SARA corpus statistics

Cross-references	Explicit		Implicit	
Within the section	30	*55*	25	*43*
To another section	34	*69*	44	*61*

Table 1: Number of subsections containing cross-references. Numbers in regular font correspond to the SARA statutes, and numbers in italics were computed from 9 sections taken at random from the IRC.

	min		max		avg		stddev		median	
Depth of leaf	1	*1*	6	*5*	3.6	*3.4*	0.8	*0.9*	4	*4*
Depth of node	0	*0*	6	*5*	3.2	*3.1*	1.0	*1.0*	3	*3*

Table 2: Statistics about the tree structure of the statutes. Numbers in regular font correspond to the SARA statutes, and numbers in italics were computed from 9 sections taken at random from the IRC.

Vocabulary		768	*1122*								
		min		max		avg		stddev		median	
Sentence length	(in words)	1	*1*	88	*121*	16.5	*22.0*	14.9	*21.7*	12.5	*17*
Section length	sentences	2	*1*	16	*30*	8.3	*10.1*	4.7	*9.1*	9	*6*
	words	62	*25*	1151	*2423*	488.9	*567.1*	310.4	*730.7*	549	*167*

Table 3: Language statistics. Numbers in regular font correspond to the SARA statutes, and numbers in italics were computed from 9 sections taken at random from the IRC.