Vol.2014-NL-217 No.16
2014/7/4

特許抄録中の複合語を対象とした字種変化特性の分析

熊澤 侑美[†]，後藤 智範[††]

研究開発の活性化に伴って，新しい概念・モデル・理論を表わす新しい用語が出現する．外国語由来の語は，すぐには漢字標記の訳語が現れないため，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あり，結果として複数の字種で表記される用語が著しく増加する傾向にある。
本研究は NL-214 での報告内容を引き継ぎ，特許抄録に出現した多字種複合語を対象に，字種の観点から，字種並びの特性を明らかにすることを意図するものである．
本報告により，字種変化パターンと用語数、先頭字種毎のパターンと用語について顕著な特性があることを明らかにした．さらに，多字種語の構成単語と字種単位との関係についても考察した．

Analysis to Character Type Sequence of Japanese Compound Terms Appeared in Patent Abstracts

Yumi Kumazawa[†1], Tomonori Gotoh[†2]

Lots of Compound terms used in Japanese technical literatures are written with multi character types. A lot of these terms are consisted of 2 from 5 single words which are expressed with using kanji, katakana, and also alphabets respectively. These terms are increasing as new ideas appear in science, or new technologies are invented in R&D.
Our research intends to analyze to the sequence of multi character types of compound terms appeared in Japanese patent documents. Specifically, about 12 thousands compound terms extracted from patent abstracts were analyzed from character type sequence of view.
It was clear in this research that some specific character type sequence patterns appear many kinds of compound terms. Furthermore, the relation between each component word and character string with single character type in a compound term were considered.

1. はじめに

研究開発の活性化に伴って、新しい概念・モデル・理論を表わす新しい用語が出現する。外国語由来の語は、すぐには漢字標記の訳語が用いられないため、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あり、結果として複数の字種で表記される用語が著しく増加する傾向にある。

このような多字種用語の増加を鑑み、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，過去 3 年間にわたり字種特性、具体的には、構成字種、字種変化パターン観点から以下の調査・分析を行ってきた。

(1) **複数の辞書（専門用語辞典を含む）の見出し語中の多字種複合語の構成字種，字種変化（並び）**[1]
(2) **特許抄録**[2][3]
(3) **学術論文標題（理工学全般）**[4]
(4) **学術論文抄録（(3)と同一文書集合）**[5]
(5) **複数の辞書（専門用語辞典を含む）の見出し語中の先頭字種を限定(漢字，ひらがな，カタカナのみ)した多字種複合語の字種変化 [6]**

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, 2011-NL-204(3)の報告で挙げられた最終的に解析対象となった 135,572 語に対し
（1）全体として名詞(末尾が副詞的，形容詞的接尾辞でない)
（2）連体修飾語を含まない
（3）連用修飾語(句)を含まない
上記 3 点に該当する文字列をさらに除外し, 最終的に表 3.1 に挙げる 135,509 語の多字種複合語を解析の対象とした．

2.2 解析項目

(2)の特許抄録では，次の 2 項目について用語数の相対比率，累積用語数等を調査・分析した．本研究では，字種変化（字種変化数，字種変化パターン）についてより詳細に調査・分析する．具体的には，以下の項目について明らかにする．

(a) 変化数毎の用語総数

[†1] 神奈川大学大学院理学研究科
Graduate School of Science, Kanagawa University
[†2] 神奈川大学理学部情報科学科
Department of Information and Computer Sciences, Kanagawa University

(b) 字種変化パターンの種類

(c) 表 3.3 で挙げた先頭字種毎の字種変化パターン

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

表 2.1 字種記号

(1)	全角漢字	J	(6)	全角数字	N
(2)	全角カタカナ	K	(7)	半角数字	n
(3)	全角ひらがな	H	(8)	全角記号	S
(4)	全角英字	A	(9)	半角記号	s
(5)	半角英字	a			

