Vol.2011-NL-202 No.1
2011/7/15

大規模複合語データに対する構成字種解析

滝川 諒†，後藤 智範
†神奈川大大学院理学研究科，神奈川大学理学部情報科学科

日本語の学術文献のテキストにおいて，主要な概念，テーマは多字種複合語で表現されることが多い．特に学術論文，特許明細書などの専門性の高い文書では，外来語などをそのまま用いることも多く，長単位の複合語表現が文章中に多々出現する．本研究は，複数の辞書の数十万語に上る膨大な見出し語から，2字種以上からなる多字種複合語12万語を抽出し，これをコーパスとして複合語の字種特性，具体的には字種構成および字種変化パターンについて分析した．
結果として，当該コーパスに出現する多字種複合語は，約90%が2字種から構成されることが明からなった．また，字種変化パターンでは，90%の用語が上位26種で表現されていることが判明した．

Analysis to Japanese Compound Terms with Multi Character Types Extracted from Lots of Entry Terms of Several Dictionaries

Ryo Takikawa†, Tomonori Gotoh

†Graduate School of Science, Kanagawa University
Department of Information and Computer Sciences, Kanagawa University

Japanese compound terms or noun phrase are used to explain key concepts or themes in Japanese academic or technical texts. Lots of long compound terms are consisted with multi character types, not with singe character type. This paper reports the results and considerations on analysis to large scale Japanese compound terms with multi character types. The term collection contains about 120 thousands terms which were extracted from a large quantity of entry terms of seven dictionaries. These compound terms were analyzed in terms of the structure and the sequence patterns of character types.

As a result, it was found that 90% of these compound terms were constructed with the two kinds of character types, and expressed with the higher ranked 26 kinds of sequence patterns of character types. These results could be used to improvement for accuracy to morphological analysis – chunking.

1. はじめに

日本語のテキストにおいて，主要な概念，テーマは多字種複合語で表現されることが多い．特に学術論文などの専門性の高い文章においては外来語などをそのまま用いることも多く，長単位の複合語表現が文章中に多々出現する．これらの専門用語に関して様々な研究がなされてきた．対象となるデータは形態素解析機等で単語レベルに分割されたテキストから統計を用いて抽出を行う研究[1][2][3][4][5]や，日本語文法に従って抽出された用語に対する研究[6][7]ある．しかしこれらの研究は抽出対象となるコーパスサイズは小規模で，抽出される複合語数はそれほど多くない．また文字種に着目した研究[8]も過去にあるが，ひらがなを対象にしないなどの問題があった．

本研究は，複数の辞書の見出し語を対象とし，複合語の特性，具体的には字種に着目した字種構成および字種変化パターンについて明らかにしようとするものである．

2. コーパスおよび解析手順

2.1 コーパス

表2.1に挙げる7種類の辞書[9]の見出し語のうち、2字種以上からなる見出し語を対象とした．

これは，例えば，「文献情報ネットワーク」や「F型シナプス」などが対象となる．逆に単字種構成となる「国際航空通信協会」や「コンパイラ」などの用語は対象から除外した．

また辞書中には以下に挙げるような見出し語として扱うことの出来ない言葉（ノイズ）が多く存在した．

表2.1 辞書

辞書名[出版年]	見出し語
25万語医学用語大辞典[1991]	216,714
EB科学技術用語大辞典[1991]	249,830
コンピュータ用語辞典[1990]	24,305
三省堂外来語[1997]	34,782
三省堂国語辞典[1997]	83,380
電気電子用語辞典（英和）[1997]	72,858
電気電子用語辞典（和英）[1997]	79,550

(1) 非見出し語（文章）：「AB型ABO式血液型分類法による血液型の1つ．」
(2) カンマ，中点での連接：「命令アドレスレジスタ，逐次制御レジスタ」
(3) 括弧を含むもの：「特定アクセス許可（指定）」
(4) 助詞および助詞相当語を含むもの：「プログラムによる停止」

