学術論文標題に出現する多字種複合語に対する字種連接特性の分析

齋藤 恵[†]　熊澤 侑美[†]　後藤 智範[‡]

[†]神奈川大学大学院理学研究科　[‡]神奈川大学理学部情報科学科

1 はじめに

研究開発の活性化に伴い，新しいモデル・理論を表す新しい用語が出現する．外国語由来の語はカタカナ，場合によってはアルファベット表記がそのまま日本語の文書で使用される．近年，この傾向は非常に顕著で複数の字種で表記される用語が著しく増加している．

このような多字種語の増加を受け，当研究室ではコーパスとして辞書見出し語，特許抄録，学術論文標題，学術論文抄録中に出現する大量の多字種語データについて，字種特性，具体的には，構成字種，字種連接パターンの観点から調査・分析を行ってきた[1][2][3][4]．本研究は，[1]と同様のコーパスを使用し，字種連接の特性について調査・分析した結果について報告する．

表 3.1　字種変化数毎のパターン数・用語数

変化数	パターン数	比率	用語頻度	比率
2	32	0.66%	69,342	50.57%
3	132	2.71%	36,917	26.92%
4	361	7.42%	14,992	10.93%
5	685	14.09%	7,545	5.50%
6	933	19.19%	3,580	2.61%
7	874	17.97%	1,914	1.40%
8	646	13.28%	1,119	0.82%
9	437	8.99%	617	0.45%
10	254	5.22%	350	0.26%
11	181	3.72%	344	0.25%
12	108	2.22%	134	0.10%
13	69	1.42%	73	0.05%
14	44	0.90%	64	0.05%
15	32	0.66%	36	0.03%
16	25	0.51%	32	0.02%
17	14	0.29%	16	0.01%
18	9	0.19%	9	0.01%
19	8	0.16%	9	0.01%
20	4	0.08%	4	0.00%
21	5	0.10%	5	0.00%
22	3	0.06%	3	0.00%
23	3	0.06%	1	0.00%
24	1	0.02%	1	0.00%
25	3	0.06%	3	0.00%
計	4,863	100%	137,112	100%

2 コーパス・解析項目

2.1 コーパス

本報告では，2013 年 3 月の NLP 報告と同様の用語集合，137,112 語を解析対象とした[5]．

2.2 解析項目

本研究では，字種連接（字種連接数，字種連接パターン）についてより詳細に調査・分析する．具体的には，次の項目について明らかにする．

1. 連接数毎の用語総数と字種連接パターンの種類
2. 先頭字種毎の字種連接パターン
3. 字種連接パターンを構成する字種の使用頻度

字種名の表現として以下に挙げる字種記号を用い，字種の連接を字種記号の記号列として扱う．

漢字：J　カタカナ：K　ひらがな：H
全角英字：A　半角英字：a　全角数字：N
半角数字：n　全角記号：S　半角記号：S

3 結果

3.1 字種変化数毎のパターン数・用語数

字種変化数毎のパターン数および用語数を表 3.1 に示す．変化数 4～8 で出現したパターン数は全体の約 75%を占めており，出現したパターンの半分以上が変化数 4～8 であることが分かる．また，変化数 2～4 で出現用語数は全体の約 89%を占めており，出現した用語の 9 割弱は変化数 2～4 であることが分かる．

3.2 先頭字種毎の字種変化パターン数・用語数

表 3.2 に先頭字種毎の字種変化パターン数・用語数を示す．出現パターン数は半角英字，漢字，半角数字で約 78%を占めており，多字種複合語のパターンは半角英字，漢字，半角数字のいずれかで始まるものが多いことが分かる．また用語数は漢字とカタカナで約 75%を占めており，多字種複合語は漢字またはカタカナで始まるものが多いことが分かる．