例えば，AI/MS 無極性フィルター付接続カード の字種並びは 半角アルファベット，半角記号，半角アルファベット，漢字，カタカナ，漢字，カタカナ，漢字，となる．これは非常に冗長な表現形式で分析を困難とするため，表 2.1 の字種記号により，asaJKJK と表わすことにする．本研究では, (1)と同様に, 字種並びを字種変化パターンとよび, また，この例では字種は 9 回変わり，9 を字種変化数とよぶ．このように，異なった字種の並びを 9 種類の記号列として表現することにより，パターン照合的なアプローチが可能となる．

3. 結果

3.1 字種変化毎のパターン数

はじめに字種の変化数毎に出現したパターン数の結果を示す．

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	54	1.23	66,284	48.91
3	271	6.19	40,515	29.90
4	628	14.33	16,060	11.85
5	904	20.63	6,186	4.57
6	830	18.95	2,990	2.21
7	585	13.35	1,455	1.07
8	377	8.61	955	0.70
9	301	6.87	505	0.37
10	167	3.81	254	0.19
11	90	2.05	116	0.09
12	61	1.39	74	0.05
13	31	0.71	31	0.02
14	21	0.48	22	0.02
15	14	0.32	15	0.01
16	9	0.21	9	0.01
17	14	0.32	14	0.01
18	7	0.16	7	0.01
19	4	0.09	4	0.00
20	2	0.05	2	0.00
21	1	0.02	1	0.00
22	2	0.05	2	0.00
23	1	0.02	1	0.00
24	2	0.05	2	0.00
25	1	0.02	1	0.00
26	0	0.00	0	0.00
27	2	0.05	2	0.00
28	1	0.02	1	0.00
29	0	0.00	0	0.00
30	1	0.02	1	0.00
計	4,381	100	135,509	100

表 3.1 は変化数毎の「出現パターン数」と「出現用語頻度」をまとめたものである．変化数 2 ではパターン数は全体の約 1.2%を占めている. 実例として「つなぎデータ(HK)」や「音声特徴パラメータ(JK)」などが挙げられる．変化数 4～7 でパターン数は全体の約 67%を占めており, 出現したパターンの半分以上が字種変化パターン 4～7 であることが分かる．変化数 6 以降からは変化数増加に伴いパターン数と用語頻度は減少し，変化数 13 以降ではほとんどが 1 つの字種変化パターンに 1 用語出現していることが分かる. 変化数 26 と 29 の字種変化パターンは出現しなかった．

変化数 2 の用語頻度は全体の約 49%を占めており，変化数 2～4 で出現用語頻度累計比率は全体の 90%に達する．変化数 5 以降では各出現用語頻度は 5%に達せず, 変化数 8 以降では 1%に達しない．このことから多字種複合語のほとんどが変化数 2～4 であることが確認できる．

表 3.2 パターン数(理論値との比較)

変化数	出現したパターン数	理論パターン数*	出現比率(%)**
2	54	72	75.00
3	271	576	47.05
4	628	4,608	13.63
5	904	36,864	2.45
6	830	294,912	0.28
7	585	2,359,296	0.02
計	3,272	2,696,328	

* 理論パターン数:TP,変化数:Lとしたとき,$TP = 9 * 8^{L-1}$で表される
** 出現パターン数/理論パターン数

表 3.2 は変化数 7 までの出現したパターン数と各変化数で順列組合せによって導かれる理論パターン数を比較したものである．変化数 2 の出現パターン数は 54 であり，これは変化数 2 の理論パターン数の 75%のパターンが出現したこととなる。実際に出現したパターンは表 4.1 に記載した. 出現しなかったパターンは「aH」「Ha」「Hs」「HS」「Js」「Ks」「nH」「nN」「NH」「Ns」「sH」「sJ」「sK」「sn」「sN」「sS」「SH」「Ss」の 18 パターンである．