これらは全て単字種と同様対象から除外した．上記規則に基づいて抽出を行い，124,194語を分析対象データとした．

2.2 解析手順

構成字種は以下の9種類に分類し，それぞれを1文字のコードとして表記する．

抽出された複合語に対して，上記字種分類に基づき字種判別を行い，字種構成および字種変化パターンについて分析を行う．デ

ータは用語数，相対比率，累積，累積の相対比率について報告する．

(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			

3. 結果

3.1 字種構成

はじめに字種構成の結果を示す．字種構成とは，用語がどのような字種で構成されているのかを表現したものであり，字種の並びには関係しない．

表3.1 構成字種数毎の用語数

字種数	用語数	比率	累積	累積比率
2	111,112	89.47%	111,112	89.47%
3	11,532	9.29%	122,644	98.75%
4	1,409	1.13%	124,053	99.89%
5	140	0.11%	124,193	100.00%
6	1	0.00%	124,194	100.00%

表3.1は構成字種数毎の統計データである．この表から，構成字種はたかだか6字種であることが分かる．また，字種数2,3で累積比率98%と大半を占め，辞書の見出し語では，多字種複合語の構成字種数は多岐には渡らないことを示している．

表3.2は字種構成パターン毎の累積比率95%までの統計データである．表2から上位95%までの用語は全て漢字（J）またはカタカナ（K）を含んでいることが分かる．これは日本語の複合名詞を構成する主成分が漢字もしくはカタカナであることを示唆していると考えられる．

表3.2 字種構成毎の用語数

パターン	語数	比率	累積	累積比率
JK	76,479	61.58%	76,479	61.58%
HJ	20,992	16.90%	97,471	78.48%
HJK	4,005	3.22%	101,476	81.71%
aJ	3,510	2.83%	104,986	84.53%
JKS	2,818	2.27%	107,804	86.80%
nJ	2,746	2.21%	110,550	89.01%
KS	1,848	1.49%	112,398	90.50%
aK	1,632	1.31%	114,030	91.82%
nK	1,401	1.13%	115,431	92.94%
aJK	1,170	0.94%	116,601	93.89%
nJK	896	0.72%	117,497	94.61%
JS	825	0.66%	118,322	95.27%

表3.3は構成字種数毎のパターン出現比率を表したものである．字種構成パターンは，全9種の字種要素の組み合わせになるので，総数はnCrで容易に計算することができる．結果から，出現した字種構成の種類は47%と半数にも満たないことが明らかになった．

表3.3 構成字種数毎の構成パターン出現比率

字種数	出現数	総数	比率
2	17	36	47.2%
3	25	84	29.8%
4	22	126	17.5%
5	10	126	7.9%
6	1	84	1.2%

3.2 字種変化パターン

次に字種変化パターンの結果を示す．字種変化パターンとは，用語内の字種の変化を表したものであり，字種構成と異なり，出現順や複数回出現する字種も全て考慮し，字種コードを用いて用語を表現したものである．

3.2.1 用語全体

表3.4は字種変化数の統計データである．結果から2変化が最も多く，変化数が増加するにつれて用語数は減少する．変化数5の時点で全体の約99%に達することから，複合語の字種変化はあまり多くならないことが分かる．

表3.4 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	84,620	68.135%	84,620	68.135%
3	28,191	22.699%	112,811	90.835%
4	7,950	6.401%	120,761	97.236%
5	2,168	1.746%	122,929	98.981%
6	707	0.569%	123,636	99.551%
7	328	0.264%	123,964	99.815%
8	141	0.114%	124,105	99.928%
9	42	0.034%	124,147	99.962%
10	15	0.012%	124,162	99.974%
11	14	0.011%	124,176	99.986%
12	10	0.008%	124,186	99.994%
13	3	0.002%	124,189	99.996%
14	2	0.002%	124,191	99.998%
15	2	0.002%	124,193	99.999%
17	1	0.001%	124,194	100.000%

