Vol.2013-NL-214 No.16
2013/11/15

辞書見出し語中の複合語を対象とした字種変化特性の分析

熊澤 侑美[†]，斎藤恵[†]，後藤 智範[††]

近年，理工学分野の学術論文，特許などの文書の用いられる専門用語は，複数の字種で表記される複合語が多くみられる．研究・開発が増加するにつれ，複数字種表記の専門用語が増加する傾向がある．本研究は NL-202 で報告した研究内容を引き継ぎ，辞書見出し語中の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，対象用語集合には字種並びについて以下に挙げる顕著な特性があることが判明した．

(a) 字種変化は 2 から 13 で，計 590 種類以上の字種変化パターンが見られた．
(b) 全用語数のうち 95%以上が 2～ 4 字種変化パターンの形態をとっていた．
(c) 2～4 字種変化パターン構成の用語の 95%以上が，漢字またはカタカナで開始される．
(d) 全用語数から得られた字種変化パターンの 95%以上は，漢字またはカタカナで開始するパターンであった．

Analysis to Character Type Sequence of Japanese Compound Terms Extracted from Lots of Entry Terms of Several Dictionaries

Yumi Kumazawa[†1], Megumu Saito[†1], Tomonori Gotoh[†2]

Lots of Compound terms used in Japanese technical literatures and patent documents are consisted with multi character types. Technical terms written in multi character types in these texts are increasing as New Ideas appear in science, or new technologies are invented in R&D.

This research intends to analyze to the sequence of multi character types of compound terms extracted from entry terms in the multiple dictionaries. Specifically, about 11 thousands compound terms starting with one of three types of character (hiragana, katakana, and Chinese character) were analyzed from the pattern matching point of view.

The following characteristics to the compound terms set were found by this research.

(a) the range the length of character type sequence was 2 to 13.
(b) there were 599 patterns of character type sequence in the compound term set.
(c) The compound terms by over 95% were consisted of 2 to 4 in the length of character type sequence.
(d) The compound terms by over 95% were started with katakana or Chinese character, in other words, compound terms starting with hiragana were very few.

1. はじめに

近年，理工学分野の学術論文，特許などの文書に用いられる専門用語は，複数の字種で表記される複合語が多くみられる．研究・開発が増加するにつれ，複数字種表記の専門用語が増加する傾向がある．

日本語の専門用語に対して多くの研究がなされ，最近もいくつかの研究がある[1][2][3]．しかしながら，コーパスサイズは小さく，また字種並びの観点での研究は非常に少ない[5].

筆者らは，辞書見出し語，テキスト（文書の標題・抄録）中に出現する，多字種複合語について，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）[4]**
(2) **特許抄録[5][6]**
(3) **学術論文標題（理工学全般）[7]**
(4) **学術論文抄録（(3)と同一文書集合）[8]**

本研究は，(1)と同様のコーパスを使用し，2.1 節で述べる基準により縮小した用語体集合を対象に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

(1)の調査分析対象である，約 12 万語の多字種複合語集合には，以下に挙げる化学物質名のような，特定分野に限定して用いられる用語てきな狭義での「用語」として扱うには，以下のような表現が散見された．

本報告では，(1)の多字種複合語のうち，先頭が以下の字種が始まる用語 113,397 語を調査対象とした[9][10].

漢字，ひらがな，カタカナ

2.2 解析項目

(1)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．

(a) 構成字種（構成字種数，構成字種組み合わせ）
(b) 字種変化（字種変化数，字種変化パターン）

[†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

本研究では，(b)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
(b) 字種変化パターンの種類
(c) 2.1 で挙げ先頭字種毎の字種変化パターン

また，字種変化パターンの構成について考察する．

(1)と同様に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	パターン数比率	用語頻度	用語頻度比率
2	14	2.337229	78,369	69.11029
3	56	9.348915	25,902	22.84187
4	131	21.86978	6,667	5.879344
5	161	26.87813	1,689	1.489457
6	126	21.03506	537	0.473558
7	64	10.68447	145	0.127869
8	28	4.674457	67	0.059084
9	13	2.170284	15	0.013228
10	2	0.33389	2	0.001764
11	2	0.33389	2	0.001764
12	1	0.166945	1	0.000882
13	1	0.166945	1	0.000882
	599	100	113,397	100

