Lecture 13. Are Natural Languages finite-state languages? (and other questions)

Reading: Section 17.3 of PtMW: Regular Languages. pp. 471-480.
Especially in connection with Marc Hauser's work: (See links in WHISC of September 23: http://people.umass.edu/potts/whisc/whisc-2004-9-23.html#hauser .)

- Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science* 298(5598):1569-1579, 22 November.
- Thomas Bever and Mario Montalbetti. 2002. Noam's ark. *Science* 298(5598):1565-1566, 22 November.
- Mark Liberman. September 3, 2003. Update on Fitch & Hauser. Linguist List 15.2450.
- W. Tecumseh Fitch and Mark D. Hauser. 2004. Computational constraints on syntactic processing in a nonhuman primate. *Science Magazine* 303(5656):377-380, 16 January.
- Pierre Perruchet and Arnaud Rey. 2004. Does the mastery of center-embedded linguistic structures distinguish humans from nonhuman primates? To appear in the *Psychonomic Bulletin and Review*.
- Mark Liberman. January 16, 2004. Language in humans and monkeys. Language Log.
- Mark Liberman. January 16, 2004. Hi Lo Hi Lo, it's off to formal language theory we go. Language Log.
- Mark Liberman. August 31, 2004. Humans context-free, monkeys finite-state? Apparently not. Language Log.
- Greg Kochanski. 2004. Is a phrase structure grammar the important difference between humans and monkeys? A comment on 'Computational constraints on syntactic processing in a nonhuman primate'.
- Ray Jackendoff and Steven Pinker. In press. The faculty of language: What's special about it? *Cognition*.

0. Review.

First, let's do on the board three finite-state automata that will be relevant to the ensuing discussion.

1. $(ab)^n$, i.e. $\{(ab)^n | n>0\}$ (This is the language the tamarin monkeys reportedly learned.)

2. {ab, aabb, aaabbb, aaaabbbb} (This is a finite sublanguage of the non-finite-state language a^nb^n.)

3. $aA^*a \cup bA^*b$, where A = {a,b}, i.e. the set of all strings of a's and b's of length ≥ 2 which begin and end with the same symbol. (This is a finite-state language with some long-distance dependency, but no center-embedding, showing that we have to separate those issues.)

Second, let's recall the **closure properties** of the set of finite-state languages (= regular languages, = languages which can be generated by a type 3 grammar, alias left-linear or right-linear grammar.) These were discussed in Lecture 12; they can all be shown by construction algorithms on finite-state automata.

Assume a fixed alphabet *A*. So all the languages to be considered are subsets of *A**.

Union: If *X, Y* are regular languages, then so is $X \cup Y$.
Concatenation: If *X, B* are regular languages, then so is *XB*.
Kleene star: If *X* is a regular language, then so is *X**.
Complementation: If *X* is a regular language, then $A^* - X$ is a regular language.
Intersection: If *X, Y* are regular languages, then so is $X \cap Y$.

And given that the empty language ∅ and the universal language *A** are regular languages (accepted by 1-state finite state automata, the simplest ones possible), we can conclude that the class of regular languages over any fixed alphabet is a Boolean algebra.

1. Inadequacy of Type 3 grammars for natural languages: Classic examples

Is English a regular language? No. The typical proof (as in PtMW, 478-9) uses closure under intersection plus the pumping lemma (pumping theorem).

First let's practice using the pumping lemma some more to show that some languages on the alphabet {a,b} are not regular languages.

Example 1 (repeated from Lecture 12): $\{a^n b^n \mid n > 0\}$: the set of all strings of n a's followed by n b's. Good strings: ab, aabb, aaabbb, aaaabbbb, … . Bad strings: aa, aabbb, abab, bbaa, …

Recall what the pumping lemma says: (= theorem 17.2 in PtMW)

> **Theorem 17.2** If *L* is an infinite fal over alphabet *A*, then there are strings $x,y,z \in A^*$ such that $y \neq e$ and $xy^n z \in L$ for all $n \geq 0$.

For example 1, the $a^n b^n$ languages, we argued that there can be no such string *y*, no matter what we take for *x* and *z*. We argued that there are only three possibilities for *y*: either it's some string of a's, or some string of b's, or some string of a's followed by some string of b's. And then it's immediately obvious that if you take some grammatical string containing such a string *y* and then allow that string *y* to 'loop', i.e. to repeat 2 or more times, the resulting string will NOT be of the form $a^n b^n$.

