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

田代 征嗣†　　滝川 諒†　　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

日本語の理工学分野のテキストにおいて，主要な概念，テーマは専門用語として多字種から構成される複合語で表現されることが多い．長単位の複合語表現が文章中に多々出現する．

日本語の専門用語に関して様々な研究がなされてきたが，これらの研究は用語抽出の対象とされ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かった[1][2][3][4].

当研究室では，多字種複合語／表現についてそれらの構成字種，字種変化の観点から昨年度以下について報告した．

(1) 複数の辞書見出し語[5]
(2) 特許抄録[6][7]

本研究は学術論文抄録に出現する多字種複合語に対して，上記報告と同様の観点から分析を行うものである．

2. 用語解析手順

2.1 コーパス

コーパスとして，国立情報学研究所(NII)のNTCIR-1で使用された学会発表文データベースの抄録約33万を使用した．

2.2 抽出手順

上述のコーパスを対象として以下手順により解析対象複合語集合を得た．

step1. 多字種複合文字列抽出プログラムによる自動抽出
↓
step 2. 乱数を用いて総異なり15万語を選択
↓
step 3. 人手によるスクリーニング

step1で使用したプログラムは，筆者らによる昨年度の特許抄録に対する多字種複合語の研究で用いたプログラムと同一である[6]．また，品詞辞書も同一で，EDRに電子辞書に基づくものである[8]．この段階で，113万語の総異なり多字種複合文字列を得た．

人手によるスクリーニングでは，step1のプログラムによる誤抽出，非名詞相当表現等を削除し，名詞相当表現文字列のみを得た．

2.3 解析項目

上述の手順で得られた，128744語の多字種複合表現集合について，連接している構成単語の字種の順序，出現回数を分析の対象とした．例えば，「LCD投射型ディスプレイ」は，字種の出現順序として，”半角英字”，”漢字”，”カタカナ”，であり，字種を以下に挙げる記号で表現すると，”aJK”となる．ここでは，字種の出現順序を字種変化パターンとよび，また”a”→”J”→”K”と変化し，これを字種変化数とよぶ．この例では，字種変化数は3とする．本研究では，前研究と同様に字種を以下の9種類に分類し，個々の字種変化パターン，字種変化数について異なり用語総数を調査し，その特性を分析した．

(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

表1　字種変化数毎の用語数

変化数	用語数	比率(%)	累積用語数	累積比率(%)
2	64,176	49.84	64,176	49.84
3	35,888	27.87	100,061	77.70
4	13,845	10.75	113,904	88.45
5	7,751	6.02	121,659	94.47
6	3,137	2.44	124,797	96.91
7	1,663	1.29	126,460	98.20
8	894	0.69	127,354	98.90
9	578	0.45	127,932	99.35

3.1 用語全体

全用語では，字種変化数は2～80の総計30種類であった．表1は異なり用語数多い順に上位2～9の字種変化数毎の異なり用語数を示している．この表から，対象集合全体に対して，変化数2～6までの用語で累積比率は約96%を超え，7変化以上の多字種複合語表現は非常に少ないことがわかる．

一方，字種変化パターンについては，対象複合語集合全体では，4624種あった．上位23種で全複合語の70%，674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について，用語数の累積比率70%，上位23位までを列挙したものである．これはパターン総数の0.5%に過ぎない．残り99.5%のパターンは用語全体の30%程度しか出現しないことが分かる．またこの表から，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JKJ(3位)，KJKJ(16位)，さらには5変化(19位)の変化パターンが上位にあることがわかる．

表2　字種変化パターン毎の用語数

パターン	用語数	比率(%)	累積用語数	累積比率(%)
KJ	26415	20.51	26,415	20.51
JK	18667	14.50	45,082	35.01
JKJ	9678	7.52	54,760	42.52
aJ	5305	4.12	60,065	46.64
JHJ	4478	3.48	64,543	50.12
Ja	3000	2.33	67,543	52.45
JnJ	2887	2.24	70,430	54.69
KJK	2662	2.07	73,092	56.76
nJ	2574	2.00	75,666	58.76
JH	2081	1.62	77,747	60.37
JaJ	1609	1.25	79,356	61.62
Jna	1526	1.19	80,882	62.81
asa	1349	1.05	82,231	63.86
aK	1279	0.99	83,510	64.85
Ka	883	0.69	84,393	65.54
KJKJ	842	0.65	85,235	66.19
na	830	0.64	86,065	66.83
asaJ	829	0.64	86,894	67.48
JHJHJ	767	0.60	87,661	68.07
JKJK	747	0.58	88,408	68.65
JsJ	735	0.57	89,143	69.22
Jn	711	0.55	89,854	69.78
nsna	627	0.49	90,481	70.26