変化数 5 以降では理論パターン数が多い為に出現比率は少ない．変化数 6 以降では出現比率は 1%に達しないことが分かる．

3.2 先頭字種毎の変化パターン数

次に先頭字種毎に出現したパターン数の結果を示す．

表3.3は先頭字種毎の最長変化数と出現パターン数，用語頻度をまとめたものである．漢字から始まる多字種複合語

は変化数が最大30，出現パターン数は1,400パターン以上で用語頻度は全体の約60%を占める．

表 3.3　先頭字種毎の字種変化パターン数

先頭字種	最長変化数	パターン数	用語頻度	用語比率(%)
J	30	1,472	81,107	59.85
K	24	778	40,018	29.53
A	27	851	7,374	5.44
N	18	666	4,568	3.37
n	19	244	913	0.67
a	28	174	586	0.43
H	12	60	556	0.41
S	25	128	375	0.28
s	12	8	12	0.01
計		4,381	135,509	100

先頭字種 J と K で用語比率は全体の 89%達し，出現する多字種複合語の 89%は漢字かカタカナで始まることが分かる．全角アルファベットで始まるパターンの用語は 7,300 語以上であるが，パターン数は 851 とカタカナで始まるパターンより多い．ひらがなで始まるパターンは 60 と少ないのに対し出現した用語頻度は 550 語以上と多い．このことから先頭字種 H では 1 つのパターンで多くの用語が出現したということが分かる．

表 3.4　変化数毎の出現パターン数
先頭字種非日本語

変化数	a	A	n	N	s	S
2	7	8	6	6	2	6
3	32	45	27	36	3	25
4	40	114	33	89	2	32
5	27	180	53	130	0	22
6	23	164	36	113	0	16
7	13	122	32	94	0	9
8	5	55	18	61	0	9
9	8	64	23	52	0	5
10	7	31	4	34	0	1
11	2	26	4	15	0	1
12	3	11	2	13	1	1
13	2	9	1	8	0	0
14	0	7	1	3	0	0
15	1	3	0	2	0	0
16	1	1	0	2	0	0
17	0	4	2	6	0	0
18	1	2	1	2	0	0
19	0	1	1	0	0	0
20	0	2	0	0	0	0
21	0	0	0	0	0	0
22	0	0	0	0	0	0
23	0	0	0	0	0	0
24	0	1	0	0	0	0
25	0	0	0	0	0	1
26	0	0	0	0	0	0
27	1	1	0	0	0	0
28	1	0	0	0	0	0
29	0	0	0	0	0	0
30	0	0	0	0	0	0
計	174	851	244	666	8	128

表 3.4 と表 3.5 は先頭字種毎に各変化数に出現したパターンをまとめた表となっている．先頭字種が表 3.4 は非日本語（「a」「A」「n」「N」「s」「S」），表 3.5 は日本語（「H」「J」「K」）と分けて記載している．

先頭字種 A が変化数 4～7，先頭字種 N が変化数 5～6 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 21～23，26，29～30 では非日本語先頭字種ではパターンが存在しなかった．

表 3.5 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	5	7	7
3	9	50	44
4	21	184	113
5	16	318	158
6	3	322	153
7	3	203	109
8	1	161	67
9	0	87	62
10	1	59	30
11	0	29	13
12	1	20	9
13	0	9	2
14	0	6	4
15	0	6	2
16	0	2	3
17	0	2	0
18	0	1	0
19	0	2	0
20	0	0	0
21	0	0	1
22	0	2	0
23	0	1	0
24	0	0	1
25	0	0	0
26	0	0	0
27	0	0	0
28	0	0	0
29	0	0	0
30	0	1	0
計	60	1,472	778

先頭字種Jは変化数4～8，先頭字種Kは変化数4～7で各出現パターン数が100以上ある．先頭字種Hは変化数4～5が15パターン以上出現し，他の変化数では10パターン未満しか出現していない．変化数25～29ではどの日本語先頭字種もパターンが出現していない．

表3.6と表3.7は用語頻度について同様にまとめたものである．

変化数 2 で先頭字種 A は 2,600 語以上，先頭字種 N は 1,100 語以上それぞれ出現している．どの先頭字種についても変化数 2 の用語頻度が最も多く，変化数が増加するごとに用語頻度は減少傾向にある．これについては次の表 3.7 にも同じことがいえる．

