LING 430 Fall 2026

HW5

Due Thursday Oct 29, 11:59 pm, submit on Canvas

Submission guidelines:

- This is a written homework: work on paper or type it up, whichever you prefer.
- Submit one PDF on Canvas under Assignments > HW5; a clear scan or photo of handwriting is fine.
- Show your counts and fractions, not only the final numbers.

AI usage

If applicable, list any AI tools you used for this homework, what you asked them, and where the result went into your work. If you used none, write "none".

Declaration:

An HMM tagger from a small corpus

In this homework you will build an HMM tagger based on a small corpus made up of the following seven sentences. Each sentence has been tagged by hand with two tags, N for noun and V for verb.

```
dogs/N bark/V
cats/N watch/V dogs/N
people/N watch/V birds/N
birds/N fly/V
watch/N dogs/N bark/V
dogs/N fly/V
watch/V birds/N
```

You will work through the tagger on paper in questions 1 to 5: first the three tables, then the tagging of a new sentence. In question 6 you will compare your tagger with the one from nltk. Write every probability as a fraction, for example $6/7$.

1. The start table π

Count how many sentences begin with each tag and fill in π, the probability that a sentence begins with N or with V.

π	
N	
V	

2. The transition table A

Count the tag bigrams and fill in A, the probability of each tag given the previous tag. The rows are the previous tag. The last tag of a sentence does not begin a bigram.

A	N	V
N		
V		

3. The emission table B

Count the words under each tag and fill in B, the probability of each word given its tag.

B	dogs	cats	people	birds	watch	bark	fly
N							
V							

4. Tagging a new sentence

The sentence *watch dogs bark* has two readings: *watchdogs bark*, N N V, and *watch the dogs bark*, V N V. Fill in the six factors of each path from your tables and multiply them to a decimal. Which reading does your tagger prefer, and by how much?

N N V: $P(\text{N}) \times P(\text{watch} \mid \text{N}) \times P(\text{N} \mid \text{N}) \times P(\text{dogs} \mid \text{N}) \times P(\text{V} \mid \text{N}) \times P(\text{bark} \mid \text{V})$

$= ____ \times ____ \times ____ \times ____ \times ____ \times ____ = ____$

V N V: $P(\text{V}) \times P(\text{watch} \mid \text{V}) \times P(\text{N} \mid \text{V}) \times P(\text{dogs} \mid \text{N}) \times P(\text{V} \mid \text{N}) \times P(\text{bark} \mid \text{V})$

$= ____ \times ____ \times ____ \times ____ \times ____ \times ____ = ____$

Answer:

5. The other paths

With two tags and three words there are eight possible tag sequences for *watch dogs bark*. Explain why the other six have probability 0.

Answer:

6. Comparing with nltk

Paste the following into a code cell in Colab and run it. It tags *watch dogs bark* and the seven sentences of the corpus with the nltk tagger from class 9a.

```
import nltk
nltk.download("averaged_perceptron_tagger_eng", quiet=True)

sentences = ["watch dogs bark",
             "dogs bark", "cats watch dogs", "people watch birds", "birds fly",
             "watch dogs bark", "dogs fly", "watch birds"]
for s in sentences:
    print(nltk.pos_tag(s.split()))
```

Treat NN, NNS and NNP as N, and VB, VBP and VBZ as V. Does nltk choose the same reading of *watch dogs bark* as your tagger? Where does it disagree with the hand tags in the corpus? Then reflect in three to five sentences: the nltk tagger was trained on a million words of newspaper text and yours on seven sentences. Which one would you trust on this sentence, and what would it take to change your mind?

Answer: