Vol.2011-NL-204 No.3
2011/11/21

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 -

滝川 諒[†]　　後藤 智範[††]

Part.2 では多字種複合語集合に対して、字種の観点からより詳細な特性を明らかにするために、連接する構成単語の字種の順序、出現回数も分析の対象とした。Part.1 で抽出された複合語 13 万 6 千語を対象に、個々の複合語に対し構成単語の字種について字種コードを付与した。得られた字種コードデータについて，用語出現頻度，相対出現比率，累積出現頻度，累積出現比率を算出した。

結果として，変化数の増加に伴う出現比率の単調減少，変化数に依存しない多変化の上位パターンの存在，字種による単語形成能力の差異と字種変化数の関係，などの特性が明らかとなった.

Quantitative Analysis to Japanese Compound Terms with Multi Character Types Appeared in Patent Texts Part.2

Ryo Takikawa[†]　　Tomonori Gotoh[††]

In part.2 of our research, over 135 thousands Japanese compound terms were used to the analysis which were the same corpus to be analyzed in part.1. Various character types used in each term were automatically discriminated by the computer program, and respectively assigned sequence of character type codes. These compound terms with the sequence of the codes were counted from the point of occurrence frequencies based on sequence patterns of the codes.

It was found that over 4000 kinds of the sequence patterns were identified in about 136,000 compounds terms. Moreover, the several facts were obtained that occurrence frequency of terms gradually decreased in proportion to length of the character type sequence, and there were few terms with short sequence pattern which have a head word with the character types except for Kanji or Katakana.

1. はじめに

Part.1[2]では、多字種複合語に対して、どのような字種(の用語)から構成されているかという観点から分析した。この観点では、ある複合語を構成する用語の字種の順序、および複数回の同一字種の出現は考慮されない。多字種複合語集合に対して、字種の観点からより詳細な特性を明らかにするためには、連接する構成単語の字種の順序、出現回数も分析の対象とすることが必要である。例えば、「データベース管理システム」は、字種構成では漢字とカタカナの2字種であるのに対し、字種の出現順序としては、”カタカナ”、 ”漢字”、 ”カタカナ”、となる。字種を以下に挙げる記号で表現すると、”KJK”となる。ここでは、字種の出現順序を字種変化とよぶ。筆者らは、複数の辞書を対象に見出し語から多字種複合語を抽出し、字種変化の観点から様々な項目についての分析結果を報告した[1]。part.2では同様の項目について分析を行う。

2. 用語解析手順

本研究part.1[2]で抽出された複合語13万6千語を対象とする．Part.1と同様に字種は以下の9種類に分類し，それぞれを1文字のコード‐字種コードとして表記する．対象複合語に対して字種判別を行い，字種コードを付与する。得られた字種コードデータについて。用語出現頻度，相対出現比率，累積出現頻度，累積出現比率を算出する。

⑴	全角漢字	J	⑹	全角数字	N
⑵	全角カタカナ	K	⑺	半角数字	n
⑶	全角ひらがな	H	⑻	全角記号	S
⑷	全角英字	A	⑼	半角記号	s
⑸	半角英字	a			

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

3. 結果

3.1 用語全体

表 1　字種変化数毎の用語数

変化数	用語数	比率 (%)	累積	累積比率 (%)
2	66423	48.85	66423	48.85
3	40710	29.94	107133	78.79
4	16112	11.85	123245	90.64
5	6219	4.57	129464	95.21
6	3002	2.21	132466	97.42
7	1467	1.08	133933	98.50
8	959	0.71	134892	99.21
9	511	0.38	135403	99.58
10	258	0.19	135661	99.77
11	116	0.09	135777	99.86

表1は字種変化数毎の複合語の出現頻度を示している．この表から，対象複合語全体に対して、変化数2～5までの用語で累積比率は95%に達しており，6変化以上の複合語は非常に少ないことがわかる。

対象複合語全体では、字種変化パターンは全4433種あった。上位171種で全複合語の90%，551種で95%を含んでいることが判明した．表2は用語数の多い字種変化パターン上位20位までを列挙している．この表から上位20位までの変化パターンで、全体の71%を含むことがわかる。これはパターン総数の12%に過ぎない。残り78%のパターンは用語全体の5%程度しか出現しないことが分かる．

またこの表から分かるように，必ずしも字種変化数の少ないものが上位に来るとは限らず，3変化や4変化，例えばJKN(3位)，JKJN(11位)などの変化パターンを採る用語は非常に多いことがわかる．