3.2 字種変化数毎の用語数

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明らかにする．

3.2.1 字種変化数2

字種変化数2のパターンは34種存在した．表3は字種変化数2で，異なり用語数の多い字種変化パターンを多い順に，累積比率99%に達する上位17種を列挙したものである．字種変化数2の用語総数に対して，上位6種で累積90%，上位9種で累積95%，さらに上位17種99%を超える．また字種変化数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47%に相当する．

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(%)	累積用語数	累積比率	全体比率
KJ	26415	41.16	26,415	41.16	20.51
JK	18667	29.09	45,082	70.25	35.01
aJ	5305	8.27	50,387	78.51	39.13
Ja	3000	4.67	53,387	83.19	41.46
nJ	2574	4.01	55,961	87.20	43.46
JH	2081	3.24	58,042	90.44	45.07
aK	1279	1.99	59,321	92.43	46.07
Ka	883	1.38	60,204	93.81	46.75
Na	830	1.29	61,034	95.10	47.40
Jn	711	1.11	61,745	96.21	47.95
An	578	0.90	62,323	97.11	48.40
HJ	285	0.44	62,608	97.56	48.62
SJ	251	0.39	62,859	97.95	48.81
nK	235	0.37	63,094	98.31	49.00
JS	225	0.35	63,319	98.66	49.17
Kn	197	0.31	63,516	98.97	49.32
Sa	102	0.16	63,618	99.13	49.40

以下に上位3位までについて，それぞれのパターンを採る用語(，表現)の実例を挙げる．

KJ　ワンステップ反応
aJ　Zippen反応
JK　三次元共有メモリ

3.2.2 字種変化数3

字種変化数3のパターンは149種存在した．表4は字種変化数3で，異なり用語数の多い字種変化パターンを多い順に，累積比率80%を超える上位16種を列挙したものである．字種変化数3の用語総数に対して，上位9種で累積70%，上位16種で累積80%，に達する．上位8位(JsJ)以下から，用語数が減少していることがわかる．

以下に上位3位までについて，それぞれのパターンを採る用語(，表現)の実例を挙げる．

JKJ　光アンプ雑音
JHJ　光増幅器付き受信機
JnJ　従来型1種

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(%)	累積用語数	累積比率(%)	全体比率
JKJ	9678	26.97	9,678	26.97	7.52
JHJ	4478	12.48	14,156	39.44	10.99
JnJ	2887	8.04	17,043	47.49	13.23
KJK	2662	7.42	19,705	54.91	15.30
JaJ	1609	4.48	21,314	59.39	16.55
Jna	1526	4.25	22,840	63.64	17.74

Asa	1349	3.76	24,189	67.40	18.78
JsJ	735	2.05	24,924	69.45	19.35
aJK	551	1.54	25,475	70.98	19.78
aKJ	524	1.46	25,999	72.44	20.19
JHK	482	1.34	26,481	73.79	20.56
KJa	480	1.34	26,961	75.13	20.94
JaK	476	1.33	27,437	76.45	21.31
naJ	457	1.27	27,894	77.73	21.66
asn	431	1.20	28,325	78.93	22.00
JKa	415	1.16	28,740	80.08	22.32

3.2.3 字種変化数4

字種変化数4のパターンは469種存在した．表4は字種変化数4で，異なり用語数の多い字種変化パターンを多い順に，累積比率50%を超える上位18種を列挙したものである．

