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

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で使用される専門用語は，専門性が高くなる，言い換えれば概念の内包が小さくなるほど多くの構成単語，また複数の字種から構成される用語となる傾向がある．また，理工学文献中には，単位，数値，数式，化学構造表現が多数出現する．これは，いずれも単一の字種ではなく複数の字種で表記される．

当研究室では，昨年来，以下のコーパスを対象として，字種の観点からこれら複数字種からなる複合語に対し，表層的，全体的特性を明らかにしてきた．

(1) 複数の辞書見出し語[1]
(2) 特許抄録[2][3]
(3) 学術論文抄録[4]

本報告は，学術論文標題および副標題中に出現する多字種複合語について同様な観点から調査・分析するものである．

2. 用語解析手順

2.1 コーパス

筆者らによる当年次大会のもう1つの発表[4]と同様にコーパスとして，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標題および副標題約33万を使用した．

2.2 抽出手順

抽出手順は，上述の文献[4]と同様，すなわち同一の抽出プログラム，同一の品詞辞書[5]を使用した．ただし，step1の前処理として，標題と副標題とを区別する"-", " - "の連続に対し，空白で置換するという処理をした[6]．

2.3 解析項目

上述の手順により，上記学術論文標題から140,014語の複数字種から構成される総異なり用語を得た．本研究では，文献[5]と同様の観点，すなわち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，その全体的特性を調査・分析した．

3. 結果

3.1 用語全体

全用語では，字種変化数は2～25の総計24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．

表1　字種変化数毎の異なり用語数

変化数	用語数	比率 (%)	累積用語数	累積比率 (%)
2	70078	50.05	70078	50.05
3	38003	27.14	108075	77.19
4	15530	11.08	123590	88.27
5	7813	5.59	131413	93.86
6	3727	2.66	135142	96.52
7	1978	1.41	137120	97.93
8	1152	0.82	138272	98.76
9	628	0.45	138900	99.20

一方，字種変化パターンについて，対象複合語(,表現)集合全体では，5081種あった．上位種で全複合語の90%，188種で95%を含んでいることが判明した．

表2は異なり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18位までを列挙したものである．これはパターン総数の0.4%に過ぎない．残り99.6%のパターンは用語全体の30%程度にしか出現しないことが分かる．この表から，上位4位(字種変化パターンaJ)以下から，用語数は大幅に低下することがわかる．また，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，4変化，例えばJKJKの変化パターンは8位にある．

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	27941	19.96	27941	19.96
JK	24098	17.21	52039	37.17
JKJ	13159	9.40	65198	46.57
aJ	6946	4.96	72144	51.53
KJK	4766	3.40	76910	54.93
JHJ	3162	2.26	80072	57.19
Ja	2480	1.77	82552	58.96

JKJK	2055	1.47	84607	60.43
aK	2034	1.45	86641	61.88
JaJ	1845	1.32	88486	63.20
KJKJ	1759	1.26	90245	64.45
nJ	1710	1.22	91955	65.68
asa	1576	1.13	93531	66.80
aJK	1356	0.97	94887	67.77
JH	1350	0.96	96237	68.73
JsJ	1298	0.93	97535	69.66
aKJ	1158	0.83	98693	70.49

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35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14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5種で累積90%，上位7種で累積約95%，さらに上位14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の累積比率90%は用語集合全体の45%を超える．このことは，2字種でかつ特定の字種変化パターンを採る用語が，非常に多いことを示している．

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	27941	39.87	27,941	39.87	19.96
JK	24098	34.39	52,039	74.26	37.17
aJ	6946	9.91	58,985	84.17	42.13
Ja	2480	3.54	61,465	87.71	43.90
aK	2034	2.90	63,499	90.61	45.35
nJ	1710	2.44	65,209	93.05	46.57
JH	1350	1.93	66,559	94.98	47.54
Ka	824	1.18	67,383	96.15	48.13
HJ	489	0.70	67,872	96.85	48.48
an	427	0.61	68,299	97.46	48.78
na	354	0.51	68,653	97.97	49.03
Js	339	0.48	68,992	98.45	49.28
Jn	317	0.45	69,309	98.90	49.50
nK	127	0.18	69,436	99.08	49.59

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　アーキテクチャ情報
JK　電気インピーダンス
aJ　z領域

3.2.2 字種変化数3

字種変化数3のパターンは141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90%を超える上位19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7種で累積71%，上位11種で累積81%，上位19種で90%に達する．上位1位(JKJ)のみが極端に多く，2位(KJK)はその1/3ほどである．以下暫時減少してゆくことがわかる．

