学術論文抄録に出現する多字種複合語に対する字種連接特性の分析

熊澤 侑美† 齋藤 恵† 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，字種構成，字種変化パターン観点から調査・分析を行ってきた[1][2][3][4]．本研究は，(2)と同様のコーパスを使用し，字種変化の特性について，調査・分析した結果について報告する．

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から 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）[6]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化パターン
3. 字種連接化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次に挙げる字種記号を用い，字種の変化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 1.1 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は変化数 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．

表 1.1 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	用語数	比率
2	31	0.71%	62,861	50.97%
3	145	3.31%	33,518	27.18%
4	449	10.24%	13,030	10.57%
5	815	18.58%	7,282	5.90%
6	882	20.11%	2,970	2.41%
7	660	15.05%	1,506	1.22%
8	469	10.69%	844	0.68%
9	297	6.77%	527	0.43%
10	203	4.63%	284	0.23%
11	147	3.35%	189	0.15%
12	89	2.03%	106	0.09%
13	57	1.30%	58	0.05%
14	42	0.96%	52	0.04%
15	27	0.62%	28	0.02%
16	18	0.41%	19	0.02%
17	14	0.32%	14	0.01%
18	10	0.23%	10	0.01%
19	6	0.14%	6	0.00%
20	4	0.09%	4	0.00%
21	5	0.11%	5	0.00%
22	3	0.07%	3	0.00%
23	1	0.02%	1	0.00%
24	3	0.07%	3	0.00%
25	4	0.09%	4	0.00%
26	1	0.02%	1	0.00%
27	1	0.02%	1	0.00%
31	1	0.02%	1	0.00%
33	1	0.02%	1	0.00%
55	1	0.02%	1	0.00%
計	4,386	100%	123,329	100%

表 3.1　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	922	21.02%	16,227	13.16%
A	0	0%	0	0%
H	30	0.68%	426	0.35%
J	1,408	32.10%	56,966	46.19%
K	708	16.14%	36,081	29.26%
n	966	22.02%	11,430	9.27%
N	0	0%	0	0%
s	51	1.16%	179	0.15%
S	301	6.86%	2,020	1.64%
計	4,386	100%	123,329	100%

3.2　先頭字種毎の字種変化パターン数・用語数

表 3.1 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 75%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．

表 3.2　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	24	6	19
4	86	83	7	48
5	151	159	11	60
6	154	181	7	57
7	133	132	5	38
8	106	110	4	35
9	77	83	4	13
10	45	52	3	11
11	44	32	2	7
12	32	25	0	4
13	17	23	0	2
14	15	13	1	2
15	6	13	0	0
16	8	7	0	0
17	3	9	0	0
18	4	4	0	0
19	3	1	0	0
20	1	1	0	0
21	2	2	0	0
22	2	1	0	0
23	0	1	0	0
24	0	2	0	0
25	1	1	0	0
26	0	1	0	0
33	1	0	0	0
55	0	1	0	0
計	922	966	51	301

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 3.2 と表 3.3 に示す．表 3.2 には先頭字種が非日本語，表 3.3 には先頭字種が日本語のものを記載した．

非日本語では，先頭字種 a および n が変化数 5〜8 でそれぞれ 100 パターン以上出現している．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．変化数 27〜32，34〜54 ではパターンが存在しなかった．

日本語では，先頭字種 J が変化数 4〜8，K が変化数 5〜7 でそれぞれ 100 パターン以上出現している．変化数 22，23，26，28〜30 ではパターンが存在しなかった．

表 3.3　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	2	6	6
3	8	35	27
4	10	130	85
5	5	273	156
6	4	315	164
7	0	246	106
8	0	152	62
9	0	81	39
10	0	64	28
11	1	45	16
12	0	23	6
13	0	11	4
14	0	9	2
15	0	5	3
16	0	1	2
17	0	1	1
18	0	2	0
19	0	2	0
20	0	2	0
21	0	1	0
24	0	1	0
25	0	2	0
27	0	0	1
31	0	1	0
計	30	1,408	708

