23. Natural Language for Communication

23.1 [washing-clothes-exercise]Read the following text once for understanding, and remember as much of it as you can. There will be a test later.

> The procedure is actually quite simple. First you arrange things into different groups. Of course, one pile may be sufficient depending on how much there is to do. If you have to go somewhere else due to lack of facilities that is the next step, otherwise you are pretty well set. It is important not to overdo things. That is, it is better to do too few things at once than too many. In the short run this may not seem important but complications can easily arise. A mistake is expensive as well. At first the whole procedure will seem complicated. Soon, however, it will become just another facet of life. It is difficult to foresee any end to the necessity for this task in the immediate future, but then one can never tell. After the procedure is completed one arranges the material into different groups again. Then they can be put into their appropriate places. Eventually they will be used once more and the whole cycle will have to be repeated. However, this is part of life.

23.2 An *HMM grammar* is essentially a standard HMM whose state variable is N (nonterminal, with values such as Det, $Adjective$, $Noun$ and so on) and whose evidence variable is W (word, with values such as is, $duck$, and so on). The HMM model includes a prior ${\textbf{P}}(N_0)$, a transition model ${\textbf{P}}(N_{t+1}|N_t)$, and a sensor model ${\textbf{P}}(W_t|N_t)$. Show that every HMM grammar can be written as a PCFG. [Hint: start by thinking about how the HMM prior can be represented by PCFG rules for the sentence symbol. You may find it helpful to illustrate for the particular HMM with values A, B for N and values x, y for W.]

23.3 Consider the following PCFG for simple verb phrases:

> 0.1: VP $\rightarrow$ Verb
>
> 0.2: VP $\rightarrow$ Copula Adjective
>
> 0.5: VP $\rightarrow$ Verb the Noun
>
> 0.2: VP $\rightarrow$ VP Adverb
>
> 0.5: Verb $\rightarrow$ is
>
> 0.5: Verb $\rightarrow$ shoots
>
> 0.8: Copula $\rightarrow$ is
>
> 0.2: Copula $\rightarrow$ seems
>
> 0.5: Adjective $\rightarrow$ **unwell**
>
> 0.5: Adjective $\rightarrow$ **well**
>
> 0.5: Adverb $\rightarrow$ **well**
>
> 0.5: Adverb $\rightarrow$ **badly**
>
> 0.6: Noun $\rightarrow$ **duck**
>
> 0.4: Noun $\rightarrow$ **well**

1. Which of the following have a nonzero probability as a VP? (i) shoots the duck well well well(ii) seems the well well(iii) shoots the unwell well badly
2. What is the probability of generating “is well well”?
3. What types of ambiguity are exhibited by the phrase in (b)?
4. Given any PCFG, is it possible to calculate the probability that the PCFG generates a string of exactly 10 words?

23.4 Consider the following simple PCFG for noun phrases:

0.6: NP $\rightarrow$ Det\ AdjString\ Noun

0.4: NP $\rightarrow$ Det\ NounNounCompound

0.5: AdjString $\rightarrow$ Adj\ AdjString

0.5: AdjString $\rightarrow$ Λ

1.0: NounNounCompound $\rightarrow$ Noun

0.8: Det $\rightarrow$ **the**

0.2: Det $\rightarrow$ **a**

0.5: Adj $\rightarrow$ **small**

0.5: Adj $\rightarrow$ **green**

0.6: Noun $\rightarrow$ **village**

0.4: Noun $\rightarrow$ **green**

where Λ denotes the empty string.

1. What is the longest NP that can be generated by this grammar? (i) three words(ii) four words(iii) infinitely many words
2. Which of the following have a nonzero probability of being generated as complete NPs? (i) a small green village(ii) a green green green(iii) a small village green
3. What is the probability of generating “the green green”?
4. What types of ambiguity are exhibited by the phrase in (c)?
5. Given any PCFG and any finite word sequence, is it possible to calculate the probability that the sequence was generated by the PCFG?

23.5 Outline the major differences between Java (or any other computer language with which you are familiar) and English, commenting on the “understanding” problem in each case. Think about such things as grammar, syntax, semantics, pragmatics, compositionality, context-dependence, lexical ambiguity, syntactic ambiguity, reference finding (including pronouns), background knowledge, and what it means to “understand” in the first place.

23.6 This exercise concerns grammars for very simple languages.

1. Write a context-free grammar for the language $a^n b^n$.
2. Write a context-free grammar for the palindrome language: the set of all strings whose second half is the reverse of the first half.
3. Write a context-sensitive grammar for the duplicate language: the set of all strings whose second half is the same as the first half.

23.7 Consider the sentence “Someone walked slowly to the supermarket” and a lexicon consisting of the following words:

$Pronoun \rightarrow \textbf{someone} \quad Verb \rightarrow \textbf{walked}$

$Adv \rightarrow \textbf{slowly} \quad Prep \rightarrow \textbf{to}$

$Article \rightarrow \textbf{the} \quad Noun \rightarrow \textbf{supermarket}$

Which of the following three grammars, combined with the lexicon, generates the given sentence? Show the corresponding parse tree(s).