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(%)	累積用語数	累積比率(%)	全体比率(%)
KJKJ	842	6.08	842	6.08	0.65
asaJ	829	5.99	1,671	12.07	1.30
JKJK	747	5.40	2,418	17.46	1.88
nsna	627	4.53	3,045	21.99	2.36
Jasa	393	2.84	3,438	24.83	2.67
JHJH	378	2.73	3,816	27.56	2.96
asnJ	374	2.70	4,190	30.26	3.25
KJHJ	261	1.89	4,451	32.15	3.46
nsnJ	250	1.81	4,701	33.95	3.65
JHJK	243	1.76	4,944	35.71	3.84
nJnJ	235	1.70	5,179	37.41	4.02
JHKJ	234	1.69	5,413	39.10	4.20
Jnsn	224	1.62	5,637	40.72	4.38

以下に上位3位までについて，それぞれのパターンを採る用語(，表現)の実例を挙げる．

KJKJ　　アクリル酸ベンジル誘導体
asaJ　　A+T含量
JKJK　　多モード光ファイバ

3.2.4 字種変化数5

字種変化数5のパターンは859種存在した．表6は字種変化数5で，異なり用語数の多い字種変化パターンを多い順に，累積比率30%を超える上位19種を列挙したものである．

以下に上位3位までについて，それぞれのパターンを採る用語(，表現)の実例を挙げる．

JHJHJ　　呼び出し時間
Jnsna　　出力24.1dBm
asasa　　DNS-Asn-GlcNAc

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(%)	累積用語数	累積比率(%)	全体比率(%)
JHJHJ	767	9.90	767	9.90	0.60
Jnsna	627	8.09	1,394	17.98	1.08
asasa	184	2.37	1,578	20.36	1.23
Jnsns	175	2.26	1,753	22.62	1.36
JKJKJ	157	2.03	1,910	24.64	1.48
JnsnJ	142	1.83	2,052	26.47	1.59
Jnasn	140	1.81	2,192	28.28	1.70
JasaJ	121	1.56	2,313	29.84	1.80
asnsn	119	1.54	2,432	31.38	1.89
JnSna	106	1.37	2,538	32.74	1.97
JnJnJ	92	1.19	2,630	33.93	2.04
JnSnJ	92	1.19	2,722	35.12	2.11
KJasa	90	1.16	2,812	36.28	2.18
nsnaJ	89	1.15	2,901	37.43	2.25
JHJna	84	1.08	2,985	38.51	2.32
nsnSa	79	1.02	3,064	39.53	2.38
KJKJK	72	0.93	3,136	40.46	2.44

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(%)	累積用語数	累積比率(%)	全体比率
JnsnSa	78	2.49	78	2.49	0.06
nsnasa	75	2.39	153	4.88	0.12
asasaJ	71	2.26	224	7.14	0.17
KJnsna	66	2.10	290	9.24	0.23
JHnsna	61	1.94	351	11.19	0.27
nJnJnJ	57	1.82	408	13.01	0.32
KJHJHJ	56	1.79	464	14.79	0.36
asnasn	41	1.31	505	16.10	0.39
asnsna	39	1.24	544	17.34	0.42
asasnJ	37	1.18	581	18.52	0.45
nasasn	33	1.05	614	19.57	0.48
nsnsKJ	33	1.05	647	20.62	0.50

3.2.5 字種変化数6

字種変化数6のパターンは919種存在した．表7は字種変化数56，異なり用語数の多い字種変化パターンを多い順に，累積比率20%を超える上位19種を列挙したものである．

以下に上位3位までについて，それぞれのパターンを採る用語(，表現)の実例を挙げる．

JnsnSa　　解像度0.1μm
nsnasa　　0.005bit/pel
asasaJ　　Ag-Sn-Cu球状合金

4. 考察

4.1 字種変化数

図1は，上述の表3から表7まで，すなわち，字種変化数2～6までの各変化数の上位10種の字種変化パターンの累積比率をグラフ化したものである．

すでに前章の表3～表7により明らかにされた次の事実がこのグラフから明瞭に読み取ることができる．すなわち，字種変化数が少ないほど，上位の特定の字種変化パターンをもつ複合語が多い．一方，字種変化数が多いほど，特定の字種変化パターンをもつ複合語は少なくなる．

