Psych156A/ Ling150
Spring 2014
Review Questions: Grammatical Categorization

(1) Terms/concepts to know: grammatical category, noun, verb, adjective, preposition, semantic bootstrapping hypothesis, frequent frame, corpora, bigram, productivity, lexical overlap, combinatorial diversity, Zipfian distribution, word rank, word frequency, observed overlap, expected overlap

(2) One way to think about a word's grammatical category (like noun or verb) is that it is really just a description of the way that word can be used in the language. Given the following contexts, say whether you think the novel word in each example (indicated by CAPITALS) is a noun (like "goblin"), a verb (like "sing"), an adjective (like "hot"), an adverb (like "dreamily"), or a preposition (like "in"). Be sure to briefly explain why you think so. (You might find it helpful to substitute words you know in place of the novel words, and see which ones fit best.)

        (a) That's a very BOFT bog.
        (b) Is Sir Didymus REKKing the Bog of Eternal Stench?
        (c) Ludo shouldn't have STROOPed his paw in the bog.
        (c) Sarah was bitten by a fairy PRING the wall.
        (d) They had to beware the BREER's tricks.
        (e) Hoggle couldn't believe how FREEMILY he was able to do it.

(3) Why can't the semantic bootstrapping hypothesis be used to categorize all the words of a language? (Hint: What mapping rules does it assume?)

(4) Here an example utterance from the imaginary Guin language.

"felgo bofty mu az berg mu merk berg felgo zu mu var berg az porto mu freggo berg felgo seech mu set berg draz pino trem felgo trem mu peri berg lootem bleeter."

        (a) What is the most frequent frame in this utterance?
        (b) Which words would this frame cluster together?
        (c) What is the second most frequent frame in this utterance?
        (d) What words would this second most frequent frame cluster together?

(5) By using the frequent frames learning strategy on a large set of Guin data, suppose you were able to identify 120 words that seem to behave like nouns. When Guin speakers are asked about them, you find out that 65 of the 120 words you identified did belong to the same category (nouns). However, from the data set you had, you should have found 35 more nouns than you actually did correctly identify. Calculate the precision and recall for the frequent frames approach on this set of Guin data. Be sure to show how you calculated each one.

(6) What does it mean if a grammatical categorization model has a high precision score, but a low recall score? Is this model likely to correctly identify a word as belonging to a particular grammatical category? Is this model likely to correctly group all the words of one category together? What would it mean if the opposite were true, and the model has a high recall score but a low precision score? (Hint: Imagine a scenario where the model groups all words into a single category.) Is this model likely to correctly identify a word as belonging to a particular grammatical category? Is this model likely to correctly group all the words of one category together?

(7) What experimental evidence do we have that very young children are sensitive to frequent frames in the input?

(8) Is it better for the categorization unit to be a frame or to be a sequential unit like a bigram?

(9) Is it better for the units that make up the frames to be words or to be more abstract units (for example, clusters of words that make up a category)?

(10) Do frequent frames always operate best over words? (Hint: Think about languages like Turkish and German.) How well do frequent frames (using any unit) operate cross-linguistically?

(11) What considerations about children's memory did the model used by Wang & Mintz (2008) make? (Think about what they were trying to improve on from Mintz (2003)'s study.) How did the model by Wang & Mintz (2008) perform, compared to the study done by Mintz (2003)?

(12) Even if a child knows the category Noun, why wouldn't we expect that child to produce all possible nouns in combination with all the determiners (like "a" and "the") that the child knows? That is, if a child knows the nouns "penguin", "kitty", and "puppy", should we expect the child to spontaneously produce all of the following in natural communication settings: "a penguin", "the penguin", "a kitty", "the kitty", "a puppy", and "the puppy"? Why or why not?

(13) What does it mean for linguistic output to have a Zipfian distribution? Why does this make it hard for us to assess everything a child knows? (Hint: Think about what this means for figuring out lexical overlap when determining knowledge about grammatical categories.)

(14) How does word rank differ from word frequency? Does a word with a fairly high rank (say 10 out of 200) need to have a very high frequency?

(15) Does the formal metric proposed by Yang (2010) calculate expected overlap or observed overlap?

(16) According to Yang's metric, is young children's naturalistic output consistent with knowing the category *Noun*?

(17) According to Yang's metric, does the specific 20 to 24-month-old child examined seem to know the category *Verb*? Does this child seem to know basic word order (that is, the way subjects, verbs, and objects go together)? Does this child seem to know a number of different lexical items that are verbs?