Ling 235 – answers to HW #4

March 2, 2005

1. There are many possible ways to change Suppes' grammar (which we'll call G) that will lead to a better fit of the data. Since the main source of poor fit in his grammar is the high frequency of the NN-type noun phrase, the easiest way to improve his grammar is to add a specific production rule that directly produces that noun phrase type. From a linguistic point of view, such a rule could be interpreted as a fixed template by which children construct saturated NPs just by combining two nouns – and the high frequency of the NN type could be considered evidence that such a template exists.

   (a)

	Production Rule	*Probability*
1.	NP $\rightarrow$N	a_1
2.	NP $\rightarrow$N + N	a_2
3.	NP $\rightarrow$AdjP	a_3
4.	NP $\rightarrow$AdjP + N	a_4
5.	NP $\rightarrow$Pro	a_5
6.	NP $\rightarrow$NP + NP	a_6
7.	AdjP $\rightarrow$AdjP + Adj	b_1
8.	AdjP $\rightarrow$Adj	b_2

   This grammar G' is identical to Suppes' except for the addition of rule 2. One nice property of this alternative grammar is that it contains the hypothesis space of Suppes' grammar, so we can compare the two grammars with likelihood-ratio and X^2 tests.

   In order to determine the maximum-likelihood estimates (MLEs) of the parameters of G', we assume (as per the problem description) a fixed bracketing/tree for each of the observed noun phrases, and get the following table:

Noun phrase	Bracketing	Observed frequency	Rule applications
N	$[_{NP}$ N$]$	1445	1
P	$[_{NP}$ P$]$	388	5
NN	$[_{NP}$ N N$]$	231	2
AN	$[_{NP}$ $[_{AdjP}$ A$]$ N$]$	135	4,8
A	$[_{NP}$ $[_{AdjP}$ A$]]$	114	3,8
PN	$[_{NP}$ $[_{NP}$ P$]$ $[_{NP}$ N$]]$	31	6,5,1
NA	$[_{NP}$ $[_{NP}$ N$]$ $[_{NP}$ $[_{AdjP}$ A$]]]$	19	6,1,3,8
NNN	$[_{NP}$ $[_{NP}$ N N$]$ $[_{NP}$ N$]]$	12	6,2,1
AA	$[_{NP}$ $[_{AdjP}$ $[_{AdjP}$ A$]$ A$]]$	10	3,7,8
NAN	$[_{NP}$ $[_{NP}$ N$]$ $[_{NP}$ $[_{AdjP}$ A$]$ N$]]$	8	6,1,4,8
AP	$[_{NP}$ $[_{NP}$ $[_{AdjP}$ A$]]$ $[_{NP}$ P$]]$	6	6,3,8,5
PPN	$[_{NP}$ $[_{NP}$ $[_{NP}$ P$]$ $[_{NP}$ P$]]$ $[_{NP}$ N$]]$	6	6,6,5,5,1
ANN	$[_{NP}$ $[_{NP}$ $[_{AdjP}$ A$]$ N$]$ $[_{NP}$ N$]]$	5	6,4,8,1
AAN	$[_{NP}$ $[_{AdjP}$ $[_{AdjP}$ A$]$ A$]$ N$]$	4	4,7,8
PA	$[_{NP}$ $[_{NP}$ P$]$ $[_{NP}$ $[_{AdjP}$ A$]]]$	4	6,5,3,8
ANA	$[_{NP}$ $[_{NP}$ $[_{AdjP}$ A$]$ N$]$ $[_{NP}$ $[_{AdjP}$ A$]]]$	3	6,6,7,8,3,8
APN	$[_{NP}$ $[_{NP}$ $[_{NP}$ $[_{AdjP}$ A$]]$ $[_{NP}$ P$]]$ $[_{NP}$ N$]]$	3	6,6,3,8,5,1
AAA	$[_{NP}$ $[_{AdjP}$ $[_{AdjP}$ $[_{AdjP}$ A$]$ A$]$ A$]]$	2	3,7,7,8
APA	$[_{NP}$ $[_{NP}$ $[_{NP}$ $[_{AdjP}$ A$]]$ $[_{NP}$ P$]]$ $[_{NP}$ $[_{AdjP}$ A$]]]$	2	6,6,3,8,5,3,8
NPP	$[_{NP}$ $[_{NP}$ $[_{NP}$ $[_{AdjP}$ N$]]$ $[_{NP}$ P$]]$ $[_{NP}$ P$]]$	2	6,6,1,5,5
PAA	$[_{NP}$ $[_{NP}$ $[_{NP}$ P$]$ $[_{NP}$ $[_{AdjP}$ $[_{AdjP}$ A$]$ A$]]]]$	2	6,5,3,7,8
PAN	$[_{NP}$ $[_{NP}$ P$]$ $[_{NP}$ $[_{AdjP}$ A$]$ N$]]$	2	6,5,4,8

