

Intro to Cognitive & Information Sciences, Homework 3
Due in Class on May 5

1. Using Google, verify that the string *wine that is sweet is good* does not occur on the web (except in this assignment), but that *beer in cans are allowed* does. What does that tell you about the relationship between (intuitive) acceptability and use? (No more than 50 words).
2. Consider the string *wine that is sweet is*. Google it to verify that it does occur. Now consider the two strings *is wine that is sweet* and *is wine that sweet is*. Google them to determine whether these occur on the web.

Now go to <http://www.americancorpus.org/>. This is a user-friendly interface to a 385-million word corpus of contemporary American English. You will need to register, and we recommend taking the “brief tour” before getting started using it. Using the pull-down menu labeled “POS LIST”, check for the frequency of the following three schematic strings:

- i. Singular Noun – *that* – *is* – Adjective – *is*
- ii. *is* – Singular Noun – *that* – *is* – Adjective
- iii. *is* – Singular Noun – *that* – Adjective – *is*

How are the results of these searches relevant to the Poverty of the Stimulus argument? (No more than 150 words).

3. None of the following three strings appears in this corpus:
 - i. *wine that is sweet is good*
 - ii. *is wine that is sweet good*
 - iii. *is wine that sweet is good*

It is nevertheless possible to get a crude estimate of the probability of each of these strings by considering the probability of each two-word substring (known as a “bigram”). That is, if a string s consists of the words $w_1 w_2 \dots w_{n-1} w_n$, then we can compute a very rough approximation of $P(s)$ by the product of $P(w_2 | w_1) P(w_3 | w_2) \dots P(w_n | w_{n-1})$. For any bigram $w_i w_{i+1}$, we can estimate $P(w_{i+1} | w_i)$ as the frequency of the string $w_i w_{i+1}$ in the corpus divided by the frequency of w_i . Using this simple method, estimate the probabilities of each of the three strings above. How do these probability estimates line up with your intuitions about the naturalness of the three strings if taken as sentences? (No more than 100 words)

4. Does this show that Chomsky was wrong when he wrote (in the assigned paper):
[T]he notion ‘probability of a sentence’ is an entirely useless one, under any known interpretation of this term. On empirical grounds, the probability of my producing some given sentence of English – say, this sentence, or the sentence “birds fly” or “Tuesday follows Monday”, or whatever – is indistinguishable from the probability of my producing a given sentence of Japanese.
Justify your answer in no more than 200 words.

5. One argument against analyzing sentences in terms of bigrams – or n-grams for any fixed positive integer n – is that there are dependencies between elements in sentences that can be arbitrarily far away from each other. In English, for example, expressions of the forms in (a) are well-formed, but not those in (b), where X and Y can be arbitrary declarative sentences:

(a) *either X or Y*
if X, then Y

(b) **either X then Y*
**if X or Y*

Anyone who claims that X is crazy

**Anyone who claims that X are crazy*

That is, *either* goes with *or*; *if* goes with *then*; and *anyone* goes with *is*. Give values (that is, strings of words) for X and Y that show that a massive inventory of 5-grams (with or without probabilities attached) would not suffice to account for all the facts of English syntax. Then write simple phrase structure rules to capture these dependencies.

6. If n-grams are not an adequate model for natural language syntax, then what is the relevance of probabilities computed on the basis of n-grams to claims about the learnability of syntax? (No more than 250 words)