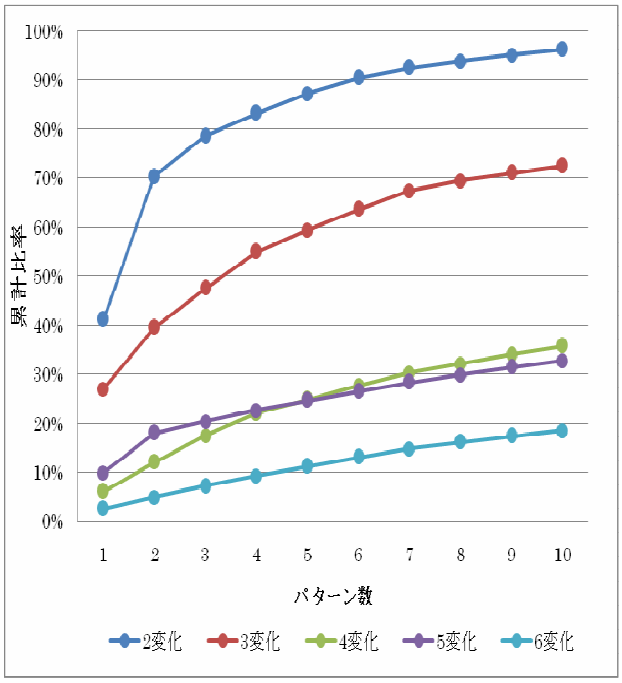

図1　字種変化数毎の複合語の累積比率

4.2 字種変化パターン

字種数毎の理論的な全変化パターンの種類Pnは，次の式で与えられる．

$$Pn = K \cdot (K-1)^{L-1}$$

K= 全字種数(= 9) , L = 字種変化パターン長

表8は，字種数毎の理論値，実際に出現した字種変化パターンの種類の比率を対比したものである．図2はこれをグラフにしたものである．

表8　字種変化数毎の複合語の出現率

字種変化数	出現率 (%)	字種変化数	出現率 (%)
2	47.22	5	2.33
3	25.87	6	0.31
4	10.18		

表1から，変化数が増加するにつれて用語数も減少するため，結果として使用されている字種変化パターンも減少する．グラフは当然の傾向を示していると言える．

5. 終わりに

本報告は，学術論文抄録中に出現する複数字種からなる複合語(および名詞相当表現)約12.8万語について，字種変化数，字種変化パターンについて，以下について明らかにした．

(a) 2回以上字種が変化する語ついて，字種変化数毎の総異なり数

(b) 字種変化数毎の字種変化パターンの種類とその総異なり数

本報告の結果は，日本語学，自然言語解析の研究者の経験的な知見を超えるものではないが，表層的，全体的な特性について客観的なデータを得ることができた．

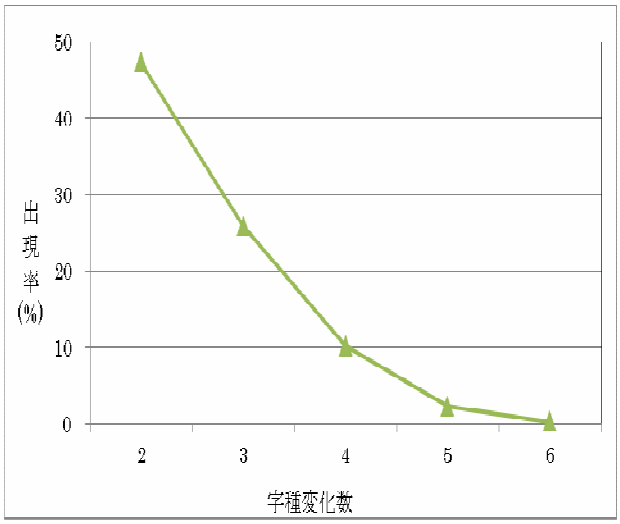

図2　字種変化数毎の複合語の出現率

謝辞

本研究では，国立情報学研究所(NII)のNTCIR-2テストコレクションを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] 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).

[2] 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).

[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] 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).

[5] 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

[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.1．自然言語処理研究会報告 2011-NL-204

[7] 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

[5] EDR日本語単語辞書．独立行政法人 日本情報通信機構(NiCT) 2002年.