From this table we can easily count the number of occurrences of each production rule, which in turn determines their MLE probabilities.

Rule	Total Count	MLE Probability
1.	1531	0.5734
2.	243	0.0910
3.	167	0.0625
4.	154	0.0576
5.	454	0.1700
6.	121	0.0453
7.	23	0.0663
8.	324	0.9337

(b) We can use the MLE estimates of rule probabilities to determine the predicted frequencies of each noun phrase type given G'. We can then use this information to conduct a goodness-of-fit test for G', as well as a model comparison of G and G'.

Noun phrase	Observed frequency	$L(o;G)$	G predicted frequency	$L(o;G')$	G' predicted frequency
N	1445	0.639	1555.6	0.573	1395.7
P	388	0.144	350.1	0.170	413.8
NN	231	0.047	113.7	0.091	221.5
AN	135	0.047	114.0	0.054	130.9
A	114	0.050	121.3	0.058	142.0
PN	31	0.011	25.6	0.004	10.8
NA	19	0.004	8.9	0.002	3.7
NNN	12	0.003	8.3	0.002	5.8
AA	10	0.0029	7.1	0.0039	9.4
NAN	8	0.0034	8.3	0.0014	3.4
AP*	6	0.0008	2.0	0.0005	1.1
PPN*	6	0.0002	0.4	0.00003	0.1
ANN	5	0.003	8.3	0.0014	3.4
AAN	4	0.0027	6.6	0.0036	8.7
PA*	4	0.0008	2.0	0.0005	1.1
ANA*	3	0.0003	0.7	0.0	0.0
APN*	3	0.00004	0.1	0.0	0.0
AAA*	2	0.00016	0.4	0.0025	0.6
APA*	2	0.0	0.0	0.0	0.0
NPP*	2	0.00016	0.4	0.00003	0.1
PAA*	2	0.00003	0.1	0.00003	0.1
PAN*	2	0.00078	1.9	0.00041	1.0
low-$\hat{\mu}$ (*)	32	0.0033	8.0	0.0017	4.1

(c) For goodness of fit, we conflate all cells with expected counts below 5 in G'. This gives us a X^2 of 223.5. If we include the residual mass of unobserved cell counts, which is a good idea for the X^2 test, we find that the expected count is 77.2, which brings our X^2 to 300.7.