表 2　字種変化パターン毎の用語数

パターン	用語数	比率 (%)	累積	累積比率 (%)
JN	23115	17.00%	23115	17.00%
KJ	14131	10.39%	37246	27.39%
JK	10388	7.64%	47634	35.03%
JKN	7726	5.68%	55360	40.71%
KN	6893	5.07%	62253	45.78%
KJN	6575	4.84%	68828	50.62%
JKJ	6063	4.46%	74891	55.08%
JNA	3507	2.58%	78398	57.66%
JA	2670	1.96%	81068	59.62%
JHJ	2310	1.70%	83378	61.32%
JKJN	2208	1.62%	85586	62.94%
Jn	2066	1.52%	87652	64.46%
KJK	1519	1.12%	89171	65.58%
JNSN	1414	1.04%	90585	66.62%
AJ	1291	0.95%	91876	67.57%
JNJ	1233	0.91%	93109	68.48%
JHJN	981	0.72%	94090	69.20%
JH	904	0.66%	94994	69.86%
KNA	739	0.54%	95733	70.41%
Kn	732	0.54%	96465	70.94%

3.2 字種変化数毎の結果

以下の節では，変化数2～6までの用語集合それぞれについて，字種変化パターンの観点から分析する．

3.2.1 変化数2

表3は変化数2を採る用語集合に対する用語頻度である．変化数2のパターンは58種存在した．上位6種で2変化の用語総数に対して累積90%，上位11種で累積95%に達する．つまり全パターン中，上位約19%のパターンが約半数を占める．また変化数2は全体の用語の中で最も多い．全体比率で見ても，変化数2の累積95%は全体の46%に当たる．そのため上記の出現頻度の偏りは用語全体に対して大きな意味を持つものである．

表 3　字種変化数2のパターン出現頻度

パターン	用語数	比率 (%)	累積	累積比率 (%)	全体比率 (%)
JN	23115	34.80	23115	34.80	17.00
KJ	14131	21.27	37246	56.07	27.39
JK	10388	15.64	47634	71.71	35.03
KN	6893	10.38	54527	82.09	40.10
JA	2670	4.02	57197	86.11	42.07
Jn	2066	3.11	59263	89.22	43.58
AJ	1291	1.94	60554	91.16	44.53
JH	904	1.36	61458	92.53	45.20
Kn	732	1.10	62190	93.63	45.74
NJ	594	0.89	62784	94.52	46.17

3.2.2　変化数3

表 4　字種変化数3のパターン出現頻度

パターン	用語数	比率 (%)	累積	累積比率 (%)	全体比率 (%)
JKN	7726	18.98	7726	18.98	5.68
KJN	6575	16.15	14301	35.13	10.52
JKJ	6063	14.89	20364	50.02	14.98
JNA	3507	8.61	23871	58.64	17.56
JHJ	2310	5.67	26181	64.31	19.25
KJK	1519	3.73	27700	68.04	20.37
JNJ	1233	3.03	28933	71.07	21.28
KNA	739	1.82	29672	72.89	21.82
JAN	689	1.69	30361	74.58	22.33
AJN	593	1.46	30954	76.04	22.76

表4は変化数3の用語出現頻度である．変化数3のパターンは276種存在した，上位27種で3変化の用語総数に対して累積90%，上位47種で累積95%に達する．つまり全パターン中，上位約17%のパターンが大半を占める．

3.2.3　変化数4

表5は変化数4の用語数を示したものである．変化数4のパターンは636種存在した．上位115種で4変化の用語総数に対して累積90%，上位212種で累積95%に達する．つまり全パターン中，上位約33%のパターンが大半を占める．

表 5　字種変化数4のパターン出現頻度

パターン	用語数	比率 (%)	累積	累積比率 (%)	全体比率 (%)
JKJN	2208	13.70	2208	13.70	1.62
JNSN	1414	8.77	3622	22.48	2.66
JHJN	981	6.09	4603	28.57	3.39
KJKN	667	4.14	5270	32.70	3.88
JKNA	642	3.98	5912	36.69	4.35
KJNA	617	3.83	6529	40.52	4.80
KNSN	577	3.58	7106	44.10	5.23
KJKJ	548	3.40	7654	47.50	5.63
JNJN	469	2.91	8123	50.41	5.97
JKJK	405	2.51	8528	52.92	6.27

3.2.4　変化数5

表 6　字種変化数5のパターン出現頻度

パターン	用語数	比率 (%)	累積	累積比率 (%)	全体比率 (%)
JHJHJ	540	8.68	540	8.68	0.40
JKNSN	390	6.27	930	14.95	0.68
KJNSN	247	3.97	1177	18.93	0.87
JKJNA	175	2.81	1352	21.74	0.99
JKJKJ	152	2.44	1504	24.18	1.11
JKJKN	124	1.99	1628	26.18	1.20
JHJNA	119	1.91	1747	28.09	1.28
KJHJN	110	1.77	1857	29.86	1.37
KJKJN	107	1.72	1964	31.58	1.44
JNSNS	104	1.67	2068	33.25	1.52