表 3.6　変化数毎の用語頻度
先頭字種非日本語

変化数	a	A	n	N	s	S
2	231	2,629	188	1,117	4	132
3	151	2,019	105	797	3	101
4	83	1,136	176	981	4	64
5	48	745	134	625	0	32
6	23	352	58	275	0	17
7	13	186	86	314	0	9
8	5	74	92	226	0	11
9	9	93	55	128	0	5
10	7	64	5	45	0	1
11	5	29	6	23	0	1
12	3	16	2	13	1	1
13	2	9	1	8	0	0
14	0	7	1	4	0	0
15	2	3	0	2	0	0
16	1	1	0	2	0	0
17	0	4	2	6	0	0
18	1	2	1	2	0	0
19	0	1	1	0	0	0
20	0	2	0	0	0	0
21	0	0	0	0	0	0
22	0	0	0	0	0	0
23	0	0	0	0	0	0
24	0	1	0	0	0	0
25	0	0	0	0	0	1
26	0	0	0	0	0	0
27	1	1	0	0	0	0
28	1	0	0	0	0	0
29	0	0	0	0	0	0
30	0	0	0	0	0	0
計	586	7,374	913	4,568	12	375

表 3.7　変化数毎の用語頻度
先頭字種日本語

変化数	H	J	K
2	285	39,318	22,380
3	159	25,770	11,410
4	73	9,722	3,821
5	28	3,424	1,150
6	5	1,679	581
7	3	520	324
8	1	376	170
9	0	118	97
10	1	84	47
11	0	36	16
12	1	28	9
13	0	9	2
14	0	6	4
15	0	6	2
16	0	2	3
17	0	2	0
18	0	1	0
19	0	2	0
20	0	0	0
21	0	0	1
22	0	2	0
23	0	1	0
24	0	0	1
25	0	0	0
26	0	0	0
27	0	0	0
28	0	0	0
29	0	0	0
30	0	1	0
計	556	81,107	40,018

先頭字種 J，K は総用語頻度が多いがその半分以上が変化数 2 と 3 で出現している．先頭字種 K は変化数 12，先頭字種 J は変化数 13 以降から各出現用語は 10 語以下となっている．実際に出現した用語は「アンドープＧａ0.88Ａｌ0.12Ａｓ活性層３(KAnsnAnsnAJN)」や「印刷ユニット１，２，３，４，５，６(JKNSNSNSNSNSN)」などが挙げられる．先頭字種 H では変化数 6 以降から用語頻度が 10 語以下となっており，「ねじ継ぎ足し式掘削具(HJHJHJ)」や「ねじ軸１８ａ～１８ｄ(HJNASNA)」などが用語として出現した．

3.3 先頭 2 字種のパターン数・用語頻度

表 3.3 から用語が特に多く出現した先頭字種は「J」,「K」,「A」であることが判明した．この 3 種の先頭字種について 2 字種目にどの字種が続くのかを調査した結果を示す．

表 3.8 先頭 2 字種までのパターン数・用語数(先頭字種 J)

先頭字種	2 字種目	パターン数	パターン数比率(%)	用語頻度	用語比率(%)
J	N	370	25.14	32,813	40.46
	K	382	25.95	31,516	38.86
	H	261	17.73	7,429	9.16
	A	219	14.88	5,450	6.72
	n	105	7.13	2,953	3.64
	S	89	6.05	668	0.82
	a	44	2.99	275	0.34
	s	2	0.14	3	0.00
		1,472	100	81,107	100

表 3.8 は先頭字種 J について 2 字種目まで着目した表である．先頭が J で次が N で始まる用語は「漢字」で始まる多字種複合語中で出現パターンは約 25%，用語頻度は約 40%を占めている．パターン数は「JN」よりも「JK」で始まるパターンの方が多いことが確認できる.「JN」「JK」「JH」「JA」「Jn」で始まる用語でパターン数の累計は約 91%，用語頻度は約 98%を占めている．表 3.9 は「JN」で始まる字種変化パターンを用語頻度が多い上位 5 パターンをまとめたものである．