表3.5 字種変化パターン毎の用語数

変化数	用語数	比率	累積	累積比率
KJ	40,179	32.35%	40,179	32.35%
JK	23,359	18.81%	63,538	51.16%
JH	8,389	6.75%	71,927	57.92%
JKJ	7,659	6.17%	79,586	64.08%
JHJ	6,957	5.60%	86,543	69.68%
KJK	3,877	3.12%	90,420	72.81%
HJ	3,480	2.80%	93,900	75.61%
aJ	2,591	2.09%	96,491	77.69%
KSKJ	1,775	1.43%	98,266	79.12%
nJ	1,528	1.23%	99,794	80.35%

KSK	1,515	1.22%	101,309	81.57%
Kn	1,342	1.08%	102,651	82.65%
JnJ	1,048	0.84%	103,699	83.50%
JHK	938	0.76%	104,637	84.25%
JHJH	935	0.75%	105,572	85.01%
aK	921	0.74%	106,493	85.75%
KJH	645	0.52%	107,138	86.27%
JKJK	638	0.51%	107,776	86.78%
HJH	555	0.45%	108,331	87.23%
KJKJ	547	0.44%	108,878	87.67%
Ka	530	0.43%	109,408	88.09%
KJHJ	515	0.41%	109,923	88.51%
JaJ	511	0.41%	110,434	88.92%
JSJ	502	0.40%	110,936	89.32%
an	482	0.39%	111,418	89.71%
aSK	472	0.38%	111,890	90.09%

表3.5は字種変化パターンによる用語数を示している．変化パターンは全1040種あり，そのうち上位26種で90%，58種で95%に達する．これは全パターンのうち5%であり，残り95%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．

3.2.2 先頭字種毎の結果

表3.6～3.18は先頭字種に着目し，それぞれ先頭字種毎に各変化パターンの用語数を明らかにしたものである．

表3.6および表3.7は先頭字種が漢字である用語の変化数と変化パターンの統計である．上位のJKの具体例としては，「一般化情報システム」，JHJとして「引張り力特性」が挙げられる．

表3.6　先頭字種漢字・変化数

パターン	用語数	比率	累積	累積比率
2	32,198	59.451%	32,198	59.451%
3	17,936	33.117%	50,134	92.568%
4	2,723	5.028%	52,857	97.596%
5	1,048	1.935%	53,905	99.531%
6	159	0.294%	54,064	99.825%
7	56	0.103%	54,120	99.928%
8	33	0.061%	54,153	99.989%
9	3	0.006%	54,156	99.994%
10	2	0.004%	54,158	99.998%
11	1	0.002%	54,159	100.000%

変化数は2変化が最も多く，長くになるにつれて順に用語数は減少する．パターンは全270種あり，上位7種で累積90%，上位12種で累積95%に達する．このことから，パターン総数は多いが，用語として頻繁に現れるパターンはあまり多くないことが分かる．

表3.8, 表3.9は先頭字種がカタカナである用語の変化数と変化パターンの統計である．「アーク溶接（KJ）」「カード／ディスク変換（KSKJ）」などがある．基本的性質は漢字と同じである．全289種からなり，上位6種で累積90%，上位12種で累積95%に達する．

表3.7　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
JK	23,359	43.13%	23,359	43.13%
JH	8,389	15.49%	31,748	58.62%
JKJ	7,659	14.14%	39,407	72.76%
JHJ	6,957	12.85%	46,364	85.60%
JnJ	1,048	1.93%	47,412	87.54%
JHK	938	1.73%	48,350	89.27%
JHJH	935	1.73%	49,285	91.00%
JKJK	638	1.18%	49,923	92.18%
JaJ	511	0.94%	50,434	93.12%
JSJ	502	0.93%	50,936	94.05%
Ja	368	0.68%	51,304	94.72%
JHJHJ	354	0.65%	51,658	95.38%