3.3 変化数毎の出現パターン数

変化数毎の出現パターン数を表 3.3 と表 3.4 に示す．表 3.3 には先頭字種が非日本語，表 3.4 には先頭字種が日本語のものを記載した．

表 3.2　先頭字種毎の字種変化パターン数・用語数

先頭字種	パターン数	比率	用語数	比率
a	1,311	26.96%	23,975	17.49%
A	0	0%	0	0%
H	34	0.70%	737	0.54%
J	1,386	28.50%	61,672	44.98%
K	846	17.40%	41,716	30.42%
n	1,075	22.11%	8,010	5.84%
N	0	0%	0	0%
s	29	0.60%	38	0.03%
S	182	3.74%	964	0.70%
計	4,863	100%	137,112	100%

表 3.3　変化数毎の出現パターン数（先頭字種非日本語）

変化数	a	n	s	S
2	6	5	1	5
3	26	20	4	17
4	83	65	2	29
5	166	121	4	33
6	236	152	4	35
7	210	186	2	26
8	174	162	1	18
9	125	126	3	8
10	80	73	5	4
11	56	63	1	4
12	36	39	0	1
13	27	23	1	1
14	18	13	1	0
15	19	7	0	1
16	13	10	0	0
17	10	3	0	0
18	6	2	0	0
19	5	2	0	0
20	3	1	0	0
21	3	1	0	0
22	2	1	0	0
23	3	0	0	0
24	1	0	0	0
25	3	0	0	0
計	1,311	1,075	29	182

非日本語では，先頭字種 a および n が変化数 5〜9 でそれぞれ 100 パターン以上出現した．先頭字種によってはパターンが出現しなかった変化数もある．

日本語では，先頭字種 J が変化数 4〜9，K が変化数 5〜8 でそれぞれ 100 パターン以上出現している．変化数 20 ではパターンが存在しなかった．

表 3.4　変化数毎の出現パターン数（先頭字種日本語）

変化数	H	J	K
2	3	6	6
3	6	31	28
4	9	104	69
5	7	201	153
6	5	291	210
7	1	288	161
8	0	188	103
9	2	119	54
10	1	66	25
11	0	39	18
12	0	23	9
13	0	10	7
14	0	11	1
15	0	4	1
16	0	2	0
17	0	1	0
18	0	1	0
19	0	0	1
21	0	1	0
計	34	1,386	846

4　考察

表 4.1 から表 4.3 は，それぞれ先頭字種が漢字，カタカナ，半角アルファベットで，字種連接パターンの特定位置（行）に特定字種（列）が存在するパターンの種類を示している[6]．具体的には，各表において字種を示す列について，左右 2 つの値が記載されているが，左の値はその位置（各行の最左の数値）にある字種のパターンの種類数，右側の値は，その位置で終了するパターンの種類数を示している．例えば，表 4.1 において，「K」（カタカナ）列の 2 行目の左側の値（415）は，先頭が漢字（J）で始まる全パターンのうち，415 パターンが 2 番目の位置に K がくることを示している，言い換えれば「JK」で始まるパターンが 415 あることを示している．また，同表で，8 行（最左の値）目で「a」列の，右側の値（25）は，J で始まる連接パターン長が 8 で，末尾が半角アルファベットであるパターンが 25 種類あることを示している．

最右行の数値は，左側が 7 種類の字種が特定位置（各行の最左の値）にある総パターン数，右側は特定の長さの総パターン数を示している．最下行で，左側は位置を考慮しない当該列の字種を含む総パターン数，右側は末尾が当該列の字種となる総パターン数を示している．前者の値は，当該字種の使用頻