表 3.1 は字種変化数毎のパターン数と用語頻度データである．変化数 2 から 5 までは変化数の増加に従い出現パターン数も増加していることが分かる．しかし，変化数 6 以降ではパターン数は減少傾向にあり，変化数 6 から 7 へはパターン数が半分ほど減少している．変化数 4，5，6 でそれぞれ 100 パターン以上出現し，全体の約 69%を占めていることが分かる．その中でも変化数 5 が 161 パターンと最もパターン数が多い．

変化数 2 で用語頻度は 78,369 と全体の約 70%を占めている．変化数 2 と 3 の用語で用語比率は 91%に達する．変化数の増加に伴い用語頻度も減少傾向にあることが分かるが, 変化数 3 から 4 の減少が特に顕著である．変化数 4 から 13 までの用語比率は 10%に満たないことが分かる．

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(出現/理論)
2	14	18	77.78
3	56	108	51.85
4	131	648	20.22
5	161	3,888	4.14
6	126	23,328	0.54
7	64	139,968	0.05

表 3.2 は出現したパターン数と理論パターン数の比較を示している．字種変化数を L として理論パターンは，以下の式で表される．

$$L = 3 \cdot 6^{L-1}$$

3: 先頭字種数， 6: 後続する字種数(A，N が出現しなかったため)

変化数 2 は 18 パターン中 14 パターンが出現している．出現しなかった 4 パターンは「Hn」,「Hs」,「HS」,「Ks」であった．つまり，先頭字種 J は全ての字種が連接していることが分かる. 変化数6と7では理論パターン数が多く, 実際に出現したパターンの比率は 1%にも満たない．

3.2 先頭字種毎の変化パターン数

表3.3は先頭字種毎の字種変化パターン数と用語頻度を示している．

表 3.3 先頭字種毎の字種変化パターン数

先頭字種	パターン数	用語頻度	用語数比率
K	289	54,258	47.85
J	269 [10]	54,159	47.76
H	41	4,980	4.39

計	599	113,397	100.00

先頭字種カタカナと漢字で始まる複合語はパターン数・用語数共に同等である．上位 2 つでパターン数累計が 550 パターンを超えている．用語頻度比率も 95%に達することが分かる．しかし，先頭字種ひらがなはパターン数 41，用語頻度 4,980 語と先頭字種カタカナ・漢字と比較すると少ない．このことから，日本語字種から始まる複合語のほとんどはカタカナや漢字のものが大半を占めていることが判明した．

表 3.4　先頭 2 字種のパターン数・用語頻度
(先頭字種:カタカナ)

先頭字種	2 字種目	パターン数		用語頻度	
K	**J**	**100**	**34.6**	**46,485**	**85.7**
	S	82	28.4	4,232	7.8
	N	40	13.8	1,702	3.1
	A	44	15.2	1,127	2.1
	H	15	5.2	661	1.2
	S	8	2.8	51	0.1
計		289		54,258	

表 3.4 はカタカナから始まる先頭 2 字種に着目し字種変化パターン数と用語頻度を列挙したものである．カタカナで始まり 2 字種目が漢字である複合語は 100 パターン，6,485 語出現した．カタカナから始まる用語は 2 字種目が漢字である用語が最も多く 100 パターン, 46,485 語出現し，用語比率も約 86%を占めている．漢字の次にパターン数と用語数が多いのが全角記号である．

表 3.5 は表 3.4 で用語頻度が最も多い 2 字種目漢字について，上位 4 パターンの用語頻度を列挙したものである．字種変化パターン KJ で 40,179 語出現しており，これは 2 字種目 J の複合語の約 86%である．先頭字種 K の用語の中では約 74%を占めている．上位 2 パターンで 2 字種目漢字での用語比率は 94%に達する．

表 3.5　2 字種目 J の用語頻度(先頭:カタカナ)

パターン	用語頻度	用語比率 *1	用語比率 *2	用語実例
KJ	40,179	86.43	74.05	アーカイバル記憶
KJK	3,877	8.34	7.15	アーク遮断ストリップ
KJH	645	1.39	1.19	アームベッド締ねじ
KJKJ	547	1.18	1.01	アーチ式フレーム構造
累積		97.34	83.39	