表3.8　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	42,470	78.274%	42,470	78.274%
3	7,138	13.156%	49,608	91.430%
4	3,567	6.574%	53,175	98.004%
5	595	1.097%	53,770	99.101%
6	359	0.662%	54,129	99.762%
7	82	0.151%	54,211	99.913%
8	32	0.059%	54,243	99.972%
9	11	0.020%	54,254	99.993%
10	1	0.002%	54,255	99.994%
11	1	0.002%	54,256	99.996%
12	1	0.002%	54,257	99.998%
13	1	0.002%	54,258	100.000%

表3.9　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	40,179	74.05%	40,179	74.05%
KJK	3,877	7.15%	44,056	81.20%
KSKJ	1,775	3.27%	45,831	84.47%
KSK	1,515	2.79%	47,346	87.26%
Kn	1,342	2.47%	48,688	89.73%
KJH	645	1.19%	49,333	90.92%
KJKJ	547	1.01%	49,880	91.93%
Ka	530	0.98%	50,410	92.91%
KJHJ	515	0.95%	50,925	93.86%
KH	383	0.71%	51,308	94.56%
KHJ	193	0.36%	51,501	94.92%
KaJ	193	0.36%	51,694	95.27%

表3.10, 表3.11は先頭字種が半角英字である用語の変化数と変化パターンの用語数を示している．「UPS単位（aJ）」「AND／OR関係（aSaJ）」などがある．基本的性質は前述の2種と同じである．全218種からなり，上位21種で累積90%，上位48種で累積95%に達する．

前述の2種と比較すると，特定の変化パターンによる偏りが少ないという性質が見られた．これは半角英字が日本語複合語の主成分にはならないことが原因であると考えられる．主成分にならない以上，先頭以外の後方部分が意味を持つ構造にならなければならず，そのために変化パターンが多岐に渡ると考えられる．

表3.10　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	4,166	60.83%	4,166	60.83%
3	1,407	20.54%	5,573	81.37%
4	819	11.96%	6,392	93.33%
5	252	3.68%	6,644	97.01%
6	97	1.42%	6,741	98.42%
7	66	0.96%	6,807	99.39%
8	23	0.34%	6,830	99.72%
9	13	0.19%	6,843	99.91%
10	5	0.07%	6,848	99.99%
11	1	0.01%	6,849	100.00%

表3.11　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	2,591	37.83%	2,591	37.83%
aK	921	13.45%	3,512	51.28%
An	482	7.04%	3,994	58.32%
aSK	472	6.89%	4,466	65.21%
aJK	270	3.94%	4,736	69.15%
aSa	205	2.99%	4,941	72.14%
aSaJ	177	2.58%	5,118	74.73%
aS	154	2.25%	5,272	76.97%
aKJ	154	2.25%	5,426	79.22%
aSKJ	142	2.07%	5,568	81.30%
anJ	127	1.85%	5,695	83.15%
aJKJ	90	1.31%	5,785	84.46%
aSaK	65	0.95%	5,850	85.41%
aSanJ	61	0.89%	5,911	86.30%
aSaS	52	0.76%	5,963	87.06%
anK	46	0.67%	6,009	87.74%
aSKJK	41	0.60%	6,050	88.33%
aSJ	35	0.51%	6,085	88.85%
aKaJ	32	0.47%	6,117	89.31%
aJHJ	30	0.44%	6,147	89.75%
asaJ	29	0.42%	6,176	90.17%
aJH	25	0.37%	6,201	90.54%
aSnJ	24	0.35%	6,225	90.89%

表3.12, 表3.13は先頭字種がひらがなである用語の変化数と変化パターンの統計である．「いぼ状皮膚炎（HJ）」「うず巻ばね（HJH）」などがある．変化数についての性質は先にあげたパターンと同様である．全41種からなり，上位4種で累積90%，上位5種で累積95%に達する．先頭ひらがなの用語は後方に漢字ないしカタカナを含むものが圧倒的に多く，全29種のうち，漢字，カタカナを含まないパターンは2種（Ha: 用語数3，HSH: 用語数1）しか存在しなかった．