Example 2. $\{x\{x^R \mid x \in \{a,b\}^*\}$: the set of all strings of a's and b's in which the second half of the string is a mirror image of the first half. All the strings are of even length. The empty string is

included in this 'mirror-image language'. Good strings: e, aa, bb, abba, aaaa, abbbba, ababbaba, … . Bad strings: a, b, aba, aabb, ababa, bbba, abbba, aaaaa, abbaba, ….

In this case it's not so easy to apply the pumping lemma directly, but we can employ a common strategy, namely first to intersect this language with a known regular language to give a language to which the pumping lemma can be applied straightforwardly. Why is this legitimate? Because we know (lecture 12; see PtMW 475-77) that the intersection of any two regular languages is regular.

So let's intersect the mirror-image language $\{x\{x^R| \ x \in \{a,b\}^*\}$ with the regular language aa*bbaa* (i.e. with the set of all strings containing one or more a's followed by exactly two b's followed by one or more a's.) The intersection gives the set $\{a^nbba^n \mid n > 0 \}$. And to this set it's easy to apply the pumping lemma in a way similar to how we did it in Example 1.

Back to English. For English, we use the technique of intersection with a regular language to find a subset of English to which we can apply the pumping lemma: then since that language is not regular, and it's the intersection of English with a regular language, we can conclude that English is not a regular language. (Note: it does NOT work just to find some subset of English that's not regular and conclude that English is not regular. Why? This is important – it's something that has marred many arguments that can be found in the literature. Similarly for proofs of non-Context-Freeness, to be discussed in Lecture 14.)

Another advantage of using the intersection strategy is that we don't have to have a complete specification of what English is to start with; we just have to get agreement that the proposed intersection language is indeed a subset of English.

So we look at the following sublanguage (PtMW p. 478-9): {the cat died, the dog died, the elephant died, the rat died, the cat the dog chased died, the elephant the rat bit died, the cat the dog the rat bit chased died, the elephant the cat the dog the rat bit chased admired died, the cat the dog the cat the rat bit bit admired died, … }

For this language, we use a particular set A of nouns (dog, cat, rat, elephant), a particular set B of transitive verbs (chased, bit, admired), and one intransitive verb (died), and a self-embedding relative clause construction.

The sentences are all of the form $(\text{the N})^n \ V_{trans}^{n-1} \ V_{intrans}$, or $x^n \ y^{n-1}$ *died,* where $x \in$ A and $y \in$ B. This language can be taken to be the intersection of English with the regular language $x^* \ y^*$ *died,* where $x \in$ A and $y \in$ B.

Caveat: In order to accept this argument, you have to believe that English contains arbitrarily deeply center-embedded relative clauses of this sort. If you put any finite upper bound on the length of the strings or the depth of center-embedding, then the result will be a finite-state language. See the Van Noord handout "Arguments for Finite-State Language Processing" for discussion of why at the very least, finite-state grammars can give good approximations to natural language, and why some researchers are inclined to go further and suggest that they may be good models of human natural language processing.

2. Issues raised in Hauser and Fitch's work.

1. Experimental evidence that some animals can learn finite state grammars but not context-free grammars, reported in Fitch and Hauser (2004).

Fitch and Hauser's abstract:

> The capacity to generate a limitless range of meaningful expressions from a finite set of elements differentiates human language from other animal communication systems. Rule systems capable of generating an infinite set of outputs ("grammars") vary in generative power. The weakest possess only local organizational principles, with regularities limited to neighboring units. We used a familiarization/discrimination paradigm to demonstrate that monkeys can spontaneously master such grammars. However, human language entails more sophisticated grammars, incorporating hierarchical structure. Monkeys tested with the same methods, syllables, and sequence lengths were unable to master a grammar at this higher, "phrase structure grammar" level.

2. From Liberman January 17, 2004.

Here are the details. Fitch and Hauser explain about their stimuli that

The FSG was $(AB)^n$, in which a random "A" syllable was always followed by a single random "B" syllable, and such pairs were repeated n times. The corresponding PSG, termed A^nB^n, generated strings with matched numbers of A and B syllables. In this grammar, n sequential "A" syllables must be followed by precisely n "B" syllables.

So the "finite state" language is $(AB)^n$ for n=2 and n=3, i.e. exactly the set of two patterns {ABAB, ABABAB}, while the "phrase structure" language is A^nB^n, for n=2 and n=3, i.e. exactly the set of two patterns {AABB, AAABBB}.