$\quad\quad\quad\quad (A):\quad\quad\quad\quad$	**$\quad\quad\quad\quad(B):\quad\quad\quad\quad$**	**$\quad\quad\quad\quad(C):\quad\quad\quad\quad$**
$S\rightarrow NP\space VP$	$S\rightarrow NP\space VP$	$S\rightarrow NP\space VP$
$NP\rightarrow Pronoun$	$NP\rightarrow Pronoun$	$NP\rightarrow Pronoun$
$NP\rightarrow Article\space Noun $	$NP\rightarrow Noun$	$NP\rightarrow Article\space NP$
$VP\rightarrow VP\space PP$	$NP\rightarrow Article\space NP$	$VP\rightarrow Verb\space Adv$
$VP\rightarrow VP\space Adv\space Adv$	$VP\rightarrow Verb\space Vmod$	$Adv\rightarrow Adv\space Adv$
$VP\rightarrow Verb$	$Vmod\rightarrow Adv\space Vmod$	$Adv\rightarrow PP$
$PP\rightarrow Prep\space NP$	$Vmod\rightarrow Adv$	$PP\rightarrow Prep\space NP$
$NP\rightarrow Noun$	$Adv\rightarrow PP$	$NP\rightarrow Noun$
$\quad$	$PP\rightarrow Prep\space NP$	$\quad$

For each of the preceding three grammars, write down three sentences of English and three sentences of non-English generated by the grammar. Each sentence should be significantly different, should be at least six words long, and should include some new lexical entries (which you should define). Suggest ways to improve each grammar to avoid generating the non-English sentences.

23.8 Collect some examples of time expressions, such as “two o’clock,” “midnight,” and “12:46.” Also think up some examples that are ungrammatical, such as “thirteen o’clock” or “half past two fifteen.” Write a grammar for the time language.

23.9 Some linguists have argued as follows:

> Children learning a language hear only *positive examples* of the language and no *negative examples*. Therefore, the hypothesis that “every possible sentence is in the language” is consistent with all the observed examples. Moreover, this is the simplest consistent hypothesis. Furthermore, all grammars for languages that are supersets of the true language are also consistent with the observed data. Yet children do induce (more or less) the right grammar. It follows that they begin with very strong innate grammatical constraints that rule out all of these more general hypotheses *a priori*.

Comment on the weak point(s) in this argument from a statistical learning viewpoint.

23.10 [chomsky-form-exercise] In this exercise you will transform $\large \varepsilon_0$ into Chomsky Normal Form (CNF). There are five steps: (a) Add a new start symbol, (b) Eliminate ϵ rules, (c) Eliminate multiple words on right-hand sides, (d) Eliminate rules of the form (${\it X}$ ${{\;}}\rightarrow{{\;}}$${\it Y}$), (e) Convert long right-hand sides into binary rules.

1. The start symbol, S, can occur only on the left-hand side in CNF. Replace ${\it S}$ everywhere by a new symbol ${\it S'}$ and add a rule of the form ${\it S}$ ${{\;}}\rightarrow{{\;}}$${\it S'}$.
2. The empty string, ϵ cannot appear on the right-hand side in CNF. $\large \varepsilon_0$ does not have any rules with ϵ, so this is not an issue.
3. A word can appear on the right-hand side in a rule only of the form (${\it X}$ ${{\;}}\rightarrow{{\;}}$*word*). Replace each rule of the form (${\it X}$ ${{\;}}\rightarrow{{\;}}$…*word* …) with (${\it X}$ ${{\;}}\rightarrow{{\;}}$…${\it W'}$ …) and (${\it W'}$ ${{\;}}\rightarrow{{\;}}$*word*), using a new symbol ${\it W'}$.
4. A rule (${\it X}$ ${{\;}}\rightarrow{{\;}}$${\it Y}$) is not allowed in CNF; it must be (${\it X}$ ${{\;}}\rightarrow{{\;}}$${\it Y}$ ${\it Z}$) or (${\it X}$ ${{\;}}\rightarrow{{\;}}$*word*). Replace each rule of the form (${\it X}$ ${{\;}}\rightarrow{{\;}}$${\it Y}$) with a set of rules of the form (${\it X}$ ${{\;}}\rightarrow{{\;}}$…), one for each rule (${\it Y}$ ${{\;}}\rightarrow{{\;}}$…), where (…) indicates one or more symbols.
5. Replace each rule of the form (${\it X}$ ${{\;}}\rightarrow{{\;}}$${\it Y}$ ${\it Z}$ …) with two rules, (${\it X}$ ${{\;}}\rightarrow{{\;}}$${\it Y}$ ${\it Z'}$) and (${\it Z'}$ ${{\;}}\rightarrow{{\;}}$${\it Z}$ …), where ${\it Z'}$ is a new symbol.

Show each step of the process and the final set of rules.