表3.12　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	3,701	74.32%	3,701	74.32%
3	828	16.63%	4,529	90.94%
4	376	7.55%	4,905	98.49%
5	47	0.94%	4,952	99.44%
6	19	0.38%	4,971	99.82%
7	7	0.14%	4,978	99.96%
8	2	0.04%	4,980	100.00%

表3.13　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	3,480	69.88%	3,480	69.88%
HJH	555	11.14%	4,035	81.02%
HJHJ	247	4.96%	4,282	85.98%
HJK	239	4.80%	4,521	90.78%
HK	218	4.38%	4,739	95.16%

表3.14, 表3.15は先頭字種が半角数字である用語の変化数と変化パターンの用語数を示している．「10進演算子（nJ）」「1‐アミノ‐2‐プロパノール（nSKSnSK）」などがある．変化数についての性質は先にあげたパターンと同様である．全154種からなり，上位20種で累積90%，上位44種で累積95%に達する．先頭半角英字と同様に変化パターンの累積は緩やかな上昇を示す．これは半角数字も複合語の主成分にはなりえないことが原因であると考えられる．

表3.14　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	1,576	52.66%	1,576	52.66%
3	710	23.72%	2,286	76.38%
4	376	12.56%	2,662	88.94%
5	121	4.04%	2,783	92.98%
7	81	2.71%	2,864	95.69%
8	44	1.47%	2,908	97.16%
6	41	1.37%	2,949	98.53%
9	11	0.37%	2,960	98.90%
11	11	0.37%	2,971	99.26%
12	9	0.30%	2,980	99.57%
10	6	0.20%	2,986	99.77%
13	2	0.07%	2,988	99.83%
14	2	0.07%	2,990	99.90%
15	2	0.07%	2,992	99.97%
17	1	0.03%	2,993	100.00%

表3.15　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
nJ	1,528	51.05%	1,528	51.05%
nSK	315	10.52%	1,843	61.58%
nJK	228	7.62%	2,071	69.19%
nSKJ	110	3.68%	2,181	72.87%
nJKJ	76	2.54%	2,257	75.41%
nJnJ	62	2.07%	2,319	77.48%
nSKSnSK	61	2.04%	2,380	79.52%
nKJ	52	1.74%	2,432	81.26%
nK	44	1.47%	2,476	82.73%
nJH	32	1.07%	2,508	83.80%
nSKSnSKJ	31	1.04%	2,539	84.83%
naJ	25	0.84%	2,564	85.67%
nJHJ	25	0.84%	2,589	86.50%
nSJ	23	0.77%	2,612	87.27%
nSKJK	22	0.74%	2,634	88.01%
nKJK	19	0.63%	2,653	88.64%
nSnJ	13	0.43%	2,666	89.07%
nJKJK	12	0.40%	2,678	89.48%
nSKSaSK	12	0.40%	2,690	89.88%
nSaSK	11	0.37%	2,701	90.24%

表3.16, 表3.17は先頭字種が全角記号である用語の変化数と変化パターンの統計である．「α運動細胞（SJ）」「β‐アドレナリン刺激剤（SKJ）」などがある．変化数についての性質は先にあげたパターンと同様である．全66種からなり，上位17種で累積90%，上位29種で累積95%に達する．先頭半角英字，半角数字と同様に，主成分にならないため変化パターンの累積の上昇は緩やかになる．

表3.16　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	508	53.36%	508	53.36%
3	171	17.96%	679	71.32%
5	105	11.03%	784	82.35%
4	88	9.24%	872	91.60%
7	36	3.78%	908	95.38%
6	32	3.36%	940	98.74%
8	7	0.74%	947	99.47%
9	4	0.42%	951	99.89%
10	1	0.11%	952	100.00%

