Vol.2011-NL-204 No.2
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.1
- 多字種複合語の抽出と構成字種の解析 -

滝川 諒[†]　　後藤 智範[††]

日本語の科学技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，複合語表現が文章中に多々出現する．1993年度の公開特許データベースの抄録から、著者らにより開発された抽出プログラムにより約16万語の多字種複合語を抽出した。さらに人手により非名詞を除去し、約13万語の多字種複合語を得た。Part1では、これらの多字種複合語について字種構成の観点から分析し、さらに辞書見出し語のそれと対比した．

結果として，構成字種数 2〜4 で累積 98%に達する，先頭字種が日本語である場合，英数字や記号よりも構成数は少なくなる，対象となる文書によって出現頻度は大きく異なる，といった特性が明らかになった．

Quantitative Analysis to Japanese Compound Terms with Multi Character Types Appeared in Patent Texts Part.1

Ryo Takikawa[†]　　Tomonori Gotoh[††]

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 extraction method to these terms appeared in the abstract texts of Japanese patent database, and the results of quantitative analyses to the terms from the aspects to structure of character types. Moreover, the comparison was done with the results and our previous research to the terms contained in lots of entry terms in the several dictionaries.

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，特許文書の抄録を対象とし，多字種複合語の抽出手法および複合語の特性，具体的には字種に着目した字種構成について明らかにしようとするものである．

2. 用語抽出手順と解析手順

本章では，対象となるコーパスから多字種複合語を抽出手法と，本研究で行う解析手順について説明する．

2.1 抽出手順

2.1.1 辞書

EDR 日本語単語辞書[9]を基にして作成した辞書を使用する．EDR をそのまま利用するだけでは，用語に不足があるため，いくつかの品詞に対して用語を追加した．具体的には，形式名詞・助詞相当語・助動詞相当語を複数追加した．

表 1 は，本研究で使用した EDR 辞書の品詞体系である．品詞体系は 2 階層の構造を持つ．またサブカテゴリは記載した詳細品詞以外にも，動詞や形容詞などの活用形の情報も含んでいる．

[†] 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

表 1　EDR における品詞分類

品詞	詳細
その他	感動詞
	記号
	助詞
	助動詞
	助動詞相当
	補助用言
サ変名	-
機能語	形式名詞
	助詞
	助詞相当語
	助動詞
	助動詞相当
接続詞	単語接続詞
	文接続詞

品詞	詳細
接頭語	形容詞的接頭
	接頭小辞
	前置助数詞
	副詞的接頭語
	連体詞的接頭
接尾語	後置助数詞
	接尾語
	単位
構文要素	述語句
	体言句
	独立句
	文
	連体修飾句
	連用修飾句

品詞	詳細
動詞	-
副詞	陳述副
	普通副
名詞	形式名
	固有名
	時詞
	数詞
	代名詞
	普通名
形容詞	-
形容動	-
語尾	動詞語
連体詞	-

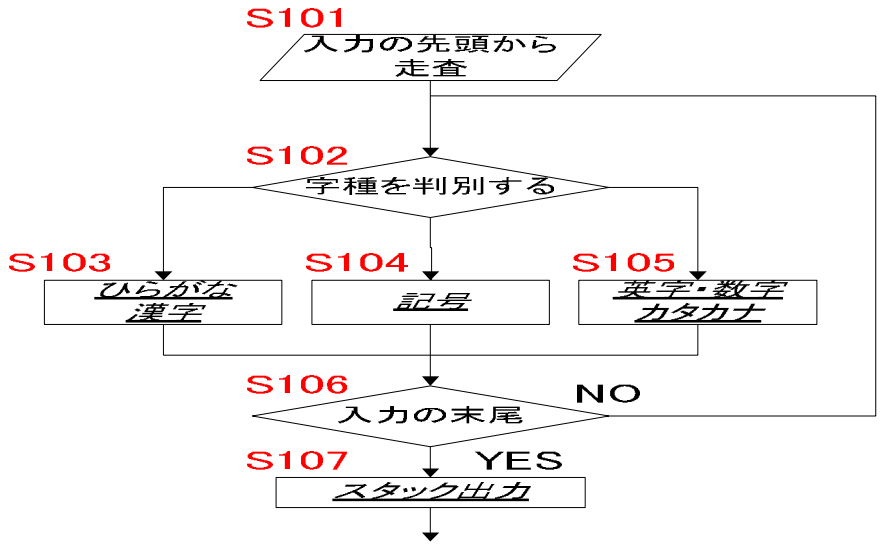

図 1　抽出流れ図①　全体の流れ

2.1.2　**抽出手順**

全体の流れ

図1に全体の流れを挙げる．基本的な流れは文章中に存在する名詞以外の品詞を分割対象文字列とし，分割対象以外の文字をスタックに積み，分割対象文字列が出現する毎にスタック内の文字列を出力することで，多字種複合語を抽出する．

与えられたテキストに対して先頭から走査を行う．（S101）先頭から 1 文字ずつ字種判別を行い（S102）各々の字種に対して処理を行う．（S103～S105）この処理を入力の末尾まで行う．（S106）最終的にスタック内に文字列が残っていることを考慮し，走査終了後にスタックを確認し，スタック状況に適した処理を行う．（S107）

ひらがな・漢字の処理

図 1-S103 のひらがな・漢字に対する処理について説明する．

まず参照位置から後方へ走査を行い，次の文字がひらがな，もしくは漢字であり続ける限り辞書引き対象文字数を増分する．この時，最大は辞書中の用語の最大文字列長までとする．

ひらがな・漢字以外の字種が出現したら，増分を中止し，参照位置から増分した値の文字数分だけ切り出し，辞書引きを行う．辞書中に切り出した文字列が存在しない場合，辞書引き対象文字数を減分し，再度辞書引きを行う．この時，減分した結果辞書引き対象文字数が 0 文字になった場合，参照位置から始まる文字列は辞書に存在しなかったことになるため，参照位置を後方へ移動し，現在参照位置の 1 文字をスタックに追加する．

辞書引きの結果，辞書中に存在している場合，切り出した文字列の長さ分だけ参照位置を後方へ移動させる．その後，品詞毎の処理を行い，分割対象であればスタック出力，分割対象でなければ切り出した文字列をスタックに追加する．

記号の処理

図 1-S104 の記号に対する処理について説明する．

記号には文章中において名詞，もしくは名詞相当語を形成するものと，そうでないものが存在する．本研究では，JIS コードに存在する記号を，(1) 分割対象記号／(2) 名詞内部記号の 2 種類に分類した．この分類の一例を表 2 に示す．この分類に従い，走査対象文字が記号の場合には 2 通りの処理を行う．分割対象であれば，スタックに追加せずにスタック出力を行い，走査開始位置を次の文字に送る．名詞内部記号であれば，スタックに追加して走査開始位置を次の文字に送る．