表6は変化数5の用語出現頻度である．変化数5のパターンは全914種存在した．上位366種で5変化の用語総数に対して累積90%，上位603種で累積95%に達する．これは全パターンの約66%である．

3.2.5　変化数6

表7は変化数6の用語出現頻度である．変化数6のパターンは全836種存在した．，

上位537種で6変化の用語総数に対して累積90%，上位686種で累積95%に達する．これは全パターンの約82%である．

表 7　字種変化数6のパターン出現頻度

パターン	用語数	比率 (%)	累積	累積比率 (%)	全体比率 (%)
JHJHJN	202	6.73	202	6.73	0.15
JNASNA	135	4.50	337	11.23	0.25
JNSNSN	122	4.07	459	15.30	0.34
JNSNJS	76	2.53	535	17.83	0.39
JHJHKN	72	2.40	607	20.23	0.45
JKJNSN	69	2.30	676	22.53	0.50
KNSNSN	55	1.83	731	24.37	0.54
KNASNA	53	1.77	784	26.13	0.58
JANSAN	50	1.67	834	27.80	0.61
KJHJHJ	48	1.60	882	29.40	0.65

3.2.6　変化数毎の累積分布

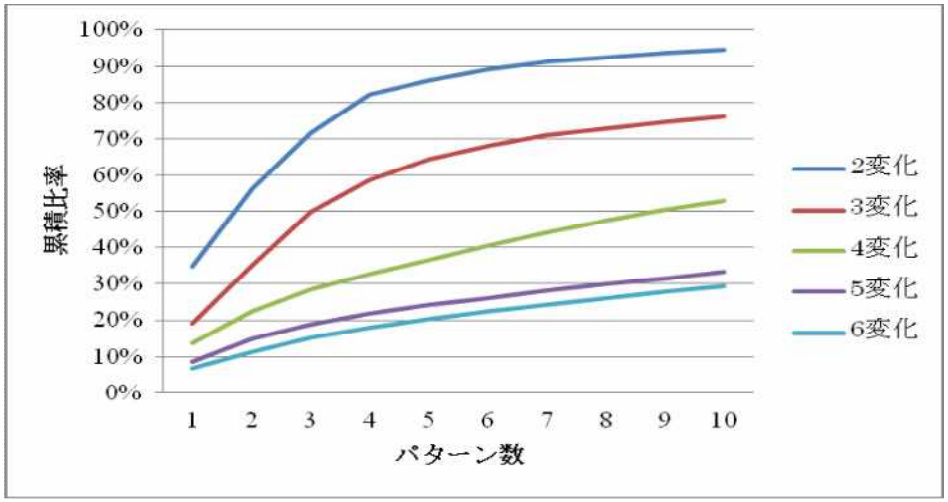

図 1　変化数毎の累積分布

図1は，変化数6までの各変化数の上位10種における累積比率をグラフにしたものである．グラフから，変化数が増えるに従って，累積比率の上昇は緩やかになることが分かる．

4.　考察

先頭字種毎に変化数毎の出現頻度および変化パターン出現頻度を分析する．また，以下の節では，変化数2～6までの用語集合それぞれについて，字種変化パターンの観点から分析する．

4.1　漢字

表 8　先頭字種漢字・変化数

変化数	用語数	比率 (%)	累積	累積比率 (%)
2	39424	48.49	39424	48.49
3	25815	31.75	65239	80.24
4	9751	11.99	74990	92.23
5	3429	4.22	78419	96.45
6	1685	2.07	80104	98.52

表8は，先頭字種が漢字の変化数毎の用語頻度である．変化数5の段階で累積96%と大半を占める．先頭字種漢字はそもそも用語全体の6割近くを占めるため，その累積比率は全体の傾向とほぼ同様の結果となる．

表9は，先頭字種漢字の変化パターン毎の出現頻度で，上位20種までを列挙している．全1490種あり，先頭の用語(単語)が漢字である全用語に対して，上位40種で累積90%に達することがわかる．表9には列挙されていないが，上位99で累積95%を占める．

表9から最上位“JN”だけで約30%を占めることがわかるが，“JN”は“亜鉛被膜７”のような複合語であり，漢字熟語に特許特有の連番表現が付加されるいる用語が多いと推測される．また3位にある“JKN”(“印刷機ドライヤ７”など）も同様に連番が付加された用語が多いと推測される．