パターン「JN」だけで用語頻度は 23,000 語を超えており，JN で始まる用語比率の約 70%，J から始まる用語中では約 28%を占めている．上位 5 パターンで JN から始まる用語累計が 90%に達する．

表 3.9　JN で始まる用語頻度上位 5 パターン

パターン	用語頻度	用語比率 (%) *	用語比率 (%) **	用例
JN	23,097	70.39	28.48	圧縮回路２４
JNA	3,505	10.68	4.32	圧力制御弁１３ａ
JNSN	1,414	4.31	1.74	圧力検出器１６，１７
JNJ	1,233	3.76	1.52	拡散層２表面
JNJN	469	1.43	0.58	第１液圧室７８
		90.57	36.64	

＊2 字種目 N 累計用語頻度(32,813)**中**

**** 先頭字種 J 累計用語頻度**(81,107)**中**

表3.10は先頭字種Kについて同様に分析した結果を示したものである．先頭が K で次が J で始まる用語は「カタカナ」で始まる複合語の中でパターン数 391，用語頻度 27,955 と最も多く出現比率も共に 50%を超えている．このことからカタカナで始まる複合語の半分以上は「カタカナ漢字」の字種並びから始まっていることが分かる．「KJ」，「KN」から始まる用語の累計で先頭字種 K の用語の 92%を占めていることが分かる．

表 3.10 先頭 2 字種までのパターン数・用語数(先頭字種 K)

先頭字種	2 字種目	パターン数	パターン数比率(%)	用語頻度	用語比率(%)
K	J	391	50.26	27.955	69.86
	N	127	16.32	9,181	22.94
	A	106	13.62	1,417	3.54
	n	43	5.53	962	2.40
	S	62	7.97	292	0.73
	H	18	2.31	110	0.27
	a	17	2.19	83	0.21
	s	14	1.80	18	0.04
		778	100	40,018	100

表 3.11 では「KJ」で始まる用語頻度上位 5 パターンをまとめたものである．パターン「KJ」のみで約 14,000 語出現し，用語比率は KJ で始まるパターン中では約 50%，先頭字種 K 中では約 35%となっている．上位 3 パターンの累計で「KJ」で始まるパターン中での用語比率は 79%以上，上位 5 パターンで「KJ」から始まるパターンの累計比率は 84%に達している．

表 3.11 KJ で始まる用語頻度上位 5 パターン

パターン	用語頻度	用語比率 (%) *	用語比率 (%) **	用例
KJ	14.125	50.53	35.30	アークセンサ回路
KJN	6,575	23.52	16.43	アーチ形処理室１
KJK	1,518	5.43	3.79	アクティブ遮音パネル
KJKN	667	2.39	1.67	アクセス制御フラグ２０９
KJNA	617	2.21	1.54	アクセス手段４０ａ
		84.07	58.73	

＊2 字種目 J 累計用語頻度(27,955)**中**

**** 先頭字種 K 累計用語頻度**(40,018)**中**

表 3.12 先頭 2 字種までのパターン数・用語数(先頭字種 A)

先頭字種	2 字種目	パターン数	パターン数比率(%)	用語頻度	用語比率(%)
A	J	125	14.69	2,715	36.82
	K	164	19.27	1,786	24.22
	N	99	11.63	1,068	14.48
	n	177	20.80	681	9.24
	S	134	15.75	588	7.97
	a	110	12.93	408	5.53
	s	38	4.47	123	1.67
	H	4	0.47	5	0.07
		851	100	7,374	100

最後に先頭字種Aについてまとめたものを表3.12に示す．用語頻度は先頭が A で次が J で始まる複合語が 2,700 語以上で用語比率は 36%に達している．しかし，用語頻度 681 語の「An」から始まる複合語は出現パターン数 177 と最も多い．この表から先頭字種 A についてはパターン数と用語頻度の関係は一方が多ければもう一方も多いというわけではないことが分かる．