先頭字種が半角記号である用語は，「$a45e」，「補写あり」，「+符号」の3用語だけで，電子辞書からの変換過程での誤りである可能性があり分析対象から除外した．

表3.17　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
SJ	266	27.94%	266	27.94%
SK	226	23.74%	492	51.68%
SKJ	117	12.29%	609	63.97%
SnKSK	67	7.04%	676	71.01%
SJK	31	3.26%	707	74.26%
SnKSKJK	31	3.26%	738	77.52%
SnKJSK	25	2.63%	763	80.15%
SKJK	20	2.10%	783	82.25%
SnSK	17	1.79%	800	84.03%
Sa	13	1.37%	813	85.40%
SaSK	12	1.26%	825	86.66%
SnJ	8	0.84%	833	87.50%
SKSKJ	8	0.84%	841	88.34%
SnaSK	7	0.74%	848	89.08%
SJKJ	6	0.63%	854	89.71%
SnKJK	6	0.63%	860	90.34%

4. 考察

4.1 字種構成

構成字種数に着目し，構成字種数毎の構成パターンの特徴について考察する．

表4.1は2字種構成のみの構成パターン比率を表したものである．2字種構成は全用語の約90%を占める．上位5種で累積95%に達する．これは2字種全17種のうち約1/3である．また最上位パターンだけで約70%であることから，構成パターンには偏りがあることが分かる．

表4.1　構成字種数毎パターン統計（2字種）

構成字種	語数	比率	累積	累積比率
JK	76,479	68.83%	76,479	68.83%
HJ	20,992	18.89%	97,471	87.72%
aJ	3,510	3.16%	100,981	90.88%
nJ	2,746	2.47%	103,727	93.35%
KS	1,848	1.66%	105,575	95.02%

表4.2は3字種構成のみの構成パターン比率を表したものである．3字種は全体の10%程度を占める．組み合わせ計算のため本来のパターン数が多く，上位9種で累積95%に達する．この上位9種は3字種全25種のうち約1/3程度であり，下位約2/3のパターンは比較的ユニークなパターンであることが分かる．

表4.3は4字種構成のみの構成パターン比率を表したものである．4字種は全体の1%程度を占める．上位6種目以降の出現比率は上位5種に対して小さく，構成に偏りがあることが分かる．

表4.4は5字種構成のみの構成パターン比率を表したものである．5字種は全体の0.1%程度を占める．最上位のパターンとそれ以外のパターンの差が大きく，全体としてユニークなパターンが多いことが分かる．

表4.2 構成字種数毎パターン統計（3字種）

構成字種	語数	比率	累積	累積比率
HJK	4,005	34.73%	4,005	34.73%
JKS	2,818	24.44%	6,823	59.17%
aJK	1,170	10.15%	7,993	69.31%
nJK	896	7.77%	8,889	77.08%
aKS	811	7.03%	9,700	84.11%
nKS	605	5.25%	10,305	89.36%
aJS	276	2.39%	10,581	91.75%
anJ	229	1.99%	10,810	93.74%
anK	184	1.60%	10,994	95.33%

表4.3 構成字種数毎パターン統計（4字種）

構成字種	語数	比率	累積	累積比率
nJKS	471	33.43%	471	33.43%
aJKS	378	26.83%	849	60.26%
anKS	162	11.50%	1,011	71.75%
anJS	128	9.08%	1,139	80.84%
anJK	111	7.88%	1,250	88.72%
HJKS	31	2.20%	1,281	90.92%
aHJK	31	2.20%	1,312	93.12%
nHJK	24	1.70%	1,336	94.82%
asJK	23	1.63%	1,359	96.45%

表4.4 構成字種数毎パターン統計（5字種）

構成字種	語数	比率	累積	累積比率
anJKS	111	79.29%	111	79.29%
nsJKS	14	10.00%	125	89.29%
aHJKS	3	2.14%	128	91.43%
anHJK	3	2.14%	131	93.57%
ansKS	3	2.14%	134	95.71%

6字種構成の構成の種類は1用語だけで，「N－1－ナフチル－N’－ジエチルエチレンジアミンしゅう酸塩」という用語で，構成字種はanHJKSである．本研究で使用された辞書には化学用語辞典は含まれてない．化学用語辞典の見出し語を分析対象とすれば，このような用語が多く出現すると考えられる．