23.11 Consider the following toy grammar:

> $S \rightarrow NP\space VP$
>
> $NP \rightarrow Noun$
>
> $NP \rightarrow NP\space and\space NP$
>
> $NP \rightarrow NP\space PP$
>
> $VP \rightarrow Verb$
>
> $VP \rightarrow VP\space and \space VP$
>
> $VP \rightarrow VP\space PP$
>
> $PP \rightarrow Prep\space NP$
>
> $Noun \rightarrow Sally\space; pools\space; streams\space; swims$
>
> $Prep \rightarrow in$
>
> $Verb \rightarrow pools\space; streams\space; swims$

1. Show all the parse trees in this grammar for the sentence “Sally swims in streams and pools.”
2. Show all the table entries that would be made by a (non-probabalistic) CYK parser on this sentence.

23.12 [exercise-subj-verb-agree] Using DCG notation, write a grammar for a language that is just like $\large \varepsilon_1$, except that it enforces agreement between the subject and verb of a sentence and thus does not generate ungrammatical sentences such as “I smells the wumpus.”

23.13 Consider the following PCFG:

$S \rightarrow NP \space VP[1.0] $

$NP \rightarrow \textit{Noun}[0.6] \space|\space \textit{Pronoun}[0.4] $

$VP \rightarrow \textit{Verb} \space NP[0.8] \space|\space \textit{Modal}\space \textit{Verb}[0.2]$

$\textit{Noun} \rightarrow \textbf{can}[0.1] \space|\space \textbf{fish}[0.3] \space|\space ...$

$\textit{Pronoun} \rightarrow \textbf{I}[0.4] \space|\space ...$

$\textit{Verb} \rightarrow \textbf{can}[0.01] \space|\space \textbf{fish}[0.1] \space|\space ...$

$\textit{Modal} \rightarrow \textbf{can}[0.3] \space|\space ...$

The sentence “I can fish” has two parse trees with this grammar. Show the two trees, their prior probabilities, and their conditional probabilities, given the sentence.

23.14 An augmented context-free grammar can represent languages that a regular context-free grammar cannot. Show an augmented context-free grammar for the language $a^nb^nc^n$. The allowable values for augmentation variables are 1 and $SUCCESSOR(n)$, where n is a value. The rule for a sentence in this language is $$S(n) {{{{\;}}\rightarrow{{\;}}}}A(n) {{\;}}B(n) {{\;}}C(n) \ .$$ Show the rule(s) for each of ${\it A}$, ${\it B}$, and ${\it C}$.

23.15 Augment the $\large \varepsilon_1$ grammar so that it handles article–noun agreement. That is, make sure that “agents” and “an agent” are ${\it NP}$s, but “agent” and “an agents” are not.

23.16 Consider the following sentence (from *The New York Times,* July 28, 2008):

> Banks struggling to recover from multibillion-dollar loans on real estate are curtailing loans to American businesses, depriving even healthy companies of money for expansion and hiring.

1. Which of the words in this sentence are lexically ambiguous?
2. Find two cases of syntactic ambiguity in this sentence (there are more than two.)
3. Give an instance of metaphor in this sentence.
4. Can you find semantic ambiguity?

23.17 [washing-clothes2-exercise] Without looking back at Exercise washing-clothes-exercise, answer the following questions:

1. What are the four steps that are mentioned?
2. What step is left out?
3. What is “the material” that is mentioned in the text?
4. What kind of mistake would be expensive?
5. Is it better to do too few things or too many? Why?

23.18 Select five sentences and submit them to an online translation service. Translate them from English to another language and back to English. Rate the resulting sentences for grammaticality and preservation of meaning. Repeat the process; does the second round of iteration give worse results or the same results? Does the choice of intermediate language make a difference to the quality of the results? If you know a foreign language, look at the translation of one paragraph into that language. Count and describe the errors made, and conjecture why these errors were made.

23.19 The D_i values for the sentence in Figure mt-alignment-figure sum to 0. Will that be true of every translation pair? Prove it or give a counterexample.

23.20 (Adapted from [@Knight:1999].) Our translation model assumes that, after the phrase translation model selects phrases and the distortion model permutes them, the language model can unscramble the permutation. This exercise investigates how sensible that assumption is. Try to unscramble these proposed lists of phrases into the correct order:

1. have, programming, a, seen, never, I, language, better
2. loves, john, mary
3. is the, communication, exchange of, intentional, information brought, by, about, the production, perception of, and signs, from, drawn, a, of, system, signs, conventional, shared
4. created, that, we hold these, to be, all men, truths, are, equal, self-evident

Which ones could you do? What type of knowledge did you draw upon? Train a bigram model from a training corpus, and use it to find the highest-probability permutation of some sentences from a test corpus. Report on the accuracy of this model.

23.21 Calculate the most probable path through the HMM in Figure sr-hmm-figure for the output sequence $[C_1,C_2,C_3,C_4,C_4,C_6,C_7]$. Also give its probability.

23.22 We forgot to mention that the text in Exercise washing-clothes-exercise is entitled “Washing Clothes.” Reread the text and answer the questions in Exercise washing-clothes2-exercise. Did you do better this time? Bransford and Johnson [@Bransford+Johnson:1973] used this text in a controlled experiment and found that the title helped significantly. What does this tell you about how language and memory works?