表 2　記号分類の一例

分割対象	名詞内部
!	%
(	¥
)	&
:	-
、	℃
。	α

英数字・カタカナの処理

図 1-S105 の英数字・カタカナに対する処理について説明する．

英数字およびカタカナは先に挙げた字種とは異なり，全て複合語の内部文字として扱うため，参照位置の 1 文字をスタックに追加し，後方へ移動する．

スタック出力

図 1-S107 や前の節で述べたスタックの出力に関して，説明する．

スタック内には，分割対象以外の文字列が出現順に積まれ，参照位置が分割対象である時に出力のチェックを行う．この時，スタック内が多字種で構成されている場合，スタック内の文字列を複合語として出力する．出力後，スタックをリセットする．

また，スタック内の文字列が単字種であった場合は，出力を行わず内容をリセットし，走査に戻る．

2.1.3　**実例**

辞書最大長：27
入力：　… 液晶セル５にはＡＣ／ＤＣ変換ＩＣ８を介して外部電極端子９が接続し…

現在のスタック　= [ ]
辞書引き文字数　2〜
辞書引き　= [液晶]
品詞チェック
普通名詞(複合語内部文字列)

現在のスタック　= [液晶]
現在のスタック　= [液晶セ]
現在のスタック　= [液晶セル]
現在のスタック　= [液晶セル５]

辞書引き文字数　2〜
辞書引き　= [には]
品詞チェック
**　機能語-助詞相当語　(分割対象)**

スタックチェック　= [液晶セル５]
出力　= [液晶セル５]
スタックリセット

現在のスタック　= [ ]
現在のスタック　= [Ａ]
現在のスタック　= [ＡＣ]
現在のスタック　= [ＡＣ／]
現在のスタック　= [ＡＣ／Ｄ]
現在のスタック　= [ＡＣ／ＤＣ]
辞書引き文字数　2〜
辞書引き　= [変換]
品詞チェック
サ変名詞語幹(複合語内部文字列)

現在のスタック = [ＡＣ／ＤＣ変換]
現在のスタック = [ＡＣ／ＤＣ変換Ｉ]
現在のスタック = [ＡＣ／ＤＣ変換ＩＣ]
現在のスタック = [ＡＣ／ＤＣ変換ＩＣ８]

辞書引き文字数 10～

辞書引き = [を介して外部電極端子]
辞書引き = [を介して外部電極端]
辞書引き = [を介して外部電極]
辞書引き = [を介して外部電]
辞書引き = [を介して外部]
辞書引き = [を介して外]
辞書引き = [を介して]
品詞チェック
** 機能語-助詞相当語 (分割対象)**

スタックチェック = [ＡＣ／ＤＣ変換ＩＣ８]
出力 = [ＡＣ／ＤＣ変換ＩＣ８]
⋮

上記の流れで処理を行い，“液晶セル５”，“ＡＣ／ＤＣ変換ＩＣ８”を複合語として解析対象とする．

2.2 解析手順

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，下記字種分類に基づき字種判別を行い，字種構成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．

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			

3. 結果

3.1 抽出

2.1.1 および 2.1.2 で説明した手順に沿って，1993 年度の特許文書 347,316 件の抄録を対象として抽出を行った．
抽出された文字列は 1,677,594 語となった．この中から無作為に 150,000 語を抽出し，人手で選定を行った．選定の基準は，
（1）名詞として適当かどうか（文章表現になっていないか等）
（2）単一の名詞として適当かどうか（以上，以下などの副詞性接尾辞がないか等）
（3）名詞以外の品詞を含んでいないかどうか
（4）その他，名詞とは判断出来ないと思われるもの
上記 4 点に該当する文字列を除外し，最終的に表 3 に挙げる 135,972（二次抽出の約 90%，全体の約 8%）の多字種複合語を解析の対象とした．

表 3 先頭字種毎の抽出用語数

先頭字種	用語数	比率 (%)
漢字	81309	59.80
カタカナ	39907	29.35
全角英字	7388	5.43
全角数字	4680	3.44
半角数字	985	0.72
半角英字	590	0.43
ひらがな	557	0.41
全角記号	537	0.39
半角記号	19	0.01
合計	135972	100.00

3.2 構成字種

3.2.1 用語全体

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 4 は構成字種数毎の出現比率である．この表から，字種数 2,3,4 で累積比率 98%と

大半を占め，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．最大8字種構成は“値２－R_{i-1}×Y”(“AJKNSans”,“－”は長音記号でありカタカナ扱い)という文字列である．本研究では名詞相当語も抽出対象となるため，このような数式のような文字列も対象に含まれる．

表 4　構成字種数毎の用語数

構成数	用語数	比率 (%)	累積	累積比率 (%)
2	82318	60.540	82318	60.540
3	42863	31.523	125181	92.064
4	9260	6.810	134441	98.874
5	1373	1.010	135814	99.884
6	139	0.102	135953	99.986
7	18	0.013	135971	99.999
8	1	0.001	135972	100.000

表 5　字種構成毎の用語数

構成字種	用語数	比率 (%)	累積	累積比率 (%)
JK	33302	24.49	33302	24.49
JN	25525	18.77	58827	43.26
JKN	19689	14.48	78516	57.74
KN	7188	5.29	85704	63.03
AJN	5494	4.04	91198	67.07
AJ	4347	3.20	95545	70.27
HJ	4255	3.13	99800	73.40
JNS	2725	2.00	102525	75.40
AJKN	2669	1.96	105194	77.36
AJK	2310	1.70	107504	79.06
nJ	2149	1.58	109653	80.64
AKN	1677	1.23	111330	81.88
HJN	1621	1.19	112951	83.07
JKNS	1588	1.17	114539	84.24
HJK	1470	1.08	116009	85.32
nJK	1281	0.94	117290	86.26
AK	1121	0.82	118411	87.08
KNS	966	0.71	119377	87.80
AN	882	0.65	120259	88.44
HJKN	877	0.64	121136	89.09
nK	746	0.55	121882	89.64
NS	744	0.55	122626	90.18

表5は字種構成パターン毎の累積比率90%までの出現頻度である．表5から上位90%までの用語の多くは漢字（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 6　構成字種数毎の構成パターン出現比率

構成数	出現数	総数	比率 (%)
2	32	36	88.89
3	63	84	75.00
4	77	126	61.11
5	51	126	40.48
6	30	84	35.71
7	7	36	19.44
8	1	9	11.11

表6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で計算することができる．結果から，構成数が増えるにつれて，出現パターンの比率は減少することが分かる．

3.2.2　構成数毎

構成字種数に着目し，構成字種数毎の構成パターンの結果を報告する．

構成数 2

表7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字種構成は全複合語の約60%を占める．上位8種で累積95%に達する．これは2字種全32種のうち1/4である．また上位2パターンだけで約70%であり，以降急激な比率減少が見られることから，構成パターンには偏りがあることが分かる．

