構文統計について

木村睦子（（財）計量計画研究所）

1．概要

日本語文法の確率論的モデルを得るための第一歩として，①構文実態調査に基づく構文規則の作成，②各規則の使用頻度と階層上の推移確率とを主とする構文統計資料作成　等の作業を行った．

2．調査対象

(a) 中学校の理科の教科書一冊　　約2500文
(b) 特許公報より　　約 500文
(c) 刑法の判例より　　約 500文

(a)を基本として他を比較及び補完のために用いる．なお推移表作成の段階では，データを(a)の半分にしぼった．

3．作業手順

(1)構文規則作成

ア．例文カードと構文の木

テキストをカードに書き写し，二分割法により，構文の木を書く．構文要素としては自立語のみを採る．助動詞は原則として単項演算子とみなし，用言につけて一単位とする．ただし断定の助動詞のみは切り離す．助詞は二項演算子とみて切り離すが，treeの構成要素として扱わないといういささか変則的な方法を採る．

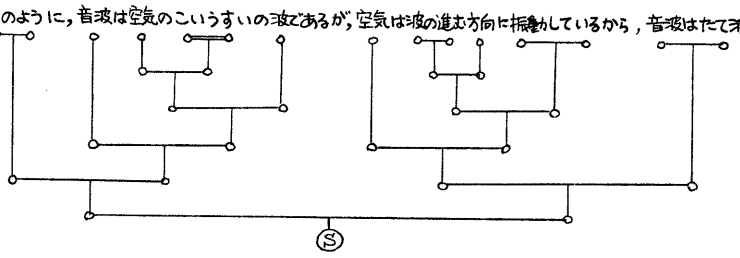
イ．コーディング

上記構文の木の枝の数を数えて括弧に置き換え，自立語を品詞に置き換え，助詞・助動詞はそのままの形でカタカナにしてコーディングする．ただし一番外側の括弧は省く．助詞や助動詞をどの位置に置くかの情報はtreeの中にはないから，コーダーがそのつど判定する．

例

（テキスト）　このように，音波は空気のこういうすいの波であるが，空気は波の進む方向に振動しているから，音波はたて波である．

（構文の木）



（コーディング結果）　（（AR N2）ニ，（N1 ハ（（N1 ノ〈AJ AJ〉）ノ N1）デアル））ガ，（（N1 ハ（（（N1 ノ V1）N1）ニ（V1 テ V2））））カラ，（N1 ハ N1 デアル））．

ウ．品詞KWIC作成と置き換え

(ア) 品詞列KWIC

(イ) 句構造の枚挙及び命名

末端の二つの語が結びついて出来た語，言い換えれば一番内側の括弧に囲まれた部分を枚挙し，頻度を調べる．前記の例で言えば，

AR N2	1
N1 ノ V1	1
V1 テ V2	1
N1 ハ N1 デアル	1
AJ AJ （並列）	1

となる．この調査はKWOCリストを用いて行う．得られた句にそれぞれ単一の名称を与える．

(ウ) 代入

データ中の上記連系部分を，括弧も含めて抹消し，新名称を代入する．これによりテキストの構造が一段浅くなったわけである．この部分は初めの3回まで計算機で，後は手作業で行った．

(イ),(ウ)をくり返すことにより，そのテキストの文構造を把握することができる．また内容的に不十分であるとはいえ，書き換え規則の集合が得られるわけであるから，いちおう構文規則が出来たと言っても差支えない．

(2)統計表作成

ア．例文カードへの規則番号の記入

イ．集計

各カードをnodeの数だけコピーし，着目箇所に印をつけて番号順にソートし，左推移表及び右推移表を作成した．出現した推移パターンの数は左右それぞれ1062，863，延べnode数は12,125（1文当り9.5）である．

4．調査結果

(1)文の長さ

資料a・b・cの文の長さの最大は，それぞれ32（文節），148，186である．また平均はデータ1（資料a）が約9.7，データ2（資料b・c）が26.6である．

(2)文の深さの分布

表1　深さの分布

資料 / 深さ	a		b		c	
	文数	%	文数	%	文数	%
0	28	(1.1)	—	—	1	(0.4)
1	114	(4.5)	7	(2.1)	3	(1.2)
2	323	(12.6)	12	(3.7)	5	(2.0)
3	513	(20.1)	29	(8.8)	2	(0.8)
4	593	(23.2)	27	(8.2)	12	(4.8)
5	452	(17.7)	38	(11.6)	7	(2.8)
6	317	(12.4)	49	(14.9)	18	(7.1)
7	148	(5.8)	35	(10.7)	21	(8.3)
8	48	(1.9)	34	(10.4)	29	(11.5)
9	13	(0.5)	31	(9.4)	30	(11.9)
10	5	(0.2)	20	(6.1)	26	(10.3)
11	—	—	14	(4.3)	22	(8.7)