表 3.13 AJ で始まる用語頻度上位 5 パターン

パターン	用語頻度	用語比率 (%) *	用語比率 (%) **	用例
AJ	1.291	47.55	17.51	ＡＦＣ処理
AJN	593	21.84	8.04	ＡＦＣ無同期分離回路１１
AJK	182	6.70	2.47	ＡＦ周波数リスト
AJKN	87	3.20	1.18	ＣＰＵ搭載パッケージ１
		79.30	29.20	

＊2 字種目 J 累計用語頻度(27,15)**中**

**** 先頭字種 A 累計用語頻度**(7,374)**中**

表 3.13 では「AJ」で始まる用語頻度上位 4 パターンをまとめたものである．パターン AJ で「AJ」で始まる複合語の用語比率 47%，先頭字種 A 中では 17%に達している．用語頻度 2 位以降では用語頻度は 600 語以下となりパターン AJKN は用語頻度が 87 語で「AJ」中ではおよそ 3%しか占めていない．

4. 考察

4.1 2 字種変化数パターン

2011-NL-204(3)では字種変化数 2 のパターンについて上位 10 パターンを記載した[3]。次の表 4.1 は変化数 2 で出現した全パターンの用語頻度をまとめたものである．

パターン JN は 23,097 語出現している．これは 2 変化パターンの用語頻度の約 35%を占めている．用語頻度が多い上位 4 パターン(JN，KJ，JK，KN)で用語頻度累計比率は約 82%を占めている．これらの実例として「圧力板２２(JN)」，「予測システム(JK)」などが挙げられる．

表 4.1　変化数 2 のパターン(54 種類)

パターン	用語頻度	比率 (%) *	パターン	用語頻度	比率 (%) *	パターン	用語頻度	比率 (%) *
JN	23,097	34.85	JS	142	0.21	SN	19	0.03
KJ	14,125	21.31	Ja	125	0.19	AS	17	0.03
JK	10,369	15.64	aJ	103	0.16	HK	16	0.02
KN	6,891	10.40	an	78	0.12	KS	16	0.02
JA	2,668	4.03	na	73	0.11	As	14	0.02
Jn	2,065	3.12	Na	60	0.09	nK	14	0.02
AJ	1,291	1.95	NS	47	0.07	as	9	0.01
JH	852	1.29	nA	43	0.06	aN	8	0.01
Kn	732	1.10	KH	32	0.05	aA	6	0.01
NJ	594	0.90	nS	31	0.05	Hn	6	0.01
KA	557	0.84	Sa	31	0.05	aS	5	0.01
AN	541	0.82	SK	28	0.04	Sn	4	0.01
AK	415	0.63	HN	27	0.04	HA	3	0.00
NA	252	0.38	Ka	27	0.04	sa	3	0.00
HJ	233	0.35	SA	27	0.04	AH	2	0.00
Aa	176	0.27	nJ	25	0.04	ns	2	0.00
An	173	0.26	SJ	23	0.03	Nn	1	0.00
NK	163	0.25	aK	22	0.03	sA	1	0.00
						計	66,284	100

* 各パターンの用語頻度/2変化の総用語頻度(66,284)

変化数2の出現パターン全54種類と変化数3以上のパターンに対して照合を行った．表 4.2 は出現回数をまとめたものである．長変化パターンに最も出現している「NS」は 47 語しか出現していない．用語頻度が最も多い「JN」は長変化パターンには 939 回しか出現していない．

表 4.2　2 変化パターンの 3 変化以上の
パターン中での出現回数

パターン	出現回数	パターン	出現回数	パターン	出現回数
NS	1137	AK	421	nK	150
SN	1122	Sn	366	Sa	146
KJ	1033	HJ	355	Kn	138
JN	939	JS	341	aJ	120
JK	870	AN	337	Nn	120
KN	662	Aa	277	HK	116
SA	583	Jn	259	aA	106
nS	537	an	246	Ja	85
NA	515	na	244	aK	75
An	509	nA	230	As	73
NJ	508	SJ	225	HN	63
KA	501	KS	222	Ka	62
JH	494	SK	213	sA	31
JA	483	as	209	KH	29
AS	470	nJ	204	HA	27
ns	439	sa	186	Hn	23
NK	437	aS	177	aN	20
AJ	422	Na	161	AH	8