表 4.1　パターン長毎の字種使用頻度（先頭字種：J）

	J		K		H		a		n		s		S		総計	
2	0	0	415	1	161	1	349	1	253	1	144	1	64	1	1386	6
3	571	6	108	5	11	3	197	5	123	5	324	4	46	3	1380	31
4	170	25	214	19	111	9	323	22	275	14	215	10	41	5	1349	104
5	375	74	175	53	25	8	253	38	123	16	252	10	42	2	1245	201
6	222	110	195	81	45	7	263	62	157	24	147	4	15	3	1044	291
7	223	134	153	92	13	2	145	38	77	19	118	2	24	1	753	288
8	115	72	93	59	11	1	101	27	70	29	71	0	4	0	465	188
9	90	52	47	36	3	1	59	21	31	7	40	2	7	0	277	119
10	35	28	35	18	1	0	51	9	17	11	17	0	2	0	158	66
11	35	22	10	6	1	0	16	6	14	5	15	0	1	0	92	39
12	9	7	11	8	0	0	18	6	7	2	8	0	0	0	53	23
13	7	3	5	4	0	0	3	1	6	2	9	0	0	0	13	7
14	2	1	4	4	1	0	10	4	2	2	1	0	0	0	14	2
15	2	1	2	2	0	0	0	0	2	1	3	0	0	0	9	4
計	1,859	537	1,469	390	383	32	1,792	241	1,159	138	1,366	33	246	15	8,274	1,386

表 4.2　パターン長毎の字種使用頻度（先頭字種：K）

	J		K		H		a		n		s		S		総計	
2	447	1	0	0	15	1	165	1	94	1	90	1	35	1	846	6
3	90	5	204	6	62	3	132	4	123	4	189	3	40	3	840	28
4	196	21	135	16	16	5	164	13	113	9	157	4	31	1	812	69
5	167	53	139	41	29	4	158	31	95	14	130	5	25	5	743	153
6	147	84	147	69	11	1	125	37	63	17	78	1	19	1	590	210
7	123	79	86	48	8	2	75	26	39	6	43	0	6	0	380	161
8	64	42	60	37	3	0	42	19	18	4	29	0	3	1	219	103
9	33	26	33	17	0	0	23	10	15	1	12	0	0	0	116	54
10	20	13	8	5	0	0	14	7	3	0	17	0	0	0	62	25
11	9	6	10	5	0	0	12	6	4	1	1	0	1	0	37	18
12	8	6	3	3	0	0	3	0	0	0	5	0	0	0	19	9
13	1	1	4	4	0	0	3	2	1	0	1	0	0	0	10	7
計	1,306	337	832	254	144	16	917	156	570	57	754	14	160	12	4,683	846

度を，また前者に対する後者の比の大きい字種は，複合語での主辞となる役割が高く，小さい字種は役割が低いことを示唆している可能性がある．

4.1　漢字（J）

漢字から始まり末尾が半角アルファベットであるパターンは 240 パターン以上あった．これは，漢字から始まりパターン中に半角アルファベットを含むパターンの約 13%，漢字から始まるパターンの約 17%に相当する．同じく漢字から始まり末尾も漢字であるパターンは 500 パターン以上存在し，これはパターン中に漢字を含むパターンの約 29%，漢字から始まるパターンの約 39%に相当する．半角数字を含むパターンは 1100 パターン以上存在しているが，半角数字が末尾であるパターンは 130 パターン程しかなく比率は約 12%であり，パターン中には多く含まれているが末尾であることはそれほど多くないと考えられる．

4.2　カタカナ（K）

カタカナから始まりパターン中に漢字を含むパターンはおよそ 1300 パターン存在した．その中で末尾が漢字であるパターンは 330 パターン以上で約 26%に相当する．カタカナから始まるパターン中では 40%に相当する．パターン中に半角アルファベットを含むパターンは 900 パターン以上存在し，そのうち半角アルファベットが末尾であるものはおよそ 156 パターンと約 17%であった．カタカナから始ま