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	13159	34.63	13,159	34.63	9.40
KJK	4766	12.54	17,925	47.17	12.80
JHJ	3162	8.32	21,087	55.49	15.06
JaJ	1845	4.85	22,932	60.34	16.38
asa	1576	4.15	24,508	64.49	17.50
aJK	1356	3.57	25,864	68.06	18.47
JsJ	1298	3.42	27,162	71.47	19.40
aKJ	1158	3.05	28,320	74.52	20.23
JnJ	905	2.38	29,225	76.90	20.87
JaK	898	2.36	30,123	79.26	21.51
JKa	664	1.75	30,787	81.01	21.99
KJa	619	1.63	31,406	82.64	22.43
nJK	569	1.50	31,975	84.14	22.84
asn	545	1.43	32,520	85.57	23.23
KaJ	479	1.26	32,999	86.83	23.57
naJ	388	1.02	33,387	87.85	23.85
asK	291	0.77	33,678	88.62	24.05
JHK	273	0.72	33,951	89.34	24.25
anJ	263	0.69	34,214	90.03	24.44

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　等式プログラム処理系
KJK　ワンタッチ式ジョイント
JHJ　詰め将棋

3.2.3 変字種化数4

字種変化数4のパターンは392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60%を超える上位12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(%)	累積用号数	累積比率(%)	全体(%)
JKJK	2055	13.23	2,055	13.23	1.47
KJKJ	1759	11.33	3,814	24.56	2.72
asaJ	1138	7.33	4,952	31.89	3.54
aJKJ	562	3.62	5,514	35.51	3.94
asnJ	453	2.92	5,967	38.42	4.26
JHJK	328	2.11	6,295	40.53	4.50

asaK	324	2.09	6,619	42.62	4.73
JaJK	310	2.00	6,929	44.62	4.95
nJKJ	297	1.91	7,226	46.53	5.16
KJHJ	279	1.80	7,505	48.33	5.36
KJaJ	261	1.68	7,766	50.01	5.55
JsKJ	252	1.62	8,018	51.63	5.73
JKJa	248	1.60	8,266	53.23	5.90
Jasa	247	1.59	8,513	54.82	6.08
asKJ	237	1.53	8,750	56.34	6.25
JHJH	210	1.35	8,960	57.69	6.40
aKJK	205	1.32	9,165	59.01	6.55
JaKJ	202	1.30	9,367	60.32	6.69

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．

JKJK　代数的プロセス合成システム
KJKJ　ワード線ブースト手法
asaJ　water-uptake効果

3.2.4 字種変化数5

字種変化数5のパターンは726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50%を超える上位17種を列挙したものである．上位2位までのパターン，JHJHJとJKJKJが顕著に多く，上位3位のパターンKJKJKから暫時減少していることがわかる．

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	579	7.28	579	7.28	0.41
JKJKJ	570	7.17	1,149	14.45	0.82
KJKJK	291	3.66	1,440	18.10	1.03
JasaJ	203	2.55	1,643	20.66	1.17
asaJK	203	2.55	1,846	23.21	1.32
asaKJ	165	2.07	2,011	25.28	1.44
asasa	143	1.80	2,154	27.08	1.54
JKJHJ	127	1.60	2,281	28.68	1.63
aJKJK	124	1.56	2,405	30.24	1.72
JasaK	118	1.48	2,523	31.72	1.80
JaJKJ	107	1.35	2,630	33.07	1.88
JsJKJ	103	1.29	2,733	34.36	1.95
JKJaJ	100	1.26	2,833	35.62	2.02
nasaJ	95	1.19	2,928	36.81	2.09
asnJK	91	1.14	3,019	37.96	2.16
JnJKJ	90	1.13	3,109	39.09	2.22
JKJaK	89	1.12	3,198	40.21	2.28
JKasa	88	1.11	3,286	41.31	2.35
naJKJ	85	1.07	3,371	42.38	2.41
JasnJ	84	1.06	3,455	43.44	2.47
JHJKJ	77	0.97	3,532	44.41	2.52
JKJKa	74	0.93	3,606	45.34	2.58
asasn	74	0.93	3,680	46.27	2.63
asnKJ	72	0.91	3,752	47.17	2.68
aKJKJ	68	0.85	3,820	48.03	2.73
nsnsK	67	0.84	3,887	48.87	2.78
KJasa	64	0.80	3,951	49.67	2.82
KJsKJ	64	0.80	4,015	50.48	2.87

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　遠心力吹き付け工法
JKJKJ　液胞タンパク質プロ型前駆体
KJKJK　エンド型アルギン酸リアーゼ

3.2.5 字種変化数6

字種変化数6のパターン985種存在した．表7は字種変化数6の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25%を超える上位17種を列挙したものである．