Vol.2011-NL-204 No.2
2011/11/21

表 7　構成字種 2 のパターン出現頻度

構成字種	用語数	比率 (%)	累積	累積比率 (%)	全体比率 (%)
JK	33302	40.46	33,302	40.46	24.49
JN	25525	31.01	58,827	71.46	43.26
KN	7188	8.73	66,015	80.20	48.55
AJ	4347	5.28	70,362	85.48	51.75
HJ	4255	5.17	74,617	90.64	54.88
nJ	2149	2.61	76,766	93.26	56.46
AK	1121	1.36	77,887	94.62	57.28
AN	882	1.07	78,769	95.69	57.93

構成数 3

表 8　構成字種 3 のパターン出現頻度

構成字種	用語数	比率 (%)	累積	累積比率 (%)	全体比率 (%)
JKN	19689	45.94	19,689	45.94	14.48
AJN	5494	12.82	25,183	58.76	18.52
JNS	2725	6.36	27,908	65.12	20.52
AJK	2310	5.39	30,218	70.51	22.22
AKN	1677	3.91	31,895	74.42	23.46
HJN	1621	3.78	33,516	78.21	24.65
HJK	1470	3.43	34,986	81.64	25.73
nJK	1281	2.99	36,267	84.63	26.67
KNS	966	2.25	37,233	86.88	27.38
ANS	743	1.73	37,976	88.61	27.93
nAJ	655	1.53	38,631	90.14	28.41

表 8 は 3 字種構成のみの構成パターン出現比率を表したものである．3 字種は全体の 30%程度を占める．組み合わせ計算のため 2 字種より本来のパターン数が多く，上位 18 種で累積 95%に達する．この上位 18 種は 3 字種全 63 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．また 2 字種同様，上位 2 パターンだけで大半を占め，以降急激な比率減少が見られることから，構成パターンには偏りがあることが分かる．

構成数 4

表 9　構成字種 4 のパターン出現頻度

構成字種	用語数	比率 (%)	累積	累積比率 (%)	全体比率 (%)
AJKN	2669	28.87	2,669	28.87	1.96
JKNS	1588	17.18	4,257	46.05	3.13
HJKN	877	9.49	5,134	55.53	3.78
AJNS	715	7.73	5,849	63.27	4.30
AKNS	307	3.32	6,156	66.59	4.53
nAJK	296	3.20	6,452	69.79	4.75
AHJN	270	2.92	6,722	72.71	4.94
aAJK	182	1.97	6,904	74.68	5.08
anJK	170	1.84	7,074	76.52	5.20
nANS	141	1.53	7,215	78.04	5.31
AJKS	129	1.40	7,344	79.44	5.40
HJNS	94	1.02	7,438	80.45	5.47
nJKS	94	1.02	7,532	81.47	5.54
nJNS	84	0.91	7,616	82.38	5.60
ansS	84	0.91	7,700	83.29	5.66
nAJS	83	0.90	7,783	84.19	5.72
anAJ	83	0.90	7,866	85.08	5.79
nAJN	77	0.83	7,943	85.92	5.84
nsJK	74	0.80	8,017	86.72	5.90
nsJS	62	0.67	8,079	87.39	5.94
nJKN	59	0.64	8,138	88.03	5.99
AHJK	56	0.61	8,194	88.63	6.03
nHJK	53	0.57	8,247	89.20	6.07
nAKS	46	0.50	8,293	89.70	6.10
aANS	46	0.50	8,339	90.20	6.13

表 9 は 4 字種構成のみの構成パターン出現比率を表したものである．4 字種は全体の 7%程度を占める．上位 38 種で累積 95%に達する．この上位 38 種は 4 字種全 77 種のうち約 1/2 程度である．また先の 2,3 字種同様，上位 2 パターンで約半数を占め，構成パターンに偏りがあることが分かる．

構成数 5

表 10　構成字種 5 のパターン出現頻度

構成字種	用語数	比率 (%)	累積	累積比率 (%)	全体比率 (%)
AJKNS	371	27.85	371	27.85	0.27
AHJKN	87	6.53	458	34.38	0.34
nAJNS	77	5.78	535	40.17	0.39
nJKNS	54	4.05	589	44.22	0.43
HJKNS	53	3.98	642	48.20	0.47
AHJNS	52	3.90	694	52.10	0.51
nAJKS	43	3.23	737	55.33	0.54
nAJKN	36	2.70	773	58.03	0.57
anAJK	34	2.55	807	60.59	0.59
nsJNS	34	2.55	841	63.14	0.62
ansAJ	33	2.48	874	65.62	0.64
nsANS	31	2.33	905	67.94	0.67
ansNS	29	2.18	934	70.12	0.69
anJNS	27	2.03	961	72.15	0.71
sAJKN	26	1.95	987	74.10	0.73
ansAS	24	1.80	1,011	75.90	0.74
ansJS	24	1.80	1,035	77.70	0.76
nsJKS	21	1.58	1,056	79.28	0.78
anJKS	20	1.50	1,076	80.78	0.79
aAJKN	16	1.20	1,092	81.98	0.80
anANS	16	1.20	1,108	83.18	0.81
nAKNS	16	1.20	1,124	84.38	0.83
ansJK	15	1.13	1,139	85.51	0.84
nsAJS	13	0.98	1,152	86.49	0.85
aJKNS	13	0.98	1,165	87.46	0.86
anAJS	12	0.90	1,177	88.36	0.87
aHJNS	10	0.75	1,187	89.11	0.87
nsKNS	10	0.75	1,197	89.86	0.88
aAJNS	9	0.68	1,206	90.54	0.89

表 10 は 5 字種構成のみの構成パターン出現比率を表したものである．5 字種は全体の 1%程度を占める．最上位のパターンだけ約 27%であり，それ以外のパターンとの差が大きく，全体としてユニークなパターンが多いことが分かる．

構成数 6

表 11　構成字種 6 のパターン出現頻度

構成字種	用語数	比率 (%)	累積	累積比率 (%)	全体比率 (%)
ansJNS	15	10.79	15	10.79	0.01
nsJKNS	14	10.07	29	20.86	0.02
anAJKS	12	8.63	41	29.50	0.03
nAJKNS	11	7.91	52	37.41	0.04
AHJKNS	10	7.19	62	44.60	0.05
anAJNS	9	6.47	71	51.08	0.05
anJKNS	9	6.47	80	57.55	0.06
nsAJKN	7	5.04	87	62.59	0.06
ansJKS	6	4.32	93	66.91	0.07
ansAJN	5	3.60	98	70.50	0.07
ansAJK	5	3.60	103	74.10	0.08
nsAJNS	5	3.60	108	77.70	0.08
aAJKNS	3	2.16	111	79.86	0.08
ansAJS	3	2.16	114	82.01	0.08
ansJKN	3	2.16	117	84.17	0.09
ansAKS	2	1.44	119	85.61	0.09
nAHJKN	2	1.44	121	87.05	0.09
nsHJKN	2	1.44	123	88.49	0.09
nAHJNS	2	1.44	125	89.93	0.09
anHJNS	2	1.44	127	91.37	0.09

表 11 は 6 字種構成のみの構成パターン出現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，語数は僅かに 139 語である．全 30 パターン中 95%到達は 25 パターン目であり，全体として目立ったパターンはないことが分かります．

