ATR Japanese Corpus (Spoken Language DataBase)

		# of Morphemes		# of Characters	
ATR Corpus	**# of Sent.**	**Range**	**Ave.**	**Range**	**Ave.**
Training set	10,361	1-34	6.69	2-58	12.57
Test set	534	1-22	6.14	2-42	11.74

Corpus	APB
SUSANNE	1.256
SEC (Spoken English Corpus)	1.239
ATR (character-base)	1.341

Number of parses $= APB^n$

where n is the number of words in a sentence: Average Parse Base (APB)

Experimental Results

Models	2-42 Characters (534 sentences)						
	PA	LP	LR	BP	BR	0-CB	m-CB
B&C	89.33 (56.1%)	97.79	97.54	98.53	98.06	94.57 (72.4%)	0.11
Two-leval PCFG	62.55 (87.5%)	96.31	95.31	98.66	97.38	95.32 (67.9%)	0.09
PCFG	53.93 (89.8%)	95.64	94.48	98.77	97.31	94.76 (71.9%)	0.08
PGLR	95.32	99.06	98.47	99.53	98.73	98.50	0.03

Models	14-42 Characters (177 sentences)						
	PA	LP	LR	BP	BR	0-CB	m-CB
B&C	77.97 (61.6%)	96.54	97.66	97.20	98.35	87.01 (73.9%)	0.27
Two-level PCFG	57.63 (80.0%)	97.36	97.29	98.80	98.72	93.22 (50.0%)	0.18
PCFG	41.24 (85.6%)	95.64	95.37	98.71	98.47	91.53 (60.0%)	0.16
PGLR	91.53	99.09	99.10	99.53	99.55	96.61	0.07

LALR and CLR table-based PGLR

Models	2-42 Characters (534 sentences)						
	PA	LP	LR	BP	BR	0-CB	m-CB
PGLR(CLR)	95.13	99.04	98.40	99.46	98.61	97.57	0.04
PGLR(LALR)	95.32	99.06	98.47	99.53	98.73	98.50	0.03