表 9　先頭字種漢字・変化パターン

パターン	用語数	比率 (%)	累積	累積比率 (%)
JN	23115	28.43	23115	28.43
JK	10388	12.78	33503	41.20
JKN	7726	9.50	41229	50.71
JKJ	6063	7.46	47292	58.16
JNA	3507	4.31	50799	62.48
JA	2670	3.28	53469	65.76
JHJ	2310	2.84	55779	68.60
JKJN	2208	2.72	57987	71.32
Jn	2066	2.54	60053	73.86
JNSN	1414	1.74	61467	75.60
JNJ	1233	1.52	62700	77.11
JHJN	981	1.21	63681	78.32
JH	904	1.11	64585	79.43
JAN	689	0.85	65274	80.28
JKNA	642	0.79	65916	81.07
JKn	576	0.71	66492	81.78
JHJHJ	540	0.66	67032	82.44
JNJN	469	0.58	67501	83.02
JKA	454	0.56	67955	83.58
JAa	412	0.51	68367	84.08

4.2 カタカナ

表 10　先頭字種カタカナ・変化数

変化数	用語数	比率 (%)	累積	累積比率 (%)
2	22375	56.07	22375	56.07
3	11367	28.48	33742	84.55
4	3806	9.54	37548	94.09
5	1127	2.82	38675	96.91
6	573	1.44	39248	98.35

表10は，先頭字種がカタカナである複合語の変化数毎の頻度である．変化数5の段階で累積96%に達する．この傾向は先頭字種が漢字である複合語と同様の増加傾向をであることがわかる．

表 11　先頭字種カタカナ・変化パターン

パターン	用語数	比率 (%)	累積	累積比率 (%)
KJ	14122	35.39	14122	35.39
KN	6891	17.27	21013	52.65
KJN	6573	16.47	27586	69.13
KJK	1515	3.80	29101	72.92
KNA	735	1.84	29836	74.76
Kn	732	1.83	30568	76.60
KJKN	667	1.67	31235	78.27
KJNA	617	1.55	31852	79.82
KNSN	575	1.44	32427	81.26
KJA	552	1.38	32979	82.64
KA	551	1.38	33530	84.02
KJKJ	548	1.37	34078	85.39
KJn	457	1.15	34535	86.54
KNJ	384	0.96	34919	87.50
KAN	320	0.80	35239	88.30
KJHJ	256	0.64	35495	88.94
KJNSN	247	0.62	35742	89.56
KJH	191	0.48	35933	90.04
KJNJ	161	0.40	36094	90.45
KJAN	116	0.29	36210	90.74

表11は，先頭字種がカタカナである複合語の変化パターン毎の頻度である．先頭字種がカタカナで始まる字種変化バターンは746種存在した．表11から上位18種のパターンで累積90%に達することがわかる．表11には列挙されていないが，上位54種で累積95%になることがわかった．

先頭字種漢字同様，上位には“KN”(“アーク５”など）や，“KJN”(“ロ

ール紙２２”など）のように末尾に連番が付いた用語が散見される．このような特徴は，既述したように漢字で始まる用語と同様の特徴を有している．

4.3 ひらがな

表 12　先頭字種ひらがな・変化数

変化数	用語数	比率 (%)	累積	累積比率 (%)
2	285	51.17	285	51.17
3	159	28.55	444	79.71
4	74	13.29	518	93.00
5	28	5.03	546	98.03
6	5	0.90	551	98.92

表 13　先頭字種ひらがな・変化パターン

パターン	用語数	比率 (%)	累積	累積比率 (%)
HJ	233	41.83	233	41.83
HJN	97	17.41	330	59.25
HN	27	4.85	357	64.09
HJHJ	18	3.23	375	67.32
HK	16	2.87	391	70.20
HKN	15	2.69	406	72.89
HJNA	14	2.51	420	75.40
HJH	10	1.80	430	77.20
HJn	9	1.62	439	78.82
HJK	8	1.44	447	80.25
HKJ	8	1.44	455	81.69
HJKN	8	1.44	463	83.12
HJHJN	7	1.26	470	84.38
Hn	6	1.08	476	85.46
HNA	6	1.08	482	86.54
HNSN	5	0.90	487	87.43
HJA	4	0.72	491	88.15
HJHN	4	0.72	495	88.87
HJNSN	4	0.72	499	89.59
HA	3	0.54	502	90.13

表12は，先頭字種がひらがなで始まる複合語の変化数毎の頻度を示している．先頭字種がひらがなの複合語はpart.1の表3から全体の約0.4%と全体に対する割合が低く，用語数が少ない．さらに変化数6以上の用語は極めて少ないことがわかる．

先頭字種がひらがなで始まる複合語の変化パターンは60種存在した．表13は，先頭字種がひらがなで始まる複合語の変化パターンの上位20位までを列挙したものである．上位20種で累積90%に達することがわかる．この表には示されないが，上位30で累積95%になることが判明した．最上位パターン“HJ”とそれ以外のパターンの偏差が非常に大きいことがわかる．.“HJ”の実例として“いねむり運転”が挙げられる．