構成数 7

表 12 は 7 字種構成のみの構成パターン出現比率を表したものである．7 字種は全体の 0.01%程度を占め，語数は僅かに 18 語であり，全 7 パターンしかない．母数が少なすぎることから，特徴を考察することは難しい．

表 12　構成字種 7 のパターン出現頻度

構成字種	用語数	比率 (%)	累積	累積比率 (%)	全体比率 (%)
ansJKNS	6	33.33	6	33.33	0.00
nsAJKNS	5	27.78	11	61.11	0.01
anAJKNS	3	16.67	14	77.78	0.01
ansAKNS	1	5.56	15	83.33	0.01
ansAJKS	1	5.56	16	88.89	0.01
anHJKNS	1	5.56	17	94.44	0.01
ansAJNS	1	5.56	18	100.00	0.01

構成数 8

8 字種構成は 1 語しかなく，字種構成は“ ansAJKNS”であり，“ 値２－Ri-1×Ｙ ”という数式のような文字列である．特許文章ではこのように全角半角が混合で使われていることが多く，数式表現のような文字列では構成字種が多くなる傾向にある．

4. 辞書見出し語との比較

我々の過去の研究[10]による辞書見出し語での同様の解析結果と今回の結果を比較する．以下の表における“ テキスト ”は今回の特許抄録に対する結果を表す．

4.1 用語全体

表 13 は特許抄録と辞書見出し語の先頭字種毎の用語数の比較である．特徴として辞書見出し語には先頭字種が全角英字，全角数字，半角記号の用語は出現しない．辞書見出し語は全体として全角文字列が少なく，全半角どちらもある文字は全て半角で記述されている．対して特許抄録では，全半角は混合で出現し，意味に違いがある時とない時が存在する．そのため本研究では全半角を区別して扱う．

比率を見ると，漢字とカタカナの合計で 80%程度であることは共通である．また各字種の比率を見ても，全半角の区別を無視すればおおよその比率は同じである．唯一特徴的なのは，特許抄録にはひらがなで始まる複合語が少ない点である．これは専門分野に利用可能なひらがな名詞が少ないからであると，考えられる．

表 13　先頭字種毎の用語数

	テキスト		辞書	
先頭字種	用語数	比率 (%)	用語数	比率 (%)
漢字	81309	59.80	54159	43.61
カタカナ	39907	29.35	54258	43.69
全角英字	7388	5.43	0	0.00
全角数字	4680	3.44	0	0.00
半角数字	985	0.72	2993	2.41
半角英字	590	0.43	6849	5.51
ひらがな	557	0.41	4980	4.01
全角記号	537	0.39	952	0.77
半角記号	19	0.01	0	0.00
合計	135972		124191	

表 14　構成字種数毎の用語数

	テキスト		辞書	
構成数	比率 (%)	累積比率 (%)	比率 (%)	累積比率 (%)
2	60.540	60.540	89.466	89.466
3	31.523	92.064	9.285	98.752
4	6.810	98.874	1.135	99.886
5	1.010	99.884	0.113	99.999
6	0.102	99.986	0.001	100.000
7	0.013	99.999		
8	0.001	100.000		

表 14 は構成字種数毎の用語出現比率の比較である．どちらも 2,3 字種で 90%を超える点は共通している．特徴として，辞書見出し語に比べて構成字種数は特許抄録では少し多い傾向にある．また，辞書見出し語では最大でも 6 構成なのに対して特許抄録では最大 8 構成まで出現する．これは先に挙げた全半角の区別に依るものであると考えられる．

4.2 構成字種数毎

構成字種数毎の結果を比較する．結果のデータ数は，辞書見出し語の 95%到達を基

準として上位パターンのみを記載した．

4.2.1 **構成数** 2

表 15　構成字種 2 のパターン出現頻度

テキスト			辞書		
構成字種	比率 (%)	累積比率 (%)	構成字種	比率 (%)	累積比率 (%)
JK	40.46	40.46	JK	68.83	68.83
JN	31.01	71.46	HJ	18.89	87.72
KN	8.73	80.20	aJ	3.16	90.88
AJ	5.28	85.48	nJ	2.47	93.35
HJ	5.17	90.64	KS	1.66	95.02

表 15 は 2 字種構成のみの用語出現比率を比較したものである．上位パターンには双方共通するパターン（“ JK ”, “ HJ ”，全半角を無視すれば “ AJ ” と “ aJ ”）が多く存在する．逆に特許抄録のみに存在するパターンである “ JN ” や “ KN ” は，“ 段差面２ ” や “ アークチューブ１０ ” など漢字やカタカナの単一字種名詞に文章中の連番が付いた複合語が多く存在していることに起因する．また 95%到達も特許抄録の方が遅く，辞書見出し語に比べて出現するパターンには，ばらつきがあることが分かる．

4.2.2 **構成数** 3

表 16　構成字種 3 のパターン出現頻度

テキスト			辞書		
構成字種	比率 (%)	累積比率 (%)	構成字種	比率 (%)	累積比率 (%)
JKN	45.94	45.94	HJK	34.73	34.73
AJN	12.82	58.76	JKS	24.44	59.17
JNS	6.36	65.12	aJK	10.15	69.31
AJK	5.39	70.51	nJK	7.77	77.08
AKN	3.91	74.42	aKS	7.03	84.11
HJN	3.78	78.21	nKS	5.25	89.36
HJK	3.43	81.64	aJS	2.39	91.75
nJK	2.99	84.63	anJ	1.99	93.74
KNS	2.25	86.88	anK	1.60	95.33

表 16 は 3 字種構成のみの用語出現比率を比較したものである．上位パターンには全半角を無視すれば双方共通するパターン（“ HJK ”, “ AJK ”, “ NJK ”, “ JKS ”, “ AJN ”, “ AKN ”）が多く存在する．逆に特許抄録のみに存在するパターンである “ JKN ” や “ HJN ” は，2 字種の上位パターンである “ JK ”, “ HJ ” に連番が付いた複合語が多く存在していることに起因すると考えられる．比率で見ると，辞書見出し語 95%到達時にテキストは 87%と累積比率の上昇が緩やかである．

4.2.3 **構成数** 4

表 17　構成字種 4 のパターン出現頻度

テキスト			辞書		
構成字種	比率 (%)	累積比率 (%)	構成字種	比率 (%)	累積比率 (%)
AJKN	28.87	28.87	nJKS	33.43	33.43
JKNS	17.18	46.05	aJKS	26.83	60.26
HJKN	9.49	55.53	anKS	11.50	71.75
AJNS	7.73	63.27	anJS	9.08	80.84
AKNS	3.32	66.59	anJK	7.88	88.72
nAJK	3.20	69.79	HJKS	2.20	90.92
AHJN	2.92	72.71	aHJK	2.20	93.12
aAJK	1.97	74.68	nHJK	1.70	94.82
anJK	1.84	76.52	asJK	1.63	96.45