*1 : 2 字種目 J 中（46,485 語）での用語比率
*2 : 先頭字種 K 中（54,258 語）での用語比率

表 3.6 は漢字から始まる先頭 2 字種に着目しパターン数と用語頻度を統計したデータである．

表 3.6　先頭 2 字種のパターン**の種類**・用語頻度
(先頭字種:漢字)

先頭字種	2 字種目	パターンの種類		用語頻度	
J	**K**	**79**	**29.4**	**32,405**	**59.8**
	H	**66**	**24.5**	**18,400**	**34.0**
	n	32	11.9	1,358	2.5
	a	41	15.2	1,097	2.0
	S	40	14.9	830	1.5
	s	11	4.1	69	0.1
計		269[10]		54,159	

2 字種目がカタカナ(K)である複合語は 79 パターン，32,405 語，2 字種目がひらがな(H)である複合語は 66 字種変化パターン，18,400 語出現した．上位 2 字種で用語頻度の比率は 93%に達する．2 字種目が半角数字(n)と半角英字(a)のパターンを比較すると，用語頻度が多ければパターン数も多いというわけではないことが分かる．

表 3.7 と表 3.8 は表 3.6 で用語頻度が最も多い上位 2 字種，2 字種目がカタカナおよびひらがなについて，上位 4 パターンの用語頻度をまとめた統計データである．

表 3.7　2 字種目 K の用語頻度(先頭字種:漢字)

パターン	用語頻度	用語比率 *3	用語比率 *4	用語実例
JK	23,359	72.08	43.13	亜イレウス
JKJ	7,659	23.64	14.14	亜アンチモン酸塩化物
JKJK	638	1.97	1.18	亜ジチオン酸ナトリウム
JKJKJ	122	0.38	0.23	亜スズ酸ナトリウム試液
累積		98.07	58.68	

*3 : 2 字種目 K 中(32,405 語)での用語比率
*4 : 先頭字種 J 中(54,159 語)での用語比率

表 3.7 からパターン JK だけで 23,359 語出現していることが分かる．これは 2 字種目 K の複合語の約 72%であり，先頭字種 K の用語の中では約 43%を占めている．上位 2 パターンで 2 字種目 K 中での用語比率の 95%に達する．

表 3.8 からパターン JH で用語頻度 8,389 語，2 字種目ひらがな中では用語比率 45%に達し，先頭字種漢字の中では約 15%に達している．2 字種目ひらがなの複合語は上位 4 パターンで 2 字種目ひらがな中の比率は 93%に達する．しかし先頭字種漢字中の比率は 31%に達するだけである．

表 3.8　2 字種目 H の用語頻度(先頭字種:漢字)

パターン	用語頻度	用語比率 *5	用語比率 *6	用語実例
JH	8,389	45.59	15.49	亜鉛めっき
JHJ	6,957	37.81	12.85	亜しきい値線量
JHK	938	5.10	1.73	圧造割りダイス
JHJH	935	5.08	1.73	圧縮せん断強さ
累積		93.58	31.79	

*5 : 2 字種目 H 中(18,400 語)での用語比率
*6 : 先頭字種 J 中(54,159 語)での用語比率

表 3.9 はひらがなから始まる先頭 2 字種に着目し，字種変化パターン数と用語頻度を示している．2 字種目が漢字である複合語は 28 種の字種変化パターンとなり,用語頻度 4,718 語である．2 字種目漢字のパターンでパターン比率は 68%に達し，用語頻度比率も 94%に達している．

表 3.9　先頭 2 字種のパターンの種類・用語頻度
(先頭字種:ひらがな)

先頭字種	2 字種目	パターンの種類		用語頻度	
H	**J**	**28**	**68.3**	**4,718**	**94.7**
	K	7	17.1	250	5.0
	S	3	7.3	6	0.1
	A	2	4.9	4	0.1
	N	1	2.4	2	0.0
計		41		4,980	