長変化パターンに含まれている 2 変化パターンの実例を示す．

実例:　Ｌ／Ｃ１／Ｃ２逆変換手段３

字種変化パターン ASANSANJN　に含まれている 2 変化パターンは可能性として「AS」,「SA」,「AN」,「NS」,「SA」,「AN」,「NJ」,「JN」となる．「SA」と「AN」は 2 回ずつ出現していることが分かる．対応する文字列は「Ｌ／」,「／Ｃ」,「Ｃ１」,「１／」,「／Ｃ」,「Ｃ２」,「２逆変換手段」,「逆変換手段３」となる．この中で「Ｃ１」,「Ｃ２」,「逆変換手段３」が意味単位として妥当であると考えられる．字種変化パターン　ASANSANJN　は　AN　AN　JN となり，2 変化パターンとして AN と JN が含まれることになる．

実例:　制御タイミング生成データ

字種変化パターン　JKJK　に含まれている2変化パターンは可能性として「JK」,「KJ」,「JK」となり，対応する文字列は「制御タイミング」,「タイミング生成」,「生成データ」である．この実例は意味単位として「制御タイミング」と「生成データ」が妥当である．字種変化パターン　JKJK　は　JK　JK　となり，2 変化パターンとして JK が含まれることになる．

実例:　糸太さ自動検出機能付き刺繍ミシン

字種変化パターンは　JHJHJK　である．この字種変化パターンに含まれている 2 変化パターンは可能性として「JH」,「HJ」,「JH」,「HJ」,「JK」となる．対応する文字列は「糸太さ」,「さ自動検出機能付」,「自動検出機能付き」,「き刺繍」,「刺繍ミシン」であり，用語は「糸太さ」,「自動検出機能付き」,「刺繍ミシン」が意味単位として妥当である。字種変化パターン　JHJHJK　は　JH　JH　JK　となり，2 変化パターンとして JH と JK が含まれることになる．

実例:　基材フイルム１／剥離層２／保護層３／金属蒸着層５／接着層６

字種変化パターンは　JKNSJNSJNSJNSJN　であり、このパターンに含まれる 2 変化パターンは可能性として「JK」,「KN」,「NS」,「SJ」,「JN」,「NS」,「SJ」,「JN」,「NS」,「SJ」,「JN」,「NS」,「SJ」,「JN」となる．対応する文字列は「基材フイルム」,「フイルム１」,「１／」,「／剥離層」,「剥離層２」,「２／」,「／保護層」,「保護層３」,「３／」,「／金属蒸着層」,「金属蒸着層５」,「５／」,「／接着層」,「接着層６」となる．字種変化パターン JKNSJNSJNSJNSJN　は　JKN　S　JN　S　JN　S　JN　S　JN　となり，2 変化パターンとして JN が含まれることになる．

4.2 独立した語の構成

表 4.1 と 4.2 から以下の仮説が立てられる．

2 変化パターンは独立した語を構成している
変化パターン

この仮説に対し「ゆで卵収容室１３(HJN)」という語に対し2変化パターンを照合した結果

(1) ゆで卵収容室(HJ)

(2) 卵収容室１３(JN)

の2単語が抽出できる．しかし，両用語ともに独立したパターンではない．多字種複合語は適切な位置で字種が区切られ，区切られた単位の字種変化に該当する2変化パターンが独立した語を構成していると考えられる．

4.3 構成単語と字種変化

多字種複合語中で構成単語数と字種変化数との間に関係があると考えられる．字種変化数は2.2節で述べたものであり，構成単語数はある用語を単語単位で分割したときの単語の個数をいう．

4.2 節の例では構成単語数と字種変化数は以下のようになる．

ゆで卵収容室１３　　HJN

構成単語:2 単語(ゆで卵　収容室１３)

字種変化: HJN (3 変化)

我々はこれまでの調査から多字種複合語中の構成単語は字種変化単位で分割されていると考えていた．しかし，上記の多字種複合語では独立した語を構成しているが字種の同一によって2変化パターンに該当する用語が抽出できていない．このように多字種複合語中の独立した語と構成単語数,字種変化数の関係性には以下5つの現象が見られた．