12	—	—	7	(2.1)	18	(7.1)
13	—	—	4	(1.2)	11	(4.4)
14	1	(0.04)	8	(2.4)	12	(4.8)
15	—	—	8	(2.4)	7	(2.8)
16	—	—	3	(0.9)	12	(4.8)
17	—	—	1	(0.3)	6	(2.4)
18	—	—	1	(0.3)	5	(2.0)
19	—	—	1	(0.3)	1	(0.4)
20	—	—	—	—	1	(0.4)
21	—	—	—	—	3	(1.2)
計	2555		329		252	

(3)構文規則 (抄)

番号	規則
1	<文> ::= <動述語> [終止] <句点> |
2	<動述語> [命令] <句点> |
3	<形述語> [終止] <句点> |
4	<体述語> [終止] <句点> |
5	<力述語> <句点> |
6	<体連語> <句点>
10	<動述語> ::= ▽1 |
11	▽1 テ ▽2 |
13	<体連語> ▽1 |
14	<体連語> 'オ' <動述語> |
15	<副体連語> 'オ' <動述語> |
16	<体連語> 'ガ' <動述語> |
18	<体連語> <ニ助詞> <動述語> |
19	<副体連語> <ニ助詞> <動述語> |
20	<体連語> <デ助詞> <動述語> |
22	<体連語> <カラ助詞> <動述語> |
24	<体連語> <ト助詞> <動述語> |
27	<体連語> <係助詞> <動述語> |
28	<体連語> 'ノ' <動述語> |
29	<体連語> <ヘ助詞> <動述語> |
32	<体連語> <シテ助詞> <動述語> |
33	<副体連語> <動述語> |
34	<用修語1> <動述語> |
36	<形述語> 'ナル' |
47	<動述語> <引用結合子> <動述語> |
51	<動述語> <動述語> |
52	<形述語> <動述語> |
53	<体述語> <動述語> |
54	<動述語> <接続助詞> <動述語> |

番号	規則	
55		＜形述語＞＜接続助詞＞＜動述語＞|
56		＜体述語＞＜接続助詞＞＜動述語＞|
58		＜動述語＞‘テ’＜動述語＞|
61		＜動述語＞＜句末付加語＞|
62		＜接続詞＞＜動述語＞
80	＜形述語＞ ::=	AJ|
81		＜動述語＞‘タイ／ニクイ／ヤスイ’|
82		＜体連語＞‘ガ’＜形述語＞|
88		＜体連語＞＜二助詞＞＜形述語＞|
90		＜体連語＞‘ノ’＜形述語＞|
91		＜体連語＞＜係助詞＞＜形述語＞|
93		＜用修語１＞＜形述語＞|
100		＜動述語＞＜接続助詞＞＜形述語＞|
104		＜接続詞＞＜形述語＞
120	＜体述語＞ ::=	＜体連語＞＜断定陳述辞＞|
121		＜副体連語＞＜断定陳述辞＞|
128		＜体連語＞‘ガ’＜体述語＞|
130		＜体連語＞＜係助詞＞＜体述語＞|
148		＜動述語＞‘カラ’＜断定陳述辞＞
160	＜カ述語＞ ::=	＜体連語＞‘カ／ナノカ’|
161		＜副体連語＞‘カ／ナノカ’|
164		＜動述語＞‘カ／ノカ’|
171		＜体連語＞＜係助詞＞＜カ述語＞|
172		＜体連語＞‘トハ’＜カ述語＞
200	＜体連語＞ ::=	N1|
201		N4|
202		N7|
203		SY|
204		AR＜体連語＞|
206		＜体連語＞SY|
208		＜用修語１＞N1|
211		＜体連語＞‘ノ’＜体連語＞|
212		＜副体連語＞‘ノ’＜体連語＞|
214		＜体連語＞‘トイウ’＜体連語＞|
217		＜動述語＞＜体連語＞|
218		＜形述語＞＜体連語＞|
219		＜体述語＞＜体連語＞|

番号	規　　則
222	＜体連語の並列＞
230	＜副体連語＞ ::= N3 |
231	N6 |
232	AR N2 |
233	＜体連語＞ ‘ノ’ N2 |
236	＜動述語＞ N2 |
256	＜体連語＞ N3 |
260	＜体連語＞＜副助詞＞|
264	＜力述語＞|
268	＜形述語＞＜副助詞＞
295	＜形動語幹＞ ::= N5
310	＜用修語1＞ ::= AD|
311	＜形動語幹＞ ‘ニ’|
313	＜副体連語＞＜係助詞＞