表 17 は 4 字種構成のみの用語出現比率を比較したものである．上位パターンには全半角を無視すれば双方共通するパターン（“ AJKN ”,“ JKNS ”,“ HJKN ”,“ HJKN ”,“ AJNS ”, “ AKNS ”）が多く存在する．比率で見ると，辞書見出し語 95%到達時にテキストは 76%と累積比率の上昇は緩やかである．

4.2.4 **構成数** 5

表 18　構成字種 5 のパターン出現頻度

テキスト			辞書		
構成字種	比率 (%)	累積比率 (%)	構成字種	比率 (%)	累積比率 (%)
AJKNS	27.85	27.85	anJKS	79.29	79.29
AHJKN	6.53	34.38	nsJKS	10.00	89.29
nAJNS	5.78	40.17	aHJKS	2.14	91.43
nJKNS	4.05	44.22	anHJK	2.14	93.57
HJKNS	3.98	48.20	ansKS	2.14	95.71

表 18 は 5 字種構成のみの用語出現比率を比較したものである．最上位パターンは全半角を無視すれば双方共通するパターンである．以降のパターンは辞書見出し語においての比率が低く、比較するのはむずかしい．また全体の比率だけ見ると，辞書見出し語 95%到達時にテキストは 48%と累積比率の上昇は非常に緩やかである．

4.2.5 **構成数** 6,7,8

辞書見出し語において、構成数 6 の“ anHJKS ”の一種しか存在しない．そのため比較することは難しい。また，構成数 7,8 は辞書見出し語には存在せず，特許抄録においても合わせて 0.14%しか存在せず，これらに対する考察の必要性は低い．

4.3 先頭字種毎

次に先頭字種毎に分けて結果を比較する．これにより各字種の特性を考察する．

4.3.1 字種構成パターン数

表 19 は先頭字種毎の字種パターン数の累積 90%，95%，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．

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

表を見ると，漢字・カタカナの 95%到達時のみ，テキストの相対比率がより低い値になっていることが分かる．一般的には相対比率は辞書見出し語の方が低く，パターンに偏りがあるが，漢字・カタカナを先頭に含む複合語の場合には，相対的にばらつきが少ないということが分かる．

表 19　先頭字種毎のパターン数と相対比率

先頭字種	割合 (%)	テキスト		辞書	
		パターン	相対比率 (%)	パターン	相対比率 (%)
漢字	90	12	7.36	2	5.88
	95	19	11.66	4	11.76
	100	163	100.00	34	100.00
カタカナ	90	9	7.96	3	9.38
	95	13	11.50	5	15.63
	100	113	100.00	32	100.00
ひらがな	90	9	30.00	1	8.33
	95	12	40.00	2	16.67
	100	30	100.00	12	100.00
半角数字	90	24	36.36	5	16.67
	95	35	53.03	8	26.67
	100	66	100.00	30	100.00
半角英字	90	27	39.71	8	22.22
	95	40	58.82	12	33.33
	100	68	100.00	36	100.00
全角記号	90	33	45.83	5	27.78
	95	45	62.50	8	44.44
	100	72	100.00	18	100.00
全角数字	90	20	20.62	-	-
	95	33	34.02	-	-
	100	97	100.00	-	-
全角英字	90	23	21.50	-	-
	95	37	34.58	-	-
	100	107	100.00	-	-
半角記号	90	5	71.43	-	-
	95	6	85.71	-	-
	100	7	100.00	-	-

4.3.2 漢字

表 20 は先頭字種漢字の構成数毎の用語出現比率を表したものである．辞書見出し語では 2 字種構成のみで 94%と大半を占めるのに対し、特許抄録では 2,3 字種構成で累積 93%となる．このように特許抄録において構成字種数が大きくなる理由は連番に依り，2 字種構成や 3 字種構成の末尾に数字が付くことに起因すると考えられる．

表 21 は先頭字種漢字の構成パターン出現比率である．比較すると辞書見出し語の上位パターンは特許抄録にも存在する．また，特許抄録の上位パターンである“ JN ”,“ nJ ”, “ JKN ”, “ AJN ”など“ n ”, “ N ”を含むものは連番に依るものであると考えられ，特

許特有の傾向であると推測される．

表 20　構成字種数毎の用語数(漢字)

	テキスト		辞書	
構成数	比率 (%)	累積比率 (%)	比率 (%)	累積比率 (%)
2	63.257	63.257	94.357	94.357
3	30.176	93.434	5.323	99.681
4	5.705	99.139	0.270	99.950
5	0.764	99.903	0.050	100.000
6	0.089	99.991		
7	0.007	99.999		
8	0.001	100.000		

表 21　字種構成毎の用語数(漢字)

テキスト			辞書		
構成字種	比率 (%)	累積比率 (%)	構成字種	比率 (%)	累積比率 (%)
JN	30.53	30.53	JK	58.70	58.70
JK	20.94	51.47	HJ	30.80	89.50
JKN	13.83	65.30	HJK	3.57	93.07
AJN	5.65	70.95	nJ	2.13	95.19
HJ	4.91	75.86	aJ	1.63	96.83
AJ	3.70	79.56	JS	1.02	97.85
JNS	2.78	82.34	JKS	0.59	98.44
nJ	2.61	84.95	aJK	0.45	98.89
HJN	1.82	86.77	nJK	0.29	99.18
AJKN	1.61	88.38	anJ	0.08	99.27

4.3.3　カタカナ

表 22 は先頭字種カタカナの構成数毎の出現比率を表したものである．辞書見出し語では 2,3 字種構成のみで累積 99%と大半を占めるのに対し、特許抄録では 2,3.4 字種構成で累積 99%となる．このように特許抄録において構成字種数が大きくなる理由は漢字同様，連番に依るものであると考えられる．

表 22　構成字種数毎の用語数(カタカナ)

	テキスト		辞書	
構成数	比率 (%)	累積比率 (%)	比率 (%)	累積比率 (%)
2	61.86	61.86	89.91	89.91
3	31.03	92.89	9.36	99.26
4	6.29	99.18	0.67	99.94
5	0.72	99.90	0.06	100.00
6	0.09	99.98		
7	0.02	100.00		

表 23　字種構成毎の用語数(カタカナ)

テキスト			辞書		
構成字種	比率 (%)	累積比率 (%)	構成字種	比率 (%)	累積比率 (%)
JK	40.75	40.75	JK	82.36	82.36
JKN	20.04	60.79	JKS	4.26	86.62
KN	17.43	78.21	HJK	3.04	89.65
AKN	2.74	80.95	KS	2.99	92.64
AJKN	2.21	83.16	nK	2.50	95.14
KNS	1.87	85.03	aK	1.27	96.41
AJK	1.84	86.87	HK	0.76	97.18
nK	1.83	88.70	aJK	0.64	97.81
JKNS	1.74	90.45	nJK	0.55	98.36
HJK	1.65	92.10	aKS	0.33	98.70

表 23 は先頭字種カタカナの構成パターンの比率である．比較すると辞書見出し語の上位パターンは特許抄録にも存在する．また，特許抄録の上位パターンである n,N を含むものは漢字同様，連番に依るものであると考えられ，特許特有の傾向であると推測される．