表 3.10 は表 3.9 で用語頻度が最も多い字種，2 字種目漢字について，上位 4 パターンの用語頻度をまとめた統計データである．

表 3.10　2 字種目 J の用語頻度(先頭字種:ひらがな)

パターン	用語頻度	用語比率 *7	用語比率 *8	用語実例
HJ	3,480	73.76	69.88	あいまい化
HJH	555	11.76	11.14	あいまい誤り
HJHJ	247	5.24	4.96	あおり後掛金ハンドル
HJK	239	5.07	4.80	あさりだし機
累積		95.82	90.78	

*7 : 2 字種目 J 中(4,718 語)での用語比率
*8 : 先頭字種 H 中(4,980 語)での用語比率

パターン HJ で 3,480 語出現しており，これは 2 字種目 J の複合語の約 74%, 先頭字種 H の用語の中では約 70%を占めている．

4. 考察

4.1 短字種変化数パターン（2，3）

前章の表 3.1 に字種変化数毎のパターン数および用語数を挙げ，字種変化数が 2 または 3 について，字種変化パターンの種類数は少ないが，両者だけで，調査対象用語の 91 を占めていることを指摘した．表 4.1 に変化数 2，表 4.3 に変化数 3 の個々のパターンの用語頻度を挙げる．

表 4.1　変化数 2 のパターン(14 種類)

パターン	用語頻度	用語累積頻度	比率 *9	パターン	用語頻度	用語累積頻度	比率 *9
KJ	40,179	40,179	51.27	Ja	368	78,030	0.47
JK	23,359	63,538	29.81	HK	218	78,248	0.28
JH	8,389	71,927	10.70	Jn	73	78,321	0.09
HJ	3,480	75,407	4.44	KS	36	78,357	0.05
Kn	1,342	76,749	1.71	JS	8	78,365	0.01
Ka	530	77,279	0.68	Ha	3	78,368	0.00
KH	383	77,662	0.49	Js	1	78,369	0.00

*9 : 用語頻度/78,369*100

表 4.1 から，２字種構成の字種変化パターン，それぞれ全てが多いのではなく，「KJ」と「JK」の 2 種類，すなわち 2 変化パターン総数の 14%だけで，2 変化パターンの全用語の約 80%に達していることがわかる．

表 4.2　変化数 3 のパターン(56 種類)

パターン	用語頻度	用語累積頻度	比率 *10	パターン	用語頻度	用語累積頻度	比率 *10
JKJ	7,659	7,659	29.57	KJn	24	25,740	0.09
JHJ	6,957	14,616	26.86	Kna	20	25,760	0.08
KJK	3,877	18,493	14.97	KsJ	19	25,779	0.07
KSK	1,515	20,008	5.85	Jan	17	25,796	0.07
JnJ	1,048	21,056	4.05	KnK	14	25,810	0.05
JHK	938	21,994	3.62	KsK	14	25,824	0.05
KJH	645	22,639	2.49	JHa	9	25,833	0.03
HJH	555	23,194	2.14	JKn	8	25,841	0.03
JaJ	511	23,705	1.97	JSn	8	25,849	0.03
JSJ	502	24,207	1.94	JsK	7	25,856	0.03
HJK	239	24,446	0.92	Jna	6	25,862	0.02
KaJ	193	24,639	0.75	Jas	4	25,866	0.02
KHJ	193	24,832	0.75	JaS	4	25,870	0.02
KaK	159	24,991	0.61	JSa	4	25,874	0.02
KnJ	90	25,081	0.35	Kas	4	25,878	0.02

KSJ	89	25,170	0.34	KaS	4	25,882	0.02
Kan	77	25,247	0.30	JKS	3	25,885	0.01
KSa	67	25,314	0.26	KJS	3	25,888	0.01
JKa	65	25,379	0.25	HJa	2	25,890	0.01
KJa	65	25,444	0.25	HKH	2	25,892	0.01
JaK	50	25,494	0.19	HnJ	2	25,894	0.01
JKH	44	25,538	0.17	KHa	2	25,896	0.01
JsJ	40	25,578	0.15	HJs	1	25,897	0.00
KSn	33	25,611	0.13	HSH	1	25,898	0.00
KHK	30	25,641	0.12	HSJ	1	25,899	0.00
HKJ	25	25,666	0.10	JaH	1	25,900	0.00
JnK	25	25,691	0.10	Jsa	1	25,901	0.00
JSK	25	25,716	0.10	KaH	1	25,902	0.00

