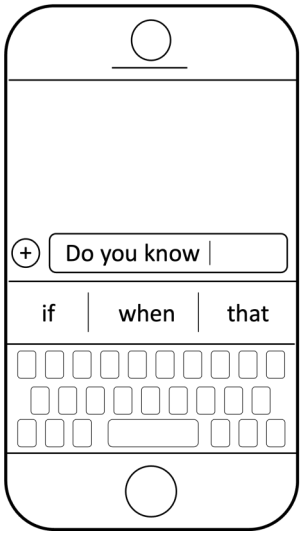
(E) Bengalese Finch Song (1/3) [5 Points]

One important feature of human language is unpredictability. If you knew what people were going to say beforehand, they wouldn't need to say it!

But language isn't completely random either. If you've ever used the auto-complete feature on a phone, you'll know that it can often predict your next word based on the words you've already typed. For instance, after the word *know*, your phone might suggest *if*, or *when*, or *that*, since those are the words that are most frequent after *know*.



This type of prediction is not just relevant for words: it can also be done with letters. For example, in English, after the letter *q*, we can predict with nearly 100% certainty that the next letter will be (you guessed it!) *u*. But for most letters in English, the next letter is less certain. After *t*, the most common next letter is *h*, which follows it around 40% of the time (mostly because of a few common words like *the*, *this*, *that*, and *then*) — but it is also frequently followed by the vowels *o*, *e*, and *i*.

For a computer to predict what letter will come next, the computer needs to store transitional probabilities, which define how likely each letter is to appear after each other letter. For example, we might say:

p(t → h) = 0.40

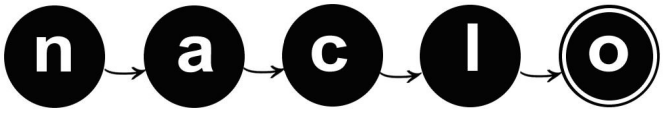
This means "the probability that *t* will be followed by *h* is 0.40." We can calculate transitional probabilities by using a piece of text. For instance, p(t → h) would be estimated as the number of times that *th* occurs in the text, divided by the number of times that *t* occurs.

Below is some text in Hawaiian, which uses 13 letters in its writing system. Spaces have been removed, and the symbol ` is a letter like any other, not punctuation — it's pronounced like the sound in the middle of the English exclamation *uh-oh*. At the top of the next page is a partially-filled table showing the steps to computing Hawaiian transitional probabilities (*L1* and *L2* stand for *letter 1* and *letter 2*).

```
ikinohihanakeakuaikalaniamekahonuahe`ano`ole
kahonuaua`oloheloheamalunanookahohonukapouli
ho`opunanaiholaka`uhaneokeakuamalunaokawai
```

E1. Fill in the blanks in the table on the next page.

E2. Suppose a Hawaiian speaker is typing, and the last letter they have typed is *k*. Based on the text above, what is the next letter most likely to be? Write your answer in the box to the right.



(E) Bengalese Finch Song (2/3)

Table described on the previous page:

L1	L2	Count of L1 followed by L2	Count of L1	P(L1 → L2)
i	h	3	8	3/8
k	u	2	12	1/6
k	e	2	12	1/6
k	i	**a.**	**b.**	**c.**
k	a	**d.**	**e.**	**f.**

Below are four sequences of letters. Two of the sequences are Hawaiian written in code (each letter has been replaced with a different letter). Don't worry about trying to figure out which letter in the code stands for which Hawaiian letter — that is not necessary for solving the problem. Note that the code used might be different between two texts of the same type—for example, *a* might stand for a different syllable in the first sequence of finch song than in the second.

The other two sequences are from the songs of two Bengalese finches, a species of songbird originally from Southeast Asia. In the case of the Bengalese finch song, the letters represent "syllables," which are the basic units that biologists have identified in their songs.

Bengalese finch song is interesting because, unlike many other types of birdsong, it is not entirely predictable — but it's still more predictable than human language. In other words, given that you hear one syllable, you can be more certain — on average — about which syllable you'll hear next, than you could be when reading human language.