表 4.3　パターン長毎の字種使用頻度（先頭字種：a）

J		K		H		a		n		s		S		総計	
2	264	1	148	1	7	1	0	0	153	1	701	1	38	1	1311	6
3	143	5	152	4	33	2	453	6	386	4	89	2	49	3	1305	26
4	265	20	146	19	23	5	298	18	113	12	396	4	38	5	1279	83
5	192	59	164	41	35	7	345	37	212	21	232	1	16	0	1196	166
6	230	96	162	71	13	2	195	41	156	21	248	2	26	3	1030	236
7	165	90	115	60	12	1	216	42	122	16	153	1	11	0	794	210
8	131	78	90	53	7	3	150	25	64	12	140	2	2	1	584	174
9	69	44	64	43	3	0	107	28	64	9	102	0	1	1	410	125
10	56	38	37	22	0	0	63	13	62	7	63	0	4	0	285	80
11	39	27	21	15	1	0	62	10	33	4	48	0	1	0	205	56
12	29	16	19	13	0	0	28	4	24	2	46	1	3	0	149	36
13	19	11	13	9	0	0	31	3	30	3	19	0	1	1	113	27
14	14	8	11	7	0	0	13	2	13	1	35	0	0	0	86	18
15	13	8	7	6	0	0	21	4	20	1	7	0	0	0	68	19
16	12	6	6	5	0	0	7	2	6	0	18	0	0	0	49	13
17	2	1	6	5	0	0	8	2	13	2	7	0	0	0	36	10
計	1,400	519	1,174	383	134	21	2,016	240	1,490	119	2,328	14	190	15	8,732	1,311

るパターン中では約 18%であった.

4.3　半角英字（a）

半角アルファベットから始まりパターン中に半角数字を含むパターンは 1450 パターン以上存在し, そのうち末尾が半角数字であるものは 110 パターン程度で 8%程しか末尾ではなかった．パターン中に半角記号を含むパターンは 2300 パターン以上存在したがそのうち末尾であったものはおよそ 10 パターンで，0.1%程しかない．半角アルファベットを含むパターンは 2000 パターン以上存在しそのうち末尾であるものはおよそ 240 パターン，比率は約 12%であった.

パターン中に漢字を含むパターンは 1400 パターン以上存在し，そのうち末尾であったものはおよそ 510 パターンであり，約 37%に相当する．半角アルファベットで始まるパターン中では約 40%であった. パターン中にカタカナを含むパターンは 1170 パターン以上でそのうち末尾であるものは 380 パターン以上で比率は約 33%であった.

5　終わりに

2.2 節に挙げた調査・分析項目について，3・4 章の表から字種連接パターンを構成する個々の字種が, 頻度，位置について字種固有の著しい特性をもっていることが明になった．字種連接についての研究結果の応用については過去にも専門用語の候補語の選定，chunking に言及したが，字種の観点からの用語検索が挙げられよう.

註・参考文献

[1] 田代征嗣, 滝川諒, 後藤智範. 学術論文標題に出現する多字種複合語に対する字種特性の解析. 第 18 回言語処理学会年次大会(NLP2012). 2012 年 3 月.

[2] 田代征嗣, 滝川諒, 後藤智範. 学術論文抄録に出現する多字種複合語に対する字種特性の解析. 第 18 回言語処理学会年次大会(NLP2012). 2012 年 3 月.

[3] 熊澤侑美, 斎藤恵, 後藤智範. 辞書見出し語中の複合語を対象とした字種特性の分析 -自然言語処理研究会報告 2013-NL-214(17), 1-6, 2013-11-15

[4] 熊澤侑美, 後藤智範. 特許抄録中に出現する多字種複合語を対象とした字種特性の分析 -自然言語処理研究会報告 2014-NL-217(16), 1-7, 2014-7-3

[5] 2013 年 3 月の NLP 報告では 140014 語であったがさらに詳細にスクリーニングした結果, 2902 語が不適切であると判明しこれらの語を除外した.

[6] 紙面の都合により，表 4.1 は最大字種連接パターン長 21 で，15 まで，同様に，表 4.2 は 19 で，13 まで，表 4.3 は 25 で，17 まで，掲載してある.