4.4 半角英字

表 14　先頭字種半角英字・変化数

変化数	用語数	比率 (%)	累積	累積比率 (%)
2	231	39.15	231	39.15
3	153	25.93	384	65.08
4	83	14.07	467	79.15
5	50	8.47	517	87.63
6	23	3.90	540	91.53

表14は，先頭字種が半角英字で始まる複合語の変化数毎の頻度である．先頭字種が半角英字で始まる複合語は，ひらがな同様全体に対する割合が約0.4%と低い．変化数6の時点で累積91%であり，今までの日本語3字種と比べると，緩やかな上昇となっている．

先頭字種が半角英字で始まる複合語の変化パターンは176種あった．表15は，先頭字種が半角英字で始まる複合語の変化パターンの上位20位までを列挙したものである．この表には示されないが，上位117種で累積90%，上位147で累積95%に達する．変化数同様今までの日本語3字種に比べ，累積比率の上昇は緩やかである．最上位のパターンでも17.46%と比率はあまり高くない．

表 15　先頭字種半角英字・変化パターン

パターン	用語数	比率 (%)	累積	累積比率 (%)
aJ	103	17.46	103	17.46
an	78	13.22	181	30.68
aK	22	3.73	203	34.41
asa	21	3.56	224	37.97
anJ	15	2.54	239	40.51
asn	13	2.20	252	42.71
aJN	12	2.03	264	44.75
aKJ	11	1.86	275	46.61
ana	11	1.86	286	48.47
anan	11	1.86	297	50.34
anSan	11	1.86	308	52.20
as	9	1.53	317	53.73
asaJ	9	1.53	326	55.25
aN	8	1.36	334	56.61
aSa	8	1.36	342	57.97
aAJ	7	1.19	349	59.15
aA	6	1.02	355	60.17
aJn	6	1.02	361	61.19
aKN	6	1.02	367	62.20
aKn	6	1.02	373	63.22

4.5 半角数字

表 16　先頭字種半角数字・変化数

変化数	用語数	比率 (%)	累積	累積比率 (%)
2	188	19.09	188	19.09
3	159	16.14	347	35.23
4	179	18.17	526	53.40
5	144	14.62	670	68.02
6	59	5.99	729	74.01

表16は，先頭字種が半角数字で始まる複合語の変化数毎の頻度である．半角英字同様，日本語字種に比べて累積上昇率は低い．またこれまでの字種とは異なり，変化数による単調増加ではない．これは先頭半角英字の複合語の多くは数値範囲表現や単位記号を付加した数値，例えば“ 0〜150℃ ”がこの特性をもつ用語である．

表 17　先頭字種半角数字・変化パターン

パターン	用語数	比率 (%)	累積	累積比率 (%)
na	73	7.41	73	7.41
nsn	56	5.69	129	13.10
nsnSnsnS	54	5.48	183	18.58
nSnS	52	5.28	235	23.86
nA	43	4.37	278	28.22
nS	31	3.15	309	31.37
nSn	31	3.15	340	34.52
nJ	25	2.54	365	37.06
nsnSnsn	25	2.54	390	39.59
nSnJ	21	2.13	411	41.73
nSnJS	20	2.03	431	43.76
nSna	19	1.93	450	45.69
nsna	18	1.83	468	47.51
nsnS	15	1.52	483	49.04
nK	14	1.42	497	50.46
nsnSnsns	13	1.32	510	51.78
nsnSnsnJS	13	1.32	523	53.10
nsnsn	11	1.12	534	54.21
nsnSnsnaS	10	1.02	544	55.23
nsnJS	8	0.81	552	56.04

先頭字種が半角数字で始まる複合語の変化パターンは250種あった．表17は先頭字種が半角数字で始まる複合語の変化パターンの上位20位までを列挙したものである．この表には示されないが，上位152種で累積90%，上位201で累積95%

となった．表17で変化数8の“nsnSnsnS”（“3.0～8.5%”など）が第3位に挙がっている点が特徴的である．他の字種では，上位パターンは変化数が少ない傾向にあるのに対して,半角数ではより変化数の多いパターンが上位に多く出現する．

4.6 半角記号

表 18　先頭字種半角記号・変化数

変化数	用語数	比率 (%)	累積	累積比率 (%)
4	5	26.32	5	26.32
2	4	21.05	9	47.37
3	4	21.05	13	68.42
5	2	10.53	15	78.95
12	2	10.53	17	89.47
6	1	5.26	18	94.74
8	1	5.26	19	100.00

表 19　先頭字種半角記号・変化パターン

パターン	用語数	比率 (%)	累積	累積比率 (%)
Sa	3	15.79	3	15.79
sana	2	10.53	5	26.32
sasJ	2	10.53	7	36.84
sAsAs	2	10.53	9	47.37
sA	1	5.26	10	52.63
sKJ	1	5.26	11	57.89
saN	1	5.26	12	63.16
San	1	5.26	13	68.42
Sas	1	5.26	14	73.68
sASs	1	5.26	15	78.95
sanans	1	5.26	16	84.21
sasansas	1	5.26	17	89.47
sasansansans	1	5.26	18	94.74
sasansasasan	1	5.26	19	100.00