Sequence A: abcbdaefbdgdabhijigdbcbhgdabkieidgdahbjaficblbfefkbcf

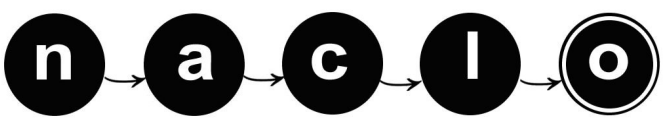
Sequence B: abacdefghahbhicdefghahbhgicdefgbjklcdefgammlcdefgajkl

Sequence C: abcdefbfcdefbfcdeaghijkbcdefafcdebcdeaghiffcdefafcdef

Sequence D: abcdedfgfdfgbahibjdkbghcbfcbffjdcbgbidgdldbgfdibjdgba

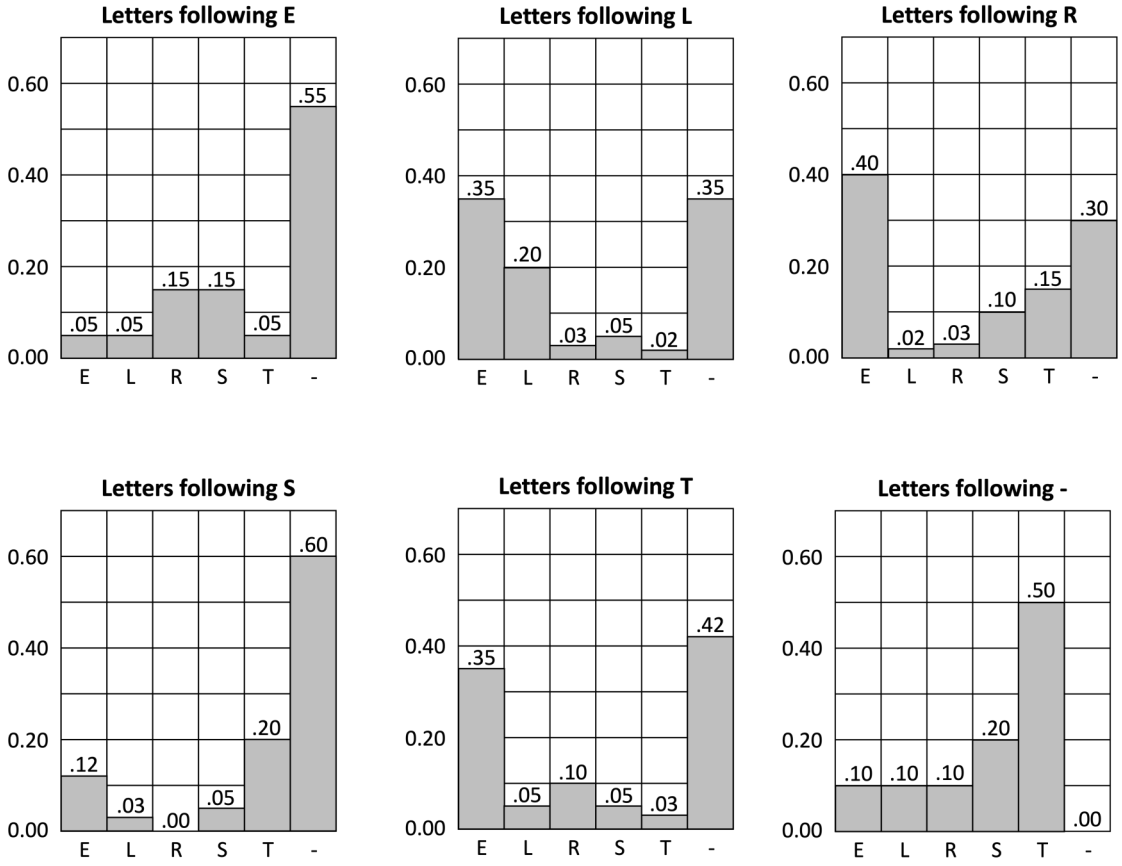
E3. Circle the two sequences that are Bengalese finch song.

Sequence A Sequence B Sequence C Sequence D



(E) Bengalese Finch Song (3/3)

Below are some plots showing transitional probabilities for a few English letters, estimated based on Wikipedia. For example, p(R → E) = 0.40. The dash symbol - is used to indicate a space.



Given a piece of text, we can label each letter with its transitional probability based on the previous letter. For instance, TREES-TREES-TREES would be labeled as shown here (with no label on the first letter, since it has no previous letter to use for determining the transitional probability):

	.10	.40	.05	.15	.60	.50	.10	.40	.05	.15	.60	.50	.10	.40	.05	.15
T	R	E	E	S	-	T	R	E	E	S	-	T	R	E	E	S

E4. Below is a phrase labeled in this way — but all of the letters are missing, leaving just the transitional probabilities! Fill in the blanks in the bottom row to complete the missing phrase. (The cell in the top left should remain blank.) In your answer, capitalization does not matter, and you can use either a space " " or a dash "-" to indicate a space.

	.40	.15	.20	.05	.35	.15	.05	.60	.10	.35	.05	.03	.35	.15	.10