(1) 先行単語と後続単語の字種が同一

(2) 単語間の連接強度

(3) 複数字種で表記される1単語

(4) ひらがな表記

(5) 記号の機能(複数単語の連結)

(1) 先行単語と後続単語の字種が同一

単語と単語の字種が同一であるために発生する現象であり，「ゆで卵収容室１３(HJN)」がこれに該当する．構成単語は「ゆで卵(HJ)」,「収容室１３(JN)」であるが，「卵」と「収容室」が「漢字(J)」の1字種でまとめられていることで2変化パターンの情報を用いて独立した用語で分割することができない．

(2) 単語間の連接強度

字種変化しているが多字種複合語ではなく1単語として扱うことが妥当であると考えられる語である.例として「ビタミンＢ1 (KAn)」などが挙げられる．

(3) 複数字種で表記される1単語

「第１アルゴリズム(JNK)」や「１位(NJ)」のように「Jn」や「NJ」序数詞の役割を持っている語である．序数詞については(2)と同様に1単語として扱うことが妥当であると考えられる．

(4) ひらがな表記

漢字の送り仮名を含む用語では送り仮名がつくたびに字種変化が発生するため字種変化数が多くなる.「誤り検出プロセス(JHJK)」のように「JH」の部分が該当し，「押し付け力制御(JHJHJ)」など「JH」が連続するパターンに多く見られる．

(5) 記号の機能(複数単語の連結)

「ー」,「-」,「/」などの記号が複数の単語を連結させる役目をしている場合である．「立上り／立下り(JHSJH)」などが挙げられる．

5. 終わりに

本研究の結果から以下のような事象が判明した．

・ 先頭がJで次がNで始まる用語は先頭字種Jの多字種複合語中で出現パターンは約25%，用語頻度は約40%を占めていた．（表3.8）

・ 先頭がKで次がJで始まる用語は先頭字種Kの複合語の中でパターン数，用語頻度の出現比率は共に50%を超えている．（表3.10）

・ 用語頻度が最も多い「JN」は長変化パターンには939回しか出現していなかった．

今回の研究では2字種変化パターンを3字種変化以上のパターンに対し照合を行い，字種変化パターンと実文字列との関係性を調査した．その結果，(1)先行単語と後続単語の字種が同一であるため2字種変化パターン単位で取り出すと単語として不適切である事例もある，(2)単語間の連接強度などにより2字種変化のパターンに該当するが1単語として扱うことが妥当な用語が存在する，(3)送り仮名の役割を持つひらがな表記，(4)複数単語の連接の役割を持つ記号の機能，の4つの特長が知見として得られ，字種変化数と複合語を構成している意味単位の単語数は必ずしも一致しないことが判明した．今後の課題として，今回判明した上述の問題も考慮したうえで，字種変化パターンと複合語の文字列についてのより詳細な調査・分析，１章で挙げた学術論文標題(理工学全般)，学術論文抄録に対して本研究と同様のアプローチを採用した調査・分析がある．

註・参考文献

1) 滝川諒，後藤智範．大規模複合語データに対する構成字種解析．自然言語処理研究会報告 2011-NL-202(1), 1-7, 2011-07-08

2) 滝川諒，後藤智範．特許抄録に出現する多字種複合語に対する字種に基づく解析 part.1- 多字種複合語の抽出と構成字種の解析 -自然言語処理研究会報告 2011-NL-204(2), 1-15, 2011-11-14

3) 滝川諒，後藤智範．特許抄録に出現する多字種複合語に対する字種に基づく解析 part.2- 字種変化パターンの解析 -自然言語処理研究会報告 2011-NL-204(3), 1-12, 2011-11-14

4) 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．

5) 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．

6) 熊澤侑美，斉藤恵，後藤智範．辞書見出し語中の複合語を対照とした字種変化特性の分析．-自然言語処理研究会報告 2013-NL214(17),2013-11-15