表18は，先頭字種半角記号の変化数毎の出現頻度である．先頭字種半角記号は用語数が極めて少ないため，特性の分析や他の字種との比較は難しい．

先頭字種が半角記号で始まる複合語の変化パターンは14種あった．表19は，全パターンを列挙しており，上位12種で累積90%，上位13で累積95%になる．最長パターンのsasansasasan は“-CO-R3-CO-NH-R6”である．先頭字種が半角記号で始まる文字列の多くはこのような化学構造式である．また先頭字種が半角記号で始まる文字列は，わずか19語と極めて少なく，全体のばらつきが非常に顕著であり，ほぼ全てユニークなパターンである．

4.7 全角記号

表 20　先頭字種全角記号・変化数

変化数	用語数	比率 (%)	累積	累積比率 (%)
2	149	28.01	149	28.01
3	156	29.32	305	57.33
4	89	16.73	394	74.06
5	64	12.03	458	86.09
6	25	4.70	483	90.79

表20は，先頭字種が全角記号で始まる複合語の変化数毎の頻度である．半角記号と比べると,“α”,“β”のような意味単位を形成する記号が多く存在するため，変化数増加による累積比率の上昇に乱れは見られない．

先頭字種が半角記号で始まる複合語の変化パターンは179種あった．表21は，上位20まで変化パターン毎の頻度を示している．この表には示されないが，上位126種で累積90%，上位152で累積95%に達する．上位のパターンは“ΔＴ”(“SA”)のような数式変数表現や,“γ線入射端面”(“SJ”)のような専門用語である．また，半角記号同様“－ＣＨ２－Ｏ－”(“SANSAS”)のような化学構造式も多く出現する．

表 21　先頭字種全角記号・変化パターン

パターン	用語数	比率 (%)	累積	累積比率 (%)
SA	33	6.20	33	6.20
Sa	31	5.83	64	12.03
SJ	29	5.45	93	17.48
SK	29	5.45	122	22.93
SKJ	23	4.32	145	27.26
SN	22	4.14	167	31.39
SAn	18	3.38	185	34.77
SAJ	12	2.26	197	37.03
SAa	12	2.26	209	39.29
SJN	12	2.26	218	40.98
SNJ	9	1.69	230	43.23
SAN	8	1.50	238	44.74
SAS	8	1.50	246	46.24
SNA	8	1.50	254	47.74
SNS	8	1.50	262	49.25
SNSNS	8	1.50	270	50.75
SJK	7	1.32	277	52.07
SASA	7	1.32	284	53.38
SNSN	6	1.13	290	54.51
SKJKJ	6	1.13	296	55.64

4.8 全角英字

表 22　先頭字種全角英字・変化数

変化数	用語数	比率 (%)	累積	累積比率 (%)
2	2631	35.61	2631	35.61
3	2019	27.33	4650	62.94
4	1140	15.43	5790	78.37
5	747	10.11	6537	88.48
6	354	4.79	6891	93.27

表22は，先頭字種が全角英字で始まる複合語の変化数毎の頻度である．英数字や記号で始まる複合語と比較すると，変化数6の時点で約93%であり累積上昇率は高い．これは全角英字が日本語字種との同様の使われ方をされていることを示唆していると考えられる．

表 23　先頭字種全角英字・変化パターン

パターン	用語数	比率 (%)	累積	累積比率 (%)
AJ	1291	17.47	1291	17.47
AJN	593	8.03	1884	25.50
AN	541	7.32	2425	32.82
AK	415	5.62	2840	38.44
AKN	329	4.45	3169	42.89
AKJ	239	3.23	3408	46.13
AJK	182	2.46	3590	48.59
Aa	176	2.38	3766	50.97
An	173	2.34	3939	53.32
AKJN	108	1.46	4047	54.78
AJKN	87	1.18	4134	55.96
AKAJ	83	1.12	4217	57.08
ASA	74	1.00	4291	58.08
ASAJ	70	0.95	4361	59.03
AKA	69	0.93	4430	59.96
ANJ	69	0.93	4499	60.90
AJKJN	69	0.93	4568	61.83
AJKJ	63	0.85	4631	62.68
ANA	59	0.80	4690	63.48
ANSN	58	0.79	4748	64.27

先頭字種が全角英字で始まる複合語の変化パターンは858種あった．表23は，先頭字種が全角英字で始まる複合語の変化パターンの上位20位までを列挙したものである．この表には示されないが，上位216種で累積90%，上位488で累積95%になる．上位パターンは“ＡＦＣ制御回路”(“AJ”)や，末尾に連番が付加