*10： 用語頻度/25,902*100

表 4.2 から，用語頻度上位の 6 種類の変化パターンで全体の 85%を占めていることがわかる．以下にこれらの字種変化パターン構成の用語実例を以下に列挙する．

JKJ: 亜アンチモン酸塩化物　　JHJ: 亜鉛めっき鋼管
KJK: アーク遮断ストリップ
KSK: コンピュータアナログインプット/アウトプット
JnJ: 悪性 2 相性中皮腫　　JHK: 圧造割りダイス

4.2 短変化数の包含

本調査の過程において，前節で挙げた，短変化数パターンが長変化パターン（4 以上）に頻出することが散見された．この事実を明らかにするために，2 および 3 変化パターンを対象に，自分自身より長いパターンに対し，文字列照合（力まかせ法）を行い，一致した位置，回数を出力し分析した．

表 4.3 に変化数 2，表 4.4 に変化数 3 について，それぞれ部分一致した長変化パターン（2 変化は変化数 3 以上のパターン，3 変化は変化数 4 以上のパターン）のパターン数を列挙したものである．

表 4.3　2 変化パターンの包含

パターン	部分一致パターン数	パターン	部分一致パターン数	パターン	部分一致パターン数
KJ	264	Ka	82	KS	121
JK	226	KH	39	JS	90
JH	136	Ja	73	Ha	12
HJ	120	HK	56	Js	17
Kn	53	Jn	63		

3 変化以上のパターン総数は，表 3.2 から 585 種ある．表 4.3 より，パターン，KJ, JK は約 50%，JH, HJ は 20%のパターンに部分パターンとして含んでいることが判明した．

表 4.4　3 変化パターンの包含

パターン	部分一致パターン数	パターン	部分一致パターン数	パターン	部分一致パターン数
JKJ	85	KJa	20	Jna	5
JHJ	84	JaK	12	Jas	3
KJK	75	JKH	14	JaS	19
KSK	53	JsJ	8	JSa	8
JnJ	28	KSn	30	Kas	5
JHK	38	KHK	10	KaS	21
KJH	41	HKJ	24	JKS	29
HJH	48	JnK	14	KJS	36
JaJ	22	JSK	25	HJa	6
JSJ	33	KJn	26	HKH	6
HJK	33	Kna	9	HnJ	2
KaJ	17	KsJ	6	KHa	0
KHJ	17	Jan	6	HJs	1
KaK	13	KnK	9	HSH	1
KnJ	10	KsK	4	HSJ	2
KSJ	20	JHa	9	JaH	0
Kan	14	JKn	9	Jsa	1
KSa	20	JSn	17	KaH	1
JKa	21	JsK	2		

4 変化以上のパターン総数は，表 3.2 から 529 種ある．表 4.4 により，パターン，JKJ, JHJ，KJK は約 15%，JH, HJ は 20%のパターンに部分パターンとして含んでいることが判明した．

4.3 正規表現

2.2 節で述べたように，多字種複合の字種並びは，9 種類からなる記号の記号列として，パターン照合的な視点によりアプローチが可能である．

図 4.1 は先頭字種ひらがなのパターン全 28 種を正規表現として表わしたものである．

^H(nJ)		^Ha($ | HJ)	
^HS(J|H|HJ)$		^H(K|KH)$	
^HK(J|JH)$		^HKJ(K|KJ)$	
^HKJ(nJ)$		^H(J|Ja|JaJ)$	
^HJ(H|HaJ)$		^HJH(J|JH|JHJ|JHJH)$	
^HJHJHJ(K|KJ)$		^HJHJH(K)$	
^HJHJ(K)$		^HJH(K|KJ|KJH)$	
^HJ(K|KH|KHJ)$		^HJK(J|JHJ)$	
^HJKJ(K|KJ)$		^HJK(SaSK)$	
^HJ(nJK)$		^HJ(s)$	
^HJ(SJKJ)$		^HJS(KJSKJ)$	