表 4.2　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	284	1	250	1	270	1	370	1	123	1	111	1	1,408	6
3	450	6	74	5	17	4	183	5	193	5	347	5	138	5	1,402	35
4	143	30	130	24	96	8	237	24	411	18	246	14	104	12	1,367	130
5	259	84	113	49	23	9	234	61	213	36	281	16	114	18	1,237	273
6	163	106	88	52	29	8	187	71	280	43	148	19	69	16	964	315
7	136	88	51	34	7	2	129	65	122	29	150	14	54	14	649	246
8	53	35	29	17	6	1	94	62	125	27	77	5	19	5	403	152
9	36	22	12	8	3	0	60	26	60	18	66	5	14	2	251	81
10	20	16	9	4	2	0	38	23	52	12	41	7	8	2	170	64
11	13	11	1	1	0	0	28	19	26	9	30	5	8	0	106	45
12	10	7	5	4	0	0	19	11	15	1	9	0	3	0	61	23
13	2	1	2	2	1	0	9	5	7	3	16	0	1	0	38	11
14	1	1	1	1	0	0	9	4	10	1	4	2	2	0	27	9
計	1,293	411	804	205	434	33	1,519	384	1,907	206	1,577	94	645	75	8,179	1,408

表 3.1　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	364	1	0	0	12	1	124	1	130	1	44	1	34	1	708	6
3	56	5	111	6	51	1	134	5	133	4	163	4	54	2	702	27
4	132	19	79	17	9	4	120	19	139	10	149	8	47	8	675	85
5	86	46	91	39	22	4	111	32	130	17	110	10	40	8	590	156
6	95	56	71	42	3	3	86	33	74	22	74	3	31	5	434	164
7	39	31	42	23	3	0	60	31	60	14	54	4	12	3	270	106
8	26	16	19	15	3	1	27	16	42	12	40	2	7	0	164	62
9	14	8	16	10	0	0	24	13	25	7	18	1	5	0	102	39
10	12	10	7	4	1	1	13	8	12	3	15	1	3	1	63	28
11	3	3	4	2	0	0	13	8	7	2	6	1	2	0	35	16
12	1	0	1	0	0	0	3	2	6	3	7	0	1	1	19	6
13	2	0	1	1	0	0	3	1	4	1	3	1	0	0	13	4
計	831	**196**	446	161	104	15	726	**173**	770	98	694	36	236	29	3,807	708

4　考察

表 4.2 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）か存在するパターンの種類を示している[7]．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 4.2 において，「K」（カタカナ）列の 2 行目の左側の値（284）は，先頭が漢字（J）で始まる全パターンのうち，284 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 284 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（62）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 62 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．

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 380 パターン以上あった．これは、漢字

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	139	1	59	1	4	1	0	0	155	1	508	1	57	1	922	6
3	74	5	79	4	16	2	289	5	319	4	82	2	57	4	916	26
4	128	21	65	18	13	4	183	15	131	13	317	7	53	8	890	86
5	92	47	76	36	14	2	217	34	203	20	167	6	35	6	804	151
6	86	51	67	35	5	2	129	33	142	18	195	10	29	5	653	154
7	78	51	38	24	4	4	129	29	117	15	113	8	20	2	499	133
8	49	37	28	17	0	0	86	23	73	15	115	12	15	2	366	106
9	32	27	16	11	0	0	72	20	63	13	70	4	7	2	260	77
10	16	14	7	5	0	0	43	12	48	11	66	3	3	0	183	45
11	15	10	3	3	0	0	40	11	46	19	32	1	2	0	138	44
12	11	10	4	4	1	0	15	5	26	9	37	3	0	0	94	31
13	10	3	1	1	0	0	23	7	18	6	11	0	0	0	63	17
14	6	4	6	6	0	0	7	3	6	2	21	0	0	0	46	15
15	3	2	2	1	0	0	7	1	14	2	5	0	0	0	31	6
計	763	293	459	173	57	15	1,261	201	1,377	153	1,766	57	279	30	5,962	922

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 25%、漢字から始まるパターンの約 27%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 400 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 31%、漢字から始まるパターンの約 29%に相当する．

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 830 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 190 パターン以上で約 23%に相当する．カタカナから始まるパターン中では 27%に相当する．

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 1300 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 150 パターン以上で 11%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 1700 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 50 パターンで、3%程しかない．半角アルファベットを含むパターンは 1200 パターン以上存在しそのうち末尾であるものはおよそ 200 パターン、比率は約 15%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 章に掲載した表から、字種連接パターンを構成する個々の字種について、頻度、位置において顕著な特性があることが判明した．はじめにで概説したように、辞書見出し語、論文標題、論文抄録、特許抄録とコーパス毎に調査・分析を進めてきたが、これらの結果の比較分析が今後の課題として挙げられる．

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連接特性の分析 -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 128774 語であったが詳細にスクリーニングした結果，5445 語が不適切であると判明しこれらの語を除外した．

[6] 従来、「字種変化」と記してきたが、「連接」の方が妥当な表現であると判断した．

[7] 紙面の都合により，表 4.2 は最大字種連接パターン長 31 で 14 まで，同様に，表 3.1 は 27 で 13 まで，表 4.3 は 33 で 15 まで，掲載している．