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(%)	累積用語数	累積比率	全体比率
asasaJ	134	3.60	134	3.60	0.10
JKJKJK	108	2.90	242	6.49	0.17
asaJKJ	79	2.12	321	8.61	0.23
asasnJ	76	2.04	397	10.65	0.28
asnasn	71	1.91	468	12.56	0.33
KJKJKJ	68	1.82	536	14.38	0.38
KJHJHJ	61	1.64	597	16.02	0.43
JHJHJK	39	1.05	636	17.06	0.45
asnJKJ	39	1.05	675	18.11	0.48
nsnSaJ	38	1.02	713	19.13	0.51
asasaK	36	0.97	749	20.10	0.53
KJasaK	35	0.94	784	21.04	0.56
nsnsKJ	31	0.83	815	21.87	0.58
JasaJK	30	0.80	845	22.67	0.60
KnsnsK	30	0.80	875	23.48	0.62
nsnaJK	30	0.80	905	24.28	0.65
asnsaJ	29	0.78	934	25.06	0.67

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．

asasaJ　Ae-Bi-O系
JKJKJK　規則構造モジュール用ジェネレータ開発システム
asaJKJ　A/I統一スイッチング方式

4. 考察

4.1変化数毎の累積分布

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．

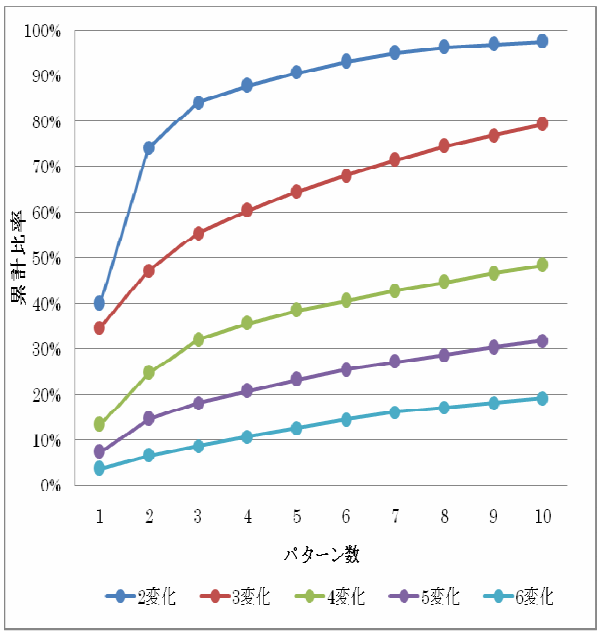

図1　字種変化数毎の累積比率

4.2 字種変化パターン

筆者らによる本年次大会のもう1つの発表[4]と同様に字種数毎の理論的な全変化パターンの種類と字種数毎の理論値，実際に出現した字種変化パターンの種類の比率を対比を試みた．

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	48.61	5	1.97
3	24.48	6	0.33
4	8.51		

文献[4]の同様の表(表8)およびグラフ(図2)と対比すると，ほとんど同様の傾向を示している．標題と抄録の相違はあるが，文書集合としては，同一のものから得ているので，当然の結果といえよう．

5. 終わりに

本報告は，学術論文標題および副題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14万語について，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以上字種変化する語ついて，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

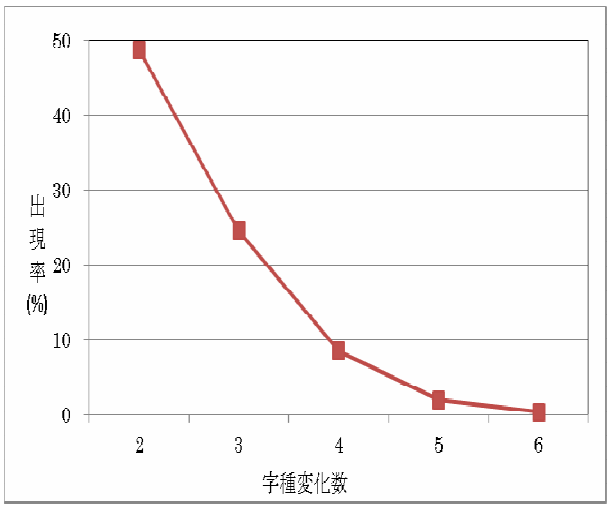

図2　字種変化数毎の複合語の出現率

本報告を含め，以下のコーパス中の用語に対して，延べ約5×10^5の語(，表現)ついて上記の観点からのデータ得たことになる．

(1) 辞書見出し語，　(2) 特許抄録

(3) 学術論文抄録，　(4) 学術論文標題

次のステップとして，(a), (b)の観点からこれらの異なったテキストから得られた用語集合の対比を試みることになろう．

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．

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．自然言語処理研究会報告 2011-NL-204

[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.1．自然言語処理研究会報告 2011-NL-204

[4] 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．言語処理学会第18回年次大会(2012), B4-6..

[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.

[6] 標題と副標題とを区別するために，以下のような様々な表現が実際に用いられており，前処理で全てについて正確に分離できなかった．

“:”, “(その1)”, “その1”‘　など

抽出プログラムの結果として，上記標記表現と副標題の先頭の用語が連接された誤った文字列が出力された．これらの文字列はスクリーニング段階で削除した．