalent to this one (and using your grammar):

   Tofu and noodles raise your blood sugar.

   (e) (4 points) Show the head word for each constituent in your parse above.