4.3.4　ひらがな

表 24 は先頭字種ひらがなの構成数毎の比率を表したものである．辞書見出し語では 2 字種構成のみで 91%と大半を占めるのに対し、特許抄録では 2,3 字種構成で累積 90%となる．このように特許抄録において構成字種数が大きくなる理由は連番に依り，2 字種構成や 3 字種構成の末尾に数字が付くことに起因すると考えられる．

表 24　構成字種数毎の用語数(ひらがな)

	テキスト		辞書	
構成数	比率 (%)	累積比率 (%)	比率 (%)	累積比率 (%)
2	56.55	56.55	91.10	91.10
3	33.39	89.95	8.80	99.90
4	8.62	98.56	0.08	99.98
5	1.08	99.64	0.02	100.00
6	0.36	100.00		

表 25　字種構成毎の用語数(ひらがな)

テキスト			辞書		
構成字種	比率 (%)	累積比率 (%)	構成字種	比率 (%)	累積比率 (%)
HJ	47.22	47.22	HJ	86.61	86.61
HJN	20.11	67.32	HJK	8.47	95.08
HN	4.85	72.17	HK	4.42	99.50
HJK	4.67	76.84	aHJ	0.16	99.66
HKN	3.23	80.07	HJS	0.10	99.76
HJKN	3.23	83.30	aH	0.06	99.82
HK	2.87	86.18	nHJ	0.04	99.86
AHJN	2.87	89.05	HJKS	0.04	99.90
nHJ	2.15	91.20	nHJK	0.04	99.94
nH	1.08	92.28	HS	0.02	99.96

表 25 は先頭字種ひらがなの構成パターンの比率である．比較すると辞書見出し語の上位パターンは特許抄録にも存在する．辞書見出し語は上位 3 種で累積 99%に到達し，残りのパターンは出現頻度が低い．特許抄録においても上位 2 種に比べて他のパターンは出現頻度が低い，また，第 2 位のパターンである“ HJN ”は最上位“ HJ ”に連番表現“ N ”が付いたものであると考えられ，実質的には“ HJ ”のみであることが推測される．

4.3.5　半角英字

表 26 は先頭字種半角英字の構成数毎の比率を表したものである．辞書見出し語では 2,3 字種構成のみで累積 92%と大半を占めるのに対し、特許抄録では 2,3,4 字種構成で累積 96%となる．このように特許抄録において構成字種数が大きくなる理由は数式表現や数値範囲表現，化学式などに依るものであると考えられる．

表 26　構成字種数毎の用語数(半角英字)

	テキスト		辞書	
構成数	比率 (%)	累積比率 (%)	比率 (%)	累積比率 (%)
2	50.51	50.51	66.07	66.07
3	33.56	84.07	26.06	92.13
4	12.20	96.27	7.08	99.21
5	2.88	99.15	0.77	99.99
6	0.85	100.00	0.01	100.00

表 27　字種構成毎の用語数(半角英字)

テキスト			辞書		
構成字種	比率 (%)	累積比率 (%)	構成字種	比率 (%)	累積比率 (%)
aJ	17.80	17.80	aJ	38.34	38.34
an	17.63	35.42	aK	13.74	52.08
as	6.61	42.03	aKS	8.85	60.93
ans	5.93	47.97	aJK	8.48	69.41
anJ	5.08	53.05	an	7.18	76.60
aJK	4.41	57.46	aS	6.25	82.84
aK	3.90	61.36	aJKS	3.64	86.48
anS	3.90	65.25	aJS	3.55	90.03
aS	2.20	67.46	anJ	2.09	92.12
aJN	2.20	69.66	anJS	1.42	93.53

表 27 は先頭字種半角英字の構成パターン出現比率である．比較すると辞書見出し語の上位パターンは特許抄録にも存在する．また，これまでの字種同様に特許抄録の上位パターンには“ n ”, “ N ”を含むものが多い．

4.3.6　半角数字

表 28 は先頭字種半角数字の構成数毎の比率を表したものである．辞書見出し語では 2,3 字種構成で 90%と大半を占めるのに対し、特許抄録では 2,3,4 字種構成で累積 95%となる．このように特許抄録において構成字種数が大きくなる理由は，本研究では数式表現を名詞相当語として扱ったことに起因すると考えられる．文章中における数式には英数字と記号の組み合わせの他に，変数名として漢字，カタカナなどの日本語字

種も用いられるため，必然的に構成字種数は大きくなりやすい．

表 28　構成字種数毎の用語数(半角数字)

	テキスト		辞書	
構成数	比率 (%)	累積比率 (%)	比率 (%)	累積比率 (%)
2	36.14	36.14	55.03	55.03
3	39.19	75.33	33.98	89.01
4	19.90	95.23	10.12	99.13
5	4.37	99.59	0.87	100.00
6	0.30	99.90		
7	0.10	100.00		

表 29　字種構成毎の用語数(半角数字)

テキスト			辞書		
構成字種	比率 (%)	累積比率 (%)	構成字種	比率 (%)	累積比率 (%)
nsS	13.20	13.20	nJ	53.29	53.29
nS	11.78	24.97	nJK	14.67	67.96
ns	7.92	32.89	nKS	13.40	81.36
an	7.82	40.71	nJKS	6.95	88.31
ansS	7.11	47.82	nHJ	2.27	90.58
nJS	5.89	53.71	nK	1.50	92.08
ans	5.69	59.39	nJS	1.40	93.48
nA	4.67	64.06	anJ	1.37	94.85
anS	4.67	68.73	anKS	1.34	96.19
nsJS	3.65	72.39	anJKS	0.74	96.93

表 29 は先頭字種半角英字の構成パターン出現比率である．これまでの字種とは異なり，特許抄録と辞書見出し語では大きく異なる結果となった．これは特許抄録では英字と数字の組み合わせによる変数名や物質名などの固有名詞に似た複合語が多く出現することに起因すると考えられる．また今回は解析対象に数式を含めたことに依り，辞書見出し語にはないパターンが多く抽出されたと考えられる．

4.3.7　全角記号

表 30 は先頭字種全角記号の構成数毎の出現比率を表したものである．先頭全角記号は双方ともに全体の比率から見て出現頻度は低く、比較が難しく，特許抄録と辞書見出し語で差異は小さい．

表 30　構成字種数毎の用語数(全角記号)

	テキスト		辞書	
構成数	比率 (%)	累積比率 (%)	比率 (%)	累積比率 (%)
2	37.99	37.99	54.10	54.10
3	48.23	86.22	34.87	88.97
4	10.99	97.21	11.03	100.00
5	2.79	100.00		

表 31　字種構成毎の用語数(全角記号)