(4)構文統計表（抄）　（頻度10以上のもののみ）

左推移表

上位規則	下位規則	頻度	推移確率	上位規則	下位規則	頻度	推移確率
S	14	96	.0751				
	18	39	.0305				
	19	31	.0243	14	200	468	.4247
	20	38	.0297		201	26	.0236
	22	19	.0149		204	61	.0554
	27	138	.1080		211	234	.2123
	32	31	.0243		212	47	.0426
	33	57	.0446		217	114	.1034
	34	41	.0321		218	48	.0436
	51	144	.1127		219	20	.0181
	54	125	.0978		222	64	.0581
	56	11	.0086			(1102)	
	58	64	.0501				
	62	86	.0673	15	256	12	.3077
	93	12	.0094		264	10	.2564
	104	13	.0102			(39)	
	130	63	.0493				
	164	72	.0563				
	171	16	.0125				
		(1278)					

上位規則	下位規則	頻度	推移確率	上位規則	下位規則	頻度	推移確率
16	200	286	.4319	28	200	38	.7917
	204	37	.0536			(48)	
	211	97	.1406	32	200	52	.3688
	212	21	.0304		204	10	.0709
	217	133	.1928		211	45	.3191
	218	22	.0319		222	12	.0851
	219	23	.0333			(141)	
	222	39	.0565	33	230	55	.2273
		(690)			231	59	.2438
18	200	372	.4850		232	30	.1240
	204	20	.0261		236	27	.1116
	211	199	.2595		260	13	.0537
	212	32	.0417		264	19	.0785
	217	47	.0613			(242)	
	218	19	.0248	34	310	243	.7147
	222	40	.0522		311	44	.1294
		(767)			313	40	.1176
19	230	22	.0873			(340)	
	231	34	.1349	36	80	63	.8077
	232	13	.0516			(78)	
	233	57	.2262	51	14	77	.2500
	236	92	.3651		16	17	.0552
		(252)			18	61	.1981
20	200	151	.5572		20	14	.0455
	204	14	.0517		27	29	.0942
	211	42	.1550		51	14	.0455
	212	10	.0369		58	41	.1331
	222	20	.0738			(308)	
		(271)		52	80	65	.7303
22	200	73	.6186			(89)	
	211	25	.2119	54	14	48	.1690
		(118)			16	43	.1514
24	200	91	.6319		18	40	.1408
	211	21	.1458		19	18	.0634
		(144)			20	13	.0458
27	200	190	.4460		27	28	.0986
	204	18	.0423		34	13	.0458
	211	94	.2207		58	19	.0669
	217	51	.1197			(284)	
	222	32	.0751	55	82	10	.5000
		(426)				(20)	
				58	10	30	.0761
					14	129	.3274

上位規則	下位規則	頻度	推移確率	上位規則	下位規則	頻度	推移確率
(58)	16	32	.0812	(164)	54	10	.1031
	18	96	.2437		61	11	.1134
	19	10	.0254		62	11	.1134
	22	10	.0254			(97)	
	24	12	.0305	171	211	12	.5000
	34	19	.0482			(24)	
	51	15	.0381				
		(394)					
81	10	20	.7407	206	200	28	.6087
		(27)				(46)	
82	200	50	.5208	208	310	13	1.0000
	211	20	.2083			(13)	
	217	16	.1667	211	200	798	.6656
		(96)			201	14	.0117
88	200	10	.5000		203	17	.0142
		(20)			204	33	.0275
90	200	14	.7778		211	99	.0826
		(18)			212	21	.0175
91	200	19	.4130		217	69	.0575
	211	12	.2609		218	10	.0083
		(46)			219	10	.0083
120	200	34	.2982		222	100	.0834
	211	15	.1315			(1199)	
	217	45	.3947	212	230	93	.4170
		(114)			231	65	.2915
121	230	11	.1392		236	19	.0852
	233	33	.4177		256	10	.0448
	236	15	.1899			(223)	
		(79)		214	200	15	.6818
122	295	47	.8103			(22)	
		(58)		217	10	65	.0985
128	200	11	.4231		11	12	.0182
		(26)			14	128	.1939
130	200	53	.4732		16	56	.0848
	211	26	.2321		18	104	.1576
	217	11	.0982		19	33	.0500
		(112)			20	31	.0470
160	202	11	.5238		22	24	.0364
		(21)			24	14	.0212
161	230	17	.7391		28	38	.0576
		(23)			32	11	.0167
164	16	12	.1237		33	14	.0212
	27	31	.3196		34	19	.0288