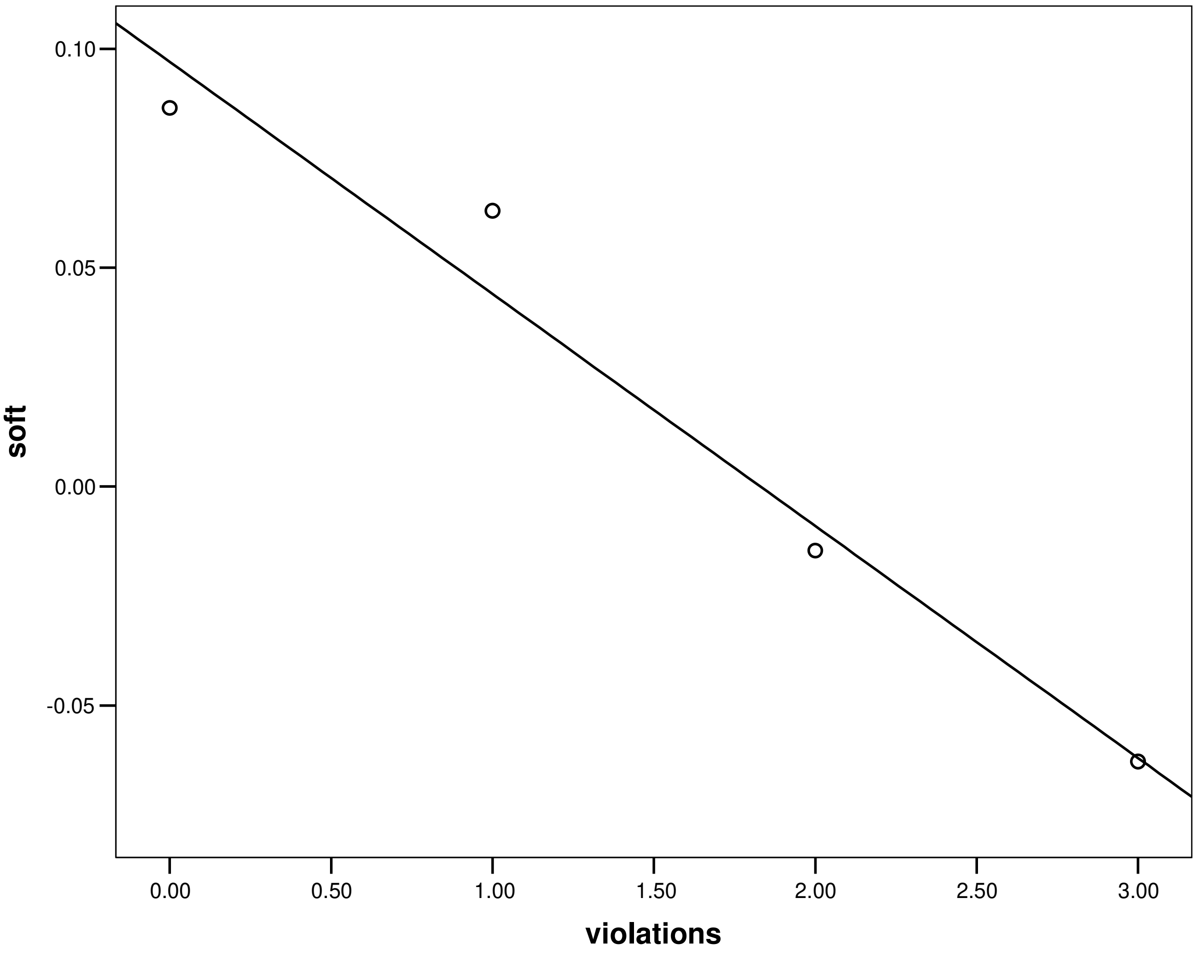
We can compare G and G' using a G^2 or X^2 test. For these tests, we'll group all the noun phrase types whose expected counts in the null-hypothesis grammar G are under 5; these are marked in the table with a *. With this grouping, it turns out that the log-likelihood of the data (as usual, ignoring the constant combinatorial factor) under G is -3513.54, and under G' it is -3495.79. One degree of freedom separates the two hypotheses (there are five free parameters in G and six in G'). So we look up the p-value of $-2 * (L(G) - L(G'))$ and find that $\chi^2(1; 35.48) \ll 0.001$. Therefore, G' is a much better fit to the data than G. Examining the difference in data likelihood for each cell, it turns out that by far the single biggest difference between G and G' is the likelihood of the NN data, consistent with our observation that G did an unsatisfactory job of generating NN noun phrases. We can also see, though, that G' has a new problem with undergenerating PN sequences; this arises from the reduced probability of the NP