4.2 字種変化

4.2.1 字種変化数毎の変化パターン

表4.5は変化数2の用語統計である．変化数2のパターンは全29種であり，上位5種で累積90%，上位7種で累積95%に達する．つまり全パターン中，上位約24%のパターンが大半を占める．また変化数2は全体の用語の中で最も多い．そのため上記の統計的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4.6は変化数3の用語統計である．変化数3のパターンは全101種であり，上位14種で累積90%，上位22種で累積95%に達する．つまり全パターン中，上位約22%のパターンが大半を占める．

表4.5 変化数毎の変化パターン・2変化

パターン	用語数	比率	累積	累積比率
KJ	40,179	47.48%	40,179	47.48%
JK	23,359	27.60%	63,538	75.09%
JH	8,389	9.91%	71,927	85.00%
HJ	3,480	4.11%	75,407	89.11%
aJ	2,591	3.06%	77,998	92.17%
nJ	1,528	1.81%	79,526	93.98%
Kn	1,342	1.59%	80,868	95.57%
aK	921	1.09%	81,789	96.65%
Ka	530	0.63%	82,319	97.28%
an	482	0.57%	82,801	97.85%

表4.6 変化数毎の変化パターン・3変化

パターン	用語数	比率	累積	累積比率
JKJ	7,659	27.17%	7,659	27.17%
JHJ	6,957	24.68%	14,616	51.85%
KJK	3,877	13.75%	18,493	65.60%
KSK	1,515	5.37%	20,008	70.97%
JnJ	1,048	3.72%	21,056	74.69%
JHK	938	3.33%	21,994	78.02%
KJH	645	2.29%	22,639	80.31%
HJH	555	1.97%	23,194	82.27%
JaJ	511	1.81%	23,705	84.09%
JSJ	502	1.78%	24,207	85.87%

表4.7 変化数毎の変化パターン・4変化

パターン	用語数	比率	累積	累積比率
KSKJ	1,775	22.32%	1,775	22.32%
JHJH	935	11.76%	2,710	34.08%
JKJK	638	8.02%	3,348	42.11%
KJKJ	547	6.88%	3,895	48.99%
KJHJ	515	6.48%	4,410	55.46%
JHJK	332	4.18%	4,742	59.64%
HJHJ	247	3.11%	4,989	62.75%
aSaJ	177	2.23%	5,166	64.97%
aSKJ	142	1.79%	5,308	66.76%
JHKJ	135	1.70%	5,443	68.46%

表4.7は変化数4の用語数を示したものである．変化数4のパターンは全218種であり，上位48種で累積90%，上位80種で累積95%に達する．つまり全パターン中，上位約36%のパターンが大半を占める．

表4.8は変化数5の用語統計である．変化数5のパターンは全273種であり，上位110種で累積90%，上位164種で累積95%に達する．これは全パターンの約60%である．

表4.9は変化数6の用語統計である．変化数6のパターンは全191種であり，上位121種で累積90%，上位156種で累積95%に達する．これは全パターンの約81%である．

表4.8　変化数毎の変化パターン・5変化

パターン	用語数	比率	累積	累積比率
JHJHJ	354	16.33%	354	16.33%
JKJKJ	122	5.63%	476	21.96%
JKJHJ	102	4.70%	578	26.66%
JHJHK	85	3.92%	663	30.58%
KJKJK	73	3.37%	736	33.95%
KSKJK	69	3.18%	805	37.13%
SnKSK	67	3.09%	872	40.22%
KSKSK	62	2.86%	934	43.08%
aSanJ	61	2.81%	995	45.89%
KSaSK	50	2.31%	1,045	48.20%

表4.9　変化数毎の変化パターン・6変化

パターン	用語数	比率	累積	累積比率
KSKSKJ	132	18.70%	132	18.70%
JHJHJH	33	4.67%	165	23.37%
KSnSKJ	28	3.97%	193	27.34%
SnKJSK	25	3.54%	218	30.88%
KnSnKJ	22	3.12%	240	33.99%
JHJHJK	21	2.97%	261	36.97%
KJHJHJ	17	2.41%	278	39.38%
KSKSJK	14	1.98%	292	41.36%
aSaSaJ	13	1.84%	305	43.20%
JKJKJK	12	1.70%	317	44.90%