上位規則	下位規則	頻度	推移確率	上位規則	下位規則	頻度	推移確率
(217)	51	16	.0242		16	52	.2694
	54	11	.0167		18	20	.1036
	58	38	.0576			(193)	
		(660)		237	82	11	.3929
218	80	186	.8378			(28)	
	90	11	.0495	256	200	33	.8049
		(222)				(41)	
219	121	46	.4220	264	164	21	.6563
	122	53	.4862			(32)	
		(109)		268	82	12	.8571
233	200	82	.6891			(14)	
	211	16	.1345	313	231	20	.3571
	222	14	.1176		236	15	.2679
		(119)				(56)	
236	14	43	.2228				
93	310	41	.7593				
	313	13	.2407				
		(54)					

右推移表

上位規則	下位規則	頻度	推移確率	上位規則	下位規則	頻度	推移確率
14	10	668	.6062		52	10	.0145
	11	112	.1016		58	23	.0333
	18	61	.0554			(690)	
	20	49	.0445	18	10	423	.5515
	24	70	.0635		11	69	.0900
	33	49	.0445		14	107	.1395
	34	28	.0254		16	107	.1395
	52	14	.0127		33	11	.0143
	58	22	.0200			(767)	
		(1102)		19	10	103	.4087
15	10	25	.6410		11	13	.0516
		(39)			14	32	.1270
16	10	406	.5884		16	25	.0992
	11	64	.0928		18	14	.0556
	14	20	.0290		27	10	.0397
	18	31	.0449		34	10	.0397
	19	11	.0159		58	11	.0437
	32	16	.0232			(252)	
	33	13	.0188	20	10	121	.4465
	34	35	.0507		11	14	.0517
	36	33	.0478		14	33	.1218

上位規則	下位規則	頻度	推移確率	上位規則	下位規則	頻度	推移確率
(20)	16	36	.1328	(33)	18	15	.0620
	18	10	.0369		20	10	.0413
	34	11	.0406		27	14	.0579
	51	10	.0369		54	13	.0537
		(271)				(242)	
22	10	33	.2797	34	10	127	.3735
	14	17	.1441		11	30	.0882
	16	27	.2288		14	22	.0647
	18	13	.1102		16	25	.0735
		(118)			18	27	.0794
24	10	125	.8681		19	13	.0382
	11	11	.0764		27	16	.0471
		(144)			36	10	.0294
27	10	42	.0986		51	10	.0294
	14	19	.0446		54	15	.0441
	16	19	.0446			(340)	
	18	53	.1244	47	10	25	.9615
	19	32	.0751			(26)	
	20	22	.0516	51	14	86	.2792
	24	10	.0235		16	32	.1039
	32	24	.0563		18	34	.1104
	33	17	.0399		19	12	.0390
	34	36	.0845		20	10	.0325
	36	13	.0305		22	10	.0325
	51	18	.0423		27	21	.0682
	53	12	.0282		32	10	.0325
	54	23	.0540		34	13	.0422
	58	38	.0892		58	39	.1266
		(426)				(308)	
28	10	42	.8750	52	10	52	.5843
		(48)			11	12	.1348
29	10	11	.6111			(89)	
		(18)		54	10	11	.0387
32	10	52	.3688		14	10	.0352
	11	14	.0993		16	54	.1901
	14	20	.1418		18	18	.0634
	16	25	.1773		19	14	.0493
	18	12	.0851		27	51	.1796
		(141)			33	11	.0387
33	10	73	.3017		34	16	.0563
	11	19	.0785		51	19	.0669
	14	24	.0992		54	22	.0775
	16	20	.0826		58	21	.0739

上位規則	下位規則	頻度	推移確率	上位規則	下位規則	頻度	推移確率
(54)		(280)		208	200	10	.7692
58	10	89	.2259			(13)	
	11	17	.0431				
	14	110	.2792	211	200	1004	.8374
	16	28	.0711		201	73	.0609
	18	55	.1396		211	20	.0167
	19	16	.0406		212	11	.0092
	33	10	.0254		218	22	.0183
	34	14	.0355		219	11	.0092
		(394)			222	24	.0200
62	14	10	.1020			(1199)	
	27	20	.2041	212	200	190	.8559
	54	12	.1224		211	10	.0450
		(98)				(222)	
82	80	82	.8913	214	200	13	.5909
		(96)				(22)	
88	80	10	.5000	217	200	569	.8634
		(20)			201	15	.0228
90	80	16	.8889		211	18	.0273
		(18)			212	15	.0228
91	80	17	.3696		218	14	.0212
		(46)			222	11	.0167
93	80	27	.5000			(659)	
		(54)		218	200	207	.9324
100	80	16	.5517			(222)	
		(29)		219	200	90	.8411
128	120	16	.6154			(107)	
		(26)		245	230	10	.5263
130	120	61	.5446			(19)	
	121	12	.1071				
	148	15	.1339				
		(112)					
171	161	13	.5417				
		(24)					
172	160	11	1.0000				
		(11)					
204	200	204	.8000				
	201	26	.1020				
	211	12	.0471				
		(255)					