された“ＣＲＴ表示装置１”(“AJN”)語が散見された．表22，23から全角英字は用語数の増加傾向は，“AJ”や“AK”のように変化数が少ないパターンが上位に位置しており，日本語字種とそれと類似している．この事実から，全角英字は日本語字種と同様な造語能力を有すると推定できるであろう．

4.9 全角数字

表 24　先頭字種全角数字・変化数

変化数	用語数	比率 (%)	累積	累積比率 (%)
2	1133	24.21	1133	24.21
4	984	21.03	2117	45.24
3	875	18.70	2992	63.93
5	632	13.50	3624	77.44
7	319	6.82	3943	84.25

表24は，先頭字種が全角数字で始まる複合語の変化数毎の頻度である．変化数の増加による累積比率の上昇は緩やかである．これは先頭字種が半角数字で始まる複合語と同様，数値範囲表現などが多く出現するために，変化数が多くなりやすい傾向にあることに起因と推測される．

先頭字種が全角数字で始まる複合語の変化パターンは669種あった．表25は，先頭字種が全角数字のので始まる複合語の変化パターンの上位20位までを列挙したものである．この表には示されないが，上位240種で累積90%，上位435で累積95%に達する．先頭字種が半角数字で始まる複合語とは異なり，後方位置に漢字を含むパターンが最上位に挙がっている．これは，例えば，“２次空気供給装置”のような複合語で，全角数字による序数詞ないし助数詞表現が多いためである．また“４．５～７．０”(NSNSNSN）のような数値範囲表現も散見された．

表 25　先頭字種全角数字・変化パターン

パターン	用語数	比率 (%)	累積	累積比率 (%)
NJ	594	12.69	594	12.69
NA	252	5.38	846	18.08
NSNS	226	4.83	1072	22.91
NSN	180	3.85	1252	26.75
NK	164	3.50	1416	30.26
NSNJ	128	2.74	1544	32.99
NKJ	126	2.69	1670	35.68
NSNSNSN	104	2.22	1774	37.91
NJK	99	2.12	1873	40.02
NSNA	99	2.12	1972	42.14
NJN	88	1.88	2060	44.02
NSNSNSNS	85	1.82	2145	45.83
NSNJS	84	1.79	2229	47.63
NAJ	78	1.67	2307	49.29
NJKJ	65	1.39	2372	50.68
Na	60	1.28	2432	51.97
NSNK	58	1.24	2490	53.21
NS	52	1.11	2542	54.32
NSNSA	51	1.09	2593	55.41
NSJ	45	0.96	2638	56.37

4.10 先頭字種毎のパターン数と相対比率

表26は先頭字種毎の変化パターン数の累積90%，95%，および100%到達時のパターン数を示したものである．相対比率は式（1）で算出する．

$$\text{相対比率} = \frac{\text{到達時パターン数}}{\text{字種毎のパターン総数}} \qquad (1)$$

表26で黄色で網掛けした値が，先頭字種ではじまる用語の字種変化バターン

総数を示してる．相対比率の値が小さいほど，ごく僅かの字種変化パターンで表記される複合語が非常に多いことを表している．具体的には，漢字およびカタカナで始まる複合語はこの傾向が顕著である．一方，相対比率の値が大きいほど，様々な字種変化パターンで表記される複合語が多いことを示している．半角または全角の記号，半角数字で始まる複合語はこの傾向が強い．言い換えれば，これらの用語は字種変化の観点から多様な表記を採る用語が相対的に多いと言えよう．

表 26　先頭字種毎のパターン数と相対比率

先頭字種	割合 (%)	パターン	相対比率 (%)	先頭字種	割合 (%)	パターン	相対比率 (%)
漢字	90	40	2.68	全角記号	90	126	70.39
	95	99	6.64		95	152	84.92
	100	1490	100		100	179	100
カタカナ	90	18	2.41	全角数字	90	240	35.87
	95	54	7.24		95	435	65.02
	100	746	100		100	669	100
ひらがな	90	20	33.33	全角英字	90	216	25.17
	95	30	50		95	488	56.88
	100	60	100		100	858	100
半角数字	90	152	60.8	半角記号	90	12	85.71
	95	201	80.4		95	13	92.86
	100	250	100		100	14	100
半角英字	90	117	66.48				
	95	147	83.52				
	100	176	100				

4.11 先頭字種毎の変化パターン累積分布

図2は先頭字種毎の変化パターンの累積出現頻度をグラフ化したものである．縦軸は先頭字種毎の用語数に対する累積出現比率であり，横軸は4.10の式（1）で表されるパターン総数に対する相対比率である．