図 4.1　ひらがなで始まる字種変化パターンの正規表現

パターンの完全一致を表現するため「^」を用いて先頭がひらがなから開始するという制約を与えている．さらに，「$」を用いることで「$」の直前の字種が文末であるパターンを受理する仕様となっている．

図4.1の正規表現をlexプログラムに書き換えたものの一部を図 4.2 に示す．lex の制約により規則の順序をパターンの長さ順に記述してある．一致しなかったパターンに対してエラー文を表示する仕様としている．

```
/*ひらがな*/
%%
^H(a|J|K)($|¥t|¥n|¥r|¥r¥n) {printf("input text is %s:2length match!¥n", yytext);} /*字種長 2*/
^H(Ja|JH|JK|Js|KH|KJ|nJ|SH|SJ)($|¥t|¥n|¥r|¥r¥n) {printf("input text is %s:3length match!¥n", yytext);} /*字種長 3*/
・
・
^H(JHJHJH|JHJHJK|JKSaSK)($|¥t|¥n|¥r|¥r¥n) {printf("input text is %s:7length match!¥n", yytext);} /*字種長 7*/
^H(JHJHJKJ|JSKJSKJ)($|¥t|¥n|¥r|¥r¥n) {printf("input text is %s:8length match!¥n", yytext);} /*字種長 8*/
.* {printf("¥nError!! input text is %s:No match!¥n", yytext);} /*エラー処理*/
%%
```

図 4.2　lex プログラムの一部

lex プログラムに対し，出現した全パターン（lex の規則と一致）と出現しないパターンを入力として与え，照合を行った．図 4.3 に実行結果の一部を示す．全ての出現したパターンは一致し，出現しないパターン（この例では HaH）は不一致となり，記述された規則が妥当であることが検証された．

```
input text is HnJ : 3 length match!
input text is Ha : 2 length match!
input text is HaHJ : 4 length match!
・
・
input text is HJs : 3 length match!
input text is HJSJKJ : 6 length match!
input text is HJSKJSKJ : 8 length match!
Error!! input text is HaH : No match!
```

図 4.3　正規表現照合結果の一部

5. 終わりに

本研究の結果から，以下の事象が判明した．

(a) 字種変化は 2 から 13 で，計 590 種類以上の字種変化パターンが見られた．

(b) 全用語数のうち 95%以上が 2～4 字種変化パターンの形態をとっていた．

(c) 2～4 字種変化パターン構成の用語の 95%以上が，漢字またはカタカナで開始される

(d) 全用語数から得られた字種変化パターンの 95%以上は，漢字またはカタカナで開始するパターンであった．

(e) カタカナはひらがなよりも全角記号と連接した複合語のほうが用語数・パターン数共に多い

(f) 特定の 2 または 3 字種変化パターンがより長い字種変化パターンの一部として含まれる．

今回の研究では，字種変化パターンを字種記号列となみし，パターン照合問題的なアプローチを試みたが，分析過程で，(a)字種変化数は構成単語数と一致しない事例もある，(b)先頭に位置するひらがなと，漢字表記される単語に後続するひらがなと同一の扱いをするのは適切ではない可能性があること，などである．字種記号列（字種変化）は，単語列（終端記号），（単語の）品詞列との関連でより多角的に捉える必要があると思われる．

今後の課題として，上述の問題を考慮したうえで，1 章で挙げた 3 種類のテキスト中の多字種複合語，すなわち特許抄録，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) 小山照夫，影浦峡，竹内孔一．日本語専門分野テキストコーパスからの複合語用語の抽出．自然言語処理研究会報告（124），55-60 （2006）．
2) 三枝 優一，古井 陽之助，速水 治夫. Web から新語を動的に獲得する形態素解析用辞書拡張方式．データベース・システム研究会報告 （6）, 77-82 （2007）．
3) 小山照夫．日本語テキストからの複合語用語抽出．情報知識学会誌 19（4）, 306-315 （2009）．
4) 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
5) 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
6) 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
7) 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．
8) 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．
9) 分野(例えば化学)によっては，このような字種限定は不適切となりうる．
10) 先の研究 4)では，当該パターン数は 270 となっているがデータ数の再確認の過程で，正しい数値は 269 であった．