テキスト			辞書		
構成字種	比率 (%)	累積比率 (%)	構成字種	比率 (%)	累積比率 (%)
JKS	7.45	7.45	JS	28.36	28.36
aS	6.89	14.34	KS	23.95	52.31
AS	6.33	20.67	JKS	19.96	72.27
NS	6.15	26.82	nKS	9.56	81.83
KS	5.96	32.77	nJKS	8.19	90.02
nAK	5.03	37.80	aKS	2.63	92.65
JNS	5.03	42.83	aS	1.47	94.12
JS	4.66	47.49	nJS	1.16	95.27
AKN	4.47	51.96	anKS	1.16	96.43
KNS	3.91	55.87	aJKS	1.05	97.48

表 31 は先頭字種全角記号の構成パターン出現比率である．比較すると辞書見出し語の上位パターンは特許抄録にも存在する．また辞書見出し語は全 18 パターンなのに対して，特許抄録では全 73 パターンと構成にばらつきが多い．また半角数字同様，数式表現も多いことから，構成パターンの種類数が多くなる傾向にあると考えられる．

4.3.8　全角英字

表 32 は先頭字種全角英字の構成字種数毎の用語出現頻度である．先頭字種全角英字は辞書見出し語には出現せず，特許抄録のみに出現する．これは辞書見出し語では全半角の区別がなく，半角で表現することに起因する．2,3,4 字種で累積 96%に到達する．この傾向はその他の字種と同様の傾向である．

表 32　構成字種数毎の用語数(全角英字)

構成数	用語数	比率 (%)	累積	累積比率 (%)
2	3030	41.01	3030	41.01
3	2926	39.60	5956	80.62
4	1178	15.94	7134	96.56
5	231	3.13	7365	99.69
6	21	0.28	7386	99.97
7	2	0.03	7388	100.00

表 33　字種構成毎の用語数(全角英字)

構成字種	用語数	比率 (%)	累積	累積比率 (%)
AJ	1341	18.15	1341	18.15
AJN	767	10.38	2108	28.53
AJK	696	9.42	2804	37.95
AN	623	8.43	3427	46.39
AK	522	7.07	3949	53.45
AKN	499	6.75	4448	60.21
AJKN	436	5.90	4884	66.11
ANS	298	4.03	5182	70.14
nA	215	2.91	5397	73.05
aA	196	2.65	5593	75.70
AJNS	169	2.29	5762	77.99
nAJ	118	1.60	5880	79.59
AS	112	1.52	5992	81.10
AKNS	109	1.48	6101	82.58
AJS	91	1.23	6192	83.81
anA	77	1.04	6269	84.85
nAS	72	0.97	6341	85.83
nAK	68	0.92	6409	86.75
AJKNS	65	0.88	6474	87.63
nAJK	50	0.68	6524	88.31
AKS	44	0.60	6568	88.90
AJKS	40	0.54	6608	89.44
nANS	40	0.54	6648	89.98

表 33 は先頭字種全角英字の字種構成毎の用語出現頻度である．上位パターンには他の字種同様 N を含むものが多い．

4.3.9　**全角数字**

表 34　構成字種数毎の用語数(全角数字)

構成数	用語数	比率 (%)	累積	累積比率 (%)
2	1988	42.48	1988	42.48
3	1978	42.26	3966	84.74
4	557	11.90	4523	96.65
5	153	3.27	4676	99.91
6	2	0.04	4678	99.96
7	2	0.04	4680	100.00

表 35　字種構成毎の用語数(全角数字)

構成字種	用語数	比率 (%)	累積	累積比率 (%)
NS	711	15.19	711	15.19
JN	703	15.02	1414	30.21
JNS	440	9.40	1854	39.62
JKN	439	9.38	2293	49.00
ANS	429	9.17	2722	58.16
AN	259	5.53	2981	63.70
KN	230	4.91	3211	68.61
KNS	199	4.25	3410	72.86
AJN	133	2.84	3543	75.71
nANS	100	2.14	3643	77.84
JKNS	97	2.07	3740	79.91
aNS	92	1.97	3832	81.88
AJNS	71	1.52	3903	83.40
AKN	62	1.32	3965	84.72
aN	61	1.30	4026	86.03
AJKN	46	0.98	4072	87.01
sNS	42	0.90	4114	87.91
AKNS	37	0.79	4151	88.70
nNS	34	0.73	4185	89.42
HJN	29	0.62	4214	90.04

表 34 は先頭字種全角数字の構成字種数毎の用語出現頻度である．先頭字種全角数字は辞書見出し語には出現せず，特許抄録のみに出現する．これは辞書見出し語では全半角の区別がなく，半角で表現することに起因する．2,3,4 字種で累積 96%に到達する．この傾向はその他の字種と同様の傾向である．

表 35 は先頭字種全角数字の構成パターン出現頻度である．先頭が全角数字である場合の多くは“０．０３５％”(“NS”)や“２００ｋΩ”(“ANS”)などの数字と単位記号で構成されるパターンである．本研究ではこのような語も複合語して対象としたために辞書見出し語にはないパターンが多く抽出されたのだと考えられる．

4.3.10 半角記号

表 36　構成字種数毎の用語数(半角記号)

構成数	用語数	比率 (%)	累積	累積比率 (%)
2	7	36.84	7	36.84
3	12	63.16	19	100.00

表 37　字種構成毎の用語数(半角記号)

構成字種	用語数	比率 (%)	累積	累積比率 (%)
Ans	7	36.84	7	36.84
As	4	21.05	11	57.89
sA	3	15.79	14	73.68
asJ	2	10.53	16	84.21
sAS	1	5.26	17	89.47
sJK	1	5.26	18	94.74
asN	1	5.26	19	100.00

表 36 および表 37 は先頭字種半角記号の出現頻度を表したものである．先頭字種半角記号は辞書見出し語には存在せず，特許抄録だけに見られるパターンである．ただし，語数は僅か 19 しか出現せず，その多くは“-CO-基”(“asJ”)のような化学構造式を表すものである．

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構成字種という観点からの特性が明らかになった．構成字種は，特許抄録・辞書見出し語ともに構成字種 2～4 だけで 98%と出現頻度に偏りが存在する．また，先頭字種毎に統計は異なり，先頭に日本語文字を持つ場合，英数字や記号を先頭に持つ複合語に比べて構成数が少ないという特性が明らかになった．さらに，その構成は，抄録では連番，単位記号や数値範囲のために英数字や記号が多く出現するなど，対象となるコーパスによって特性が異なる．この特性を調査・研究することで，文書の分野推定や分類などに用いることも可能であると考えられる。

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) 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) 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．
10) 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

Vol.2011-NL-204 No.2
2011/11/21

Vol.2011-NL-204 No.2　【正誤表】

p.4 表 3 先頭字種毎の抽出用語数

§ 3.1 表 3 先頭字種毎の抽出用語数

正

先頭字種	用語数	比率 (%)
漢字	81,309	59.80
カタカナ	39,907	29.35
全角英字	7,388	5.43
全角数字	4,680	3.44
半角数字	985	0.72
半角英字	590	0.43
ひらがな	557	0.41
全角記号	532	0.39
半角記号	19	0.01
合計	135967	