F&H motivate the lower limit on n (but not the upper one) as follows:

Because previous work demonstrates that tamarins can readily remember and precisely discriminate among strings up to three syllables in length, we restricted n to be two or three in both of the above grammars.

So it seems that these two "languages" -- intended to represent whole classes of formal grammatical power -- consisted of just two strings each, one four symbols long and the other six symbols long? Well, superficially, no -- the languages are much bigger than that, though still finite. A and B represent classes of syllables, with A being one of {ba di yo tu la mi no wu}, while B is one of {pa li mo nu ka bi do gu}. There are eight options for each class, and strings of syllables are formed by random selection without replacement, so the number of possible syllable strings in the FSG language is

8*8*7*7 + 8*8*7*7*6*6 = 116,032

and the number of possible syllable strings in the PSG language is the same.

[continuing with BHP paraphrase of Liberman:] But even though this makes a very large number of sentences, it turns out that what was really salient was that all the A syllables were spoken by a female voice and all the B syllables by a male voice, and given the task, there was

no need to pay attention to anything else. So we are effectively back to two languages of two strings each: {ABAB, ABABAB} (the "finite-state language") and {AABB, AAABBB} (the "context-free" language). As Liberman points out, there are far too many alternative explanations for the observed differences in behavior of tamarins and humans on this task to draw any strong conclusions relating to finite state vs. context-free languages.

There *was* a difference in behavior, which does need to be explained; Liberman makes some suggestions, Perruchet and Rey make others, but in any case it's clear that there are many possible hypotheses that have not been ruled out, and no real basis for an explanation based on the Chomsky hierarchy of classes of grammars, or on recursion.

3. Perruchet and Rey's abstract of their paper:

> In a recent *Science* paper, Fitch and Hauser (2004; hereafter, F&H) claimed to have demonstrated that Cotton-top Tamarins fail to learn an artificial language produced by a Phrase Structure Grammar (PSG, Chomsky, 1957) generating center-embedded sentences, while adult humans easily learn such a language. We report an experiment replicating the results of F&H in humans, but also showing that participants learned the language without exploiting in any way the center-embedded structure. When the procedure was modified to make the processing of this structure mandatory, participants no longer showed evidence of learning. We propose a simple interpretation for the difference in performance observed in F&H's task between humans and Tamarins, and argue that, beyond the specific drawbacks inherent to F&H's study, researching the source of the inability of nonhuman primates to master language within a framework built around the Chomsky's hierarchy of grammars is a conceptual dead-end.

Discussion: Even if one extended the Fitch and Hauser language to unbounded lengths, the two languages are just $(AB)^n$ (which is just an abbreviation for AB(AB)*) vs. A^nB^n: the lengths of the two halves must match, but there are no actual dependencies among the A's and the B's that would force a context-free-grammar "structure" for the language. As Liberman notes in his later blog, Greg Kuchinski already criticized F&H's claim that humans were shown to be able to learn a CF language with center-embedding, when a counting algorithm would work just as well for the given task.

P&R change the task so that for the CF language, each A has to match the corresponding B: the syllables are paired, and the grammatical strings are all of the form $A_{n1}A_{n2} \ldots A_{nk}B_{nk} \ldots B_{n2}B_{n1}$. In this case, one really does need to keep track of center-embedding: the strings are similar to mirror image strings, but as Liberman notes, even harder, since the pairing of A-B syllables is arbitrary.

And P&R found that humans couldn't learn the revised language, even when the length was just 4 or 6 syllables as in the original F&H experiment, and using the same experimental techniques as used by F&H. In their testing, they found the humans sensitive to the pitch alternation patterns (the patterns of As and Bs), but not to the dependencies between particular A's and particular B's.

Liberman suggests that while F&H's task was clearly too easy to conclude anything about acquisition of context-free languages, P&R's might be too hard because of the arbitrariness of the pairings, and suggests that it would be interesting to run an intermediate experiment where

the data have more in common with patterns of agreement – either a true mirror-image language where the relation is one of identity, or something equally systematic.

In any case, as Liberman notes in his Aug 31 2004 blog, "P&R put the ball firmly back in the court of anyone who wants to claim a relationship between the levels of the Chomsky hierarchy and the different propensities of humans and monkeys to notice things about sets of strings of spoken syllables."