このグラフから漢字およびカタカナで始まる複合語は他の字種で始まる複合語に対して，より少ないパターンで累積90%を超えることが分かる．次いで少ないのが，全角英字，全角数字，ひらがなの3種であり，それ以外の半角文字や全角記号は累積90%の到達が遅い．以上の結果から，特許抄録においては全角文字の方が高頻度の字種変化パターンで表される複合語が多いということが分かる．

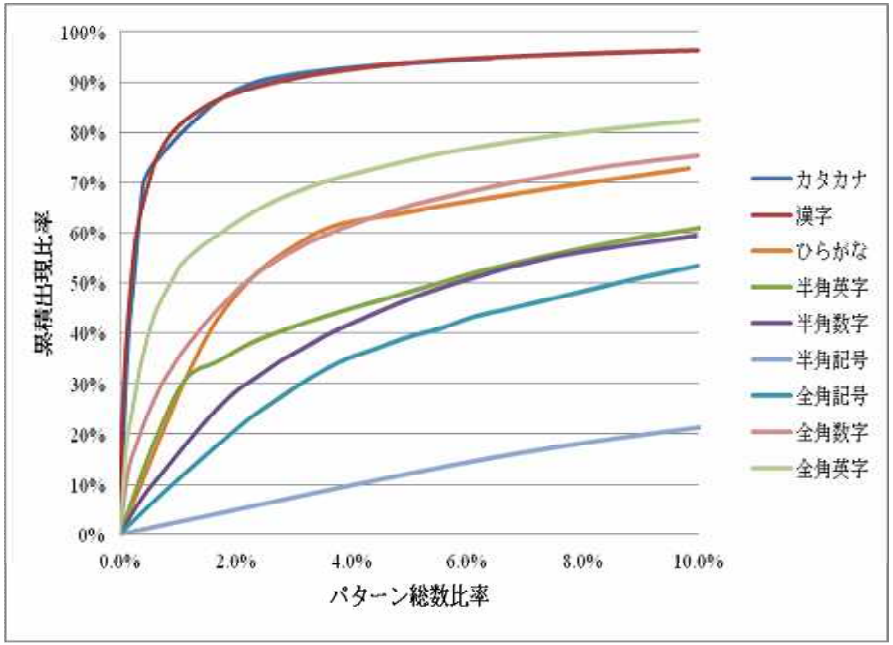

図 2　先頭字種毎の変化パターン累積分布

5. 終わりに

本研究の結果から，特許抄録などの専門分野の文書における多字種複合語の特性が明らかになった．変化数毎に見ると，変化数の増加に伴い出現比率は単調に減少することが分かった．またパターンの詳細を見ると，上位のパターンは必ずしも変化数に依存せず，より変化数の多いパターンが上位に挙がることもある．先頭字種毎に変化特性を比べると，意味単位を形成することの出来る漢字・カタカナ・ひらがなといった日本語字種は変化数・変化パターンの総種類数が他の字種に比べて少なくなる．

本研究の成果で発見された特性の一部は日本語全体に対して適用可能であると考えられる．また，文書毎の特性を変化パターンの解析という形で明らかにすることで，形態素解析やチャンキングの精度向上につながる可能性があると考えられる．

謝辞

NTCIR-4特許検索テストコレクションは国立情報学研究所(NII) の許可を得て使用させて頂きました。この場を借りて深謝いたします。

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．自然言語処理研究会報告 2011-NL-204

Vol.2011-NL-204 No.3【正誤表】
p.8 表 19 先頭字種半角記号・変化パターン

§4.6 表 19 先頭字種半角記号・変化パターン

正

パターン	用語数	比率 (%)	累積	累積比率 (%)
sa	3	15.79	3	15.79
sana	2	10.53	5	26.32
sasJ	2	10.53	7	36.84
sAsAs	2	10.53	9	47.37
sA	1	5.26	10	52.63
sKJ	1	5.26	11	57.89
saN	1	5.26	12	63.16
san	1	5.26	13	68.42
sas	1	5.26	14	73.68
sASs	1	5.26	15	78.95
sanans	1	5.26	16	84.21
sasansas	1	5.26	17	89.47
sasansansans	1	5.26	18	94.74
sasansasasan	1	5.26	19	100

誤

パターン	用語数	比率 (%)	累積	累積比率 (%)
Sa	3	15.79	3	15.79
sana	2	10.53	5	26.32
sasJ	2	10.53	7	36.84
sAsAs	2	10.53	9	47.37
sA	1	5.26	10	52.63
sKJ	1	5.26	11	57.89
saN	1	5.26	12	63.16
San	1	5.26	13	68.42
Sas	1	5.26	14	73.68
sASs	1	5.26	15	78.95
sanans	1	5.26	16	84.21
sasansas	1	5.26	17	89.47
sasansansans	1	5.26	18	94.74
sasansasasan	1	5.26	19	100