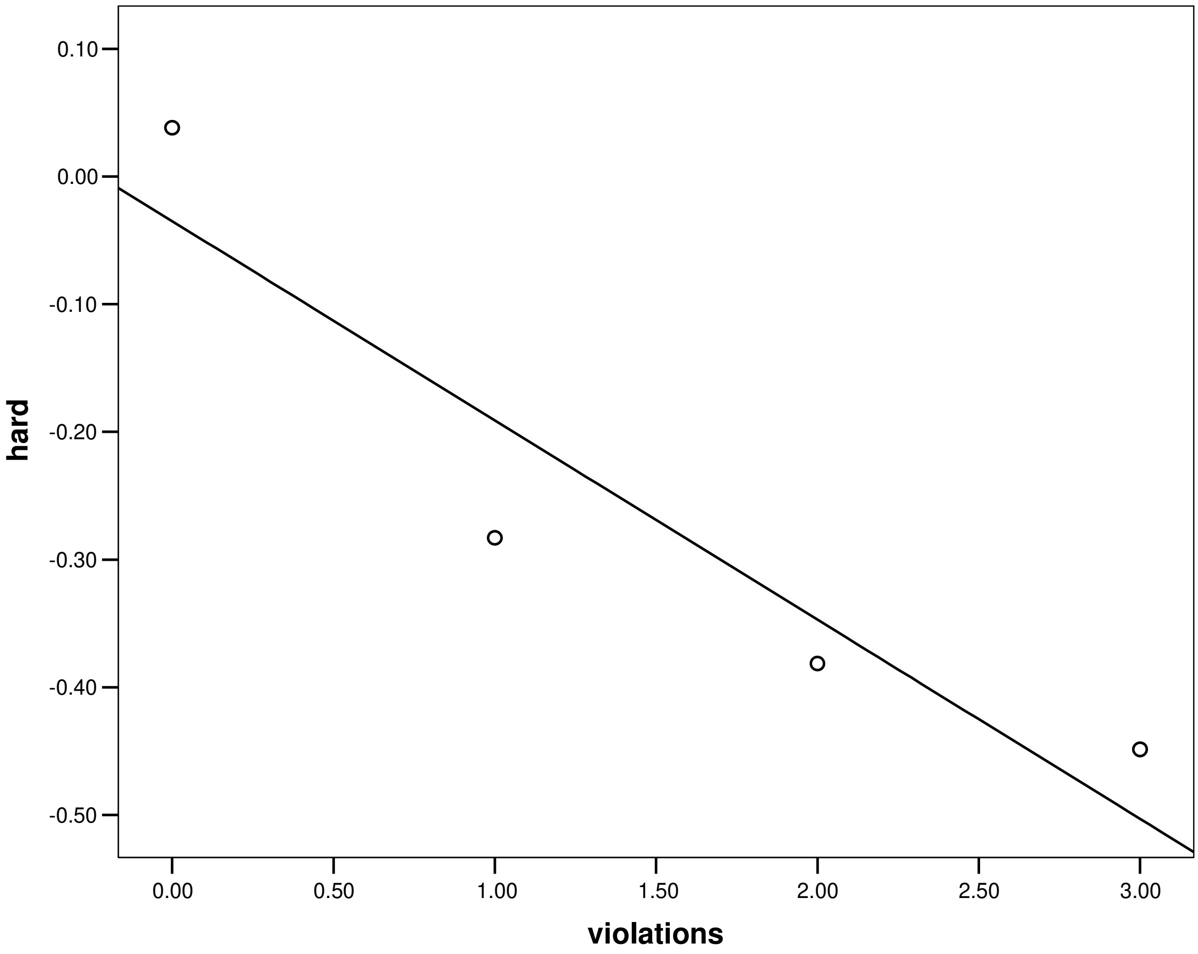
$\rightarrow$NP NP rule in G' (this rule had a higher probability in G, since it was required to generate the NN sequences). One possible further grammar refinement could be to add a new rule that allowed the Adj category to rewrite as a pronoun; perhaps sequences like *me hat* are really possessive uses that should be treated differently than free-standing pronoun uses. This would be an appropriate time to go back and look at the actual data, to see whether that idea seems plausible given the contextualized utterances.

(d) To calculate the probability mass allocated to unobserved strings, we can just subtract the mass allocated to observed strings from 1. The mass allocated for observed strings is 0.9603 for G and 0.9683 for G', leaving us with unobserved-string mass of 0.397 for G and 0.317 for G'. (Actually, these are overestimates of the unobserved-string masses, because for both G and G', some observed strings can be generated in more than one way.)

2. The soft constraint regression has an intercept of 0.097 and a slope of -0.053; the hard constraint regression has an intercept of -0.035 and a slope of -0.156. The scatterplots plus regression lines are appended.

   With only four datapoints it can be hard to evaluate the goodness of fit of a regression, but there is a clear difference between the soft and hard graphs. The linear regression seems to fit the soft constraint results fairly well, but not the hard constraint results.


0.10
0.05
0.00
-0.05
soft
0.00
0.50
1.00
1.50
2.00
2.50
3.00
violations


0.10
0.00
-0.10
-0.20
-0.30
-0.40
-0.50
hard
0.00
0.50
1.00
1.50
2.00
2.50
3.00
violations