誤

先頭字種	用語数	比率 (%)
漢字	81309	59.8
カタカナ	39907	29.35
全角英字	7388	5.43
全角数字	4680	3.44
半角数字	985	0.72
半角英字	590	0.43
ひらがな	557	0.41
全角記号	537	0.39
半角記号	19	0.01
合計	135972	

p.8 表 13 先頭字種毎の用語数

§ 4.1 表 13 先頭字種毎の用語数

正

先頭字種	テキスト 用語数	テキスト 比率 (%)	辞書 用語数	辞書 比率 (%)
漢字	81,309	59.80	54,159	43.61
カタカナ	39,907	29.35	54,258	43.69
全角英字	7,388	5.43	0	0.00
全角数字	4,680	3.44	0	0.00
半角数字	985	0.72	2,993	2.41
半角英字	590	0.43	6,849	5.51
ひらがな	557	0.41	4980	4.01
全角記号	532	0.39	952	0.77
半角記号	19	0.01	0	0.00
合計	135967		124191	

誤

先頭字種	テキスト 用語数	テキスト 比率 (%)	辞書 用語数	辞書 比率 (%)
漢字	81309	59.8	54159	43.61
カタカナ	39907	29.35	54258	43.69
全角英字	7388	5.43	0	0.00
全角数字	4680	3.44	0	0.00
半角数字	985	0.72	2993	2.41
半角英字	590	0.43	6849	5.51
ひらがな	557	0.41	4980	4.01
全角記号	537	0.39	952	0.77
半角記号	19	0.01	0	0.00
合計	135972		124191	

p.10 表 19 先頭字種毎のパターン数と相対比率

§ 4.3.1 表 19 先頭字種毎のパターン数と相対比率

正

先頭字種	割合	テキスト		辞書	
	(%)	パターン	相対比率 (%)	パターン	相対比率 (%)
漢字	90	12	7.36	2	5.88
	95	19	11.66	4	11.76
	100	163	100.00	34	100.00
カタカナ	90	9	7.96	3	9.38
	95	13	11.50	5	15.63
	100	113	100.00	32	100.00
ひらがな	90	9	30.00	1	8.33
	95	12	40.00	2	16.67
	100	30	100.00	12	100.00
半角数字	90	24	36.36	5	16.67
	95	35	53.03	8	26.67
	100	66	100.00	30	100.00
半角英字	90	27	39.71	8	22.22
	95	40	58.82	12	33.33
	100	68	100.00	36	100.00
全角記号	90	22	39.29	5	27.78
	95	32	57.14	8	44.44
	100	56	100.00	18	100.00
全角数字	90	20	20.62	-	-
	95	33	34.02	-	-
	100	97	100.00	-	-
全角英字	90	23	21.50	-	-
	95	37	34.58	-	-
	100	107	100.00	-	-
半角記号	90	5	71.43	-	-
	95	6	85.71	-	-
	100	7	100.00	-	-

誤

先頭字種	割合	テキスト		辞書	
	(%)	パターン	相対比率 (%)	パターン	相対比率 (%)
漢字	90	12	7.36	2	5.88
	95	19	11.66	4	11.76
	100	163	100	34	100
カタカナ	90	9	7.96	3	9.38
	95	13	11.5	5	15.63
	100	113	100	32	100
ひらがな	90	9	30	1	8.33
	95	12	40	2	16.67
	100	30	100	12	100
半角数字	90	24	36.36	5	16.67
	95	35	53.03	8	26.67
	100	66	100	30	100
半角英字	90	27	39.71	8	22.22
	95	40	58.82	12	33.33
	100	68	100	36	100
全角記号	90	33	45.83	5	27.78
	95	45	62.5	8	44.44
	100	72	100	18	100
全角数字	90	20	20.62	-	-
	95	33	34.02	-	-
	100	97	100	-	-
全角英字	90	23	21.5	-	-
	95	37	34.58	-	-
	100	107	100	-	-
半角記号	90	5	71.43	-	-
	95	6	85.71	-	-
	100	7	100	-	-

p.13 表 30 構成字種数毎の用語数(全角記号)

§4.3.7 表 30 構成字種数毎の用語数(全角記号)

正

構成数	テキスト		辞書	
	比率 (%)	累積比率 (%)	比率 (%)	累積比率 (%)
2	40.41	40.41	54.10	54.10
3	45.86	86.28	34.87	88.97
4	11.09	97.37	11.03	100.00
5	2.63	100.00	-	-

誤

構成数	テキスト		辞書	
	比率 (%)	累積比率 (%)	比率 (%)	累積比率 (%)
2	37.99	37.99	54.10	54.10
3	48.23	86.22	34.87	88.97
4	10.99	97.21	11.03	100.00
5	2.79	100.00	-	-

p.13 表 31 字種構成毎の用語数(全角記号)

§4.3.7 表 31 字種構成毎の用語数(全角記号)

正

テキスト			辞書		
構成字種	比率 (%)	累積比率 (%)	構成字種	比率 (%)	累積比率 (%)
AS	9.77	9.77	JS	28.36	28.36
NS	9.02	18.80	KS	23.95	52.31
JKS	8.46	27.26	JKS	19.96	72.27
ANS	7.52	34.77	nKS	9.56	81.83
nAS	7.52	42.29	nJKS	8.19	90.02
aS	7.14	49.44	aKS	2.63	92.65
KS	5.83	55.26	aS	1.47	94.12
JNS	5.83	61.09	nJS	1.16	95.27
JS	5.64	66.73	anKS	1.16	96.43
nS	3.01	69.74	aJKS	1.05	97.48

誤

テキスト			辞書		
構成字種	比率 (%)	累積比率 (%)	構成字種	比率 (%)	累積比率 (%)
JKS	7.45	7.45	JS	28.36	28.36
aS	6.89	14.34	KS	23.95	52.31
AS	6.33	20.67	JKS	19.96	72.27
NS	6.15	26.82	nKS	9.56	81.83
KS	5.96	32.77	nJKS	8.19	90.02
nAK	5.03	37.8	aKS	2.63	92.65
JNS	5.03	42.83	aS	1.47	94.12
JS	4.66	47.49	nJS	1.16	95.27
AKN	4.47	51.96	anKS	1.16	96.43
KNS	3.91	55.87	aJKS	1.05	97.48

p.15 表 37 字種構成毎の用語数(半角記号)

§4.3.10 表 37 字種構成毎の用語数(半角記号)

正

構成字種	用語数	比率 (%)	累積	累積比率 (%)
ans	7	36.84	7	36.84
as	4	21.05	11	57.89
sA	3	15.79	14	73.68
asJ	2	10.53	16	84.21
sAS	1	5.26	17	89.47
sJK	1	5.26	18	94.74
asN	1	5.26	19	100.00

誤

構成字種	用語数	比率 (%)	累積	累積比率 (%)
Ans	7	36.84	7	36.84
As	4	21.05	11	57.89
sA	3	15.79	14	73.68
asJ	2	10.53	16	84.21
sAS	1	5.26	17	89.47
sJK	1	5.26	18	94.74
asN	1	5.26	19	100.00