4.2.2 先頭字種毎の変化パターン累積分布

図4.1は，表3.7～表3.19の先頭字種毎に用語数の累積比率をグラフ化したものである．横軸は先頭字種毎の変化パターン総数に対するパターンの割合，縦軸は先頭字種毎の用語数に対する累積比率を示している．

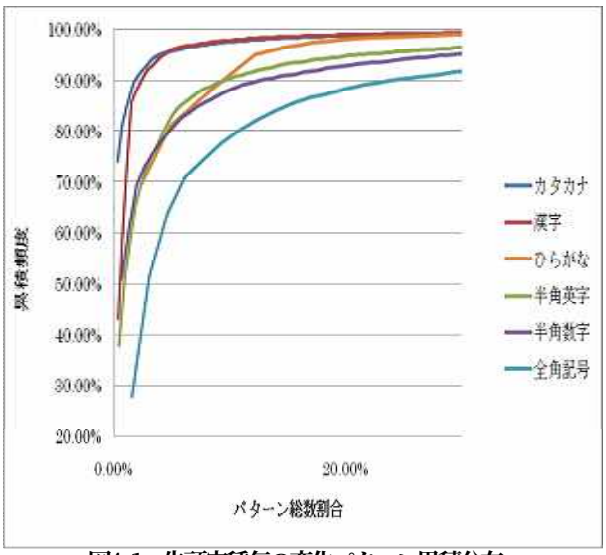

図4.1　先頭字種毎の変化パターン累積分布

このグラフから，先頭がどの字種で始まるかかにかかわらず，上位20～30%の種類の変化パターンでそれぞれの用語数の90%に達することが明確にわかる．

先頭字種が漢字またはカタカナである用語は他に比べてより少ないパターンで90%以上に達することが分かる．これらの用語は，特定非常に少ない字種変化で表現されていることを示している．先頭字種がひらがなの用語は，上位1種の時点で累積比率約70%と高く，この傾向が顕著であることがわかる．以上3字種に対して，半角英字，半角数字，全角記号で始まる用語は，多くの字種変化パターンで表現されることがわかる．これらの事実から，先頭字種が日本語文字列である場合，パターンのバラつきが少なく，逆に英数字のような字種の場合にはユニークなパターンが多いことが分かる．

5. 終わりに

本研究の結果から，日本語多字種複合語について字種構成という観点からの特性が明らかになった．調査対象とされた辞書は15年以上前のものであるが，国語辞典や外来語辞典などの一般用語を含む辞典だけでなく科学技術，医学分野の辞典も対象としているため，この特性は日本語全体に対して適用可能であると考えられる．本研究で明らかにされ列挙された字種変化パターンを形態素解析やチャンキングに用いることで，精度の向上につながる可能性があると考えられる．

註・参考文献

[1] 大畑博一，中川裕志．連接異なり語数による専門用語抽出．自然言語処理研究会報告 (29)，119-126 (2000)．
[2] 中川裕志,湯本紘彰,森辰則．出現頻度と連接頻度に基づく専門用語抽出．Journal of natural language processing 10 (1)，27-45 (2003)．
[3] 青木和夫，中山章弘，松崎剛士．形態素解析での効率的な複合語処理．自然言語処理研究会報告 (57)，1-6 (2003)．
[4] 中瀬健太，梅村恭司．Bigramの反復度を用いた技術用語抽出．情報処理学会研究報告．DD (97)，15-20 (2004)．
[5] 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)．
[6] 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)．
[7] 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)．
[8] 下畑光夫，杉尾俊之．文字種切り出しと複合語分解によるキーワード抽出．NLC，言語理解とコミュニケーション (200)，13-18 (1997)．
[9] これらの辞書は，EPwing，EB formatのCD版の電子辞書で，購入時に当研究室でテキストデータに変換したものである．