2008－NL－184 (17)
2008／3／28

辞書見出し語の7文字漢字熟語を対象とした語基構成の解析

梅木 定博, 後藤 智範
神奈川大学 理学部 情報科学科

日本語のテキストにおいて、主要な概念・テーマは漢字熟語または漢字熟語を含む名詞句に表現されることが多い。特に数文字以上の漢字熟語は、より短い漢字熟語、すなわち語基(word base)から構成され、統語的、意味的構造を有している。大規模な漢字熟語集合について、これらの構造を分析することは漢字熟語の造語構造、形態素解析、関連語の選定、未知語の推定など様々な自然言語解析に有用な基礎データを提供するものと考えられる。

本研究は、一般辞書および専門用語辞書の見出し語から7文字の漢字熟語を対象に、構成語基の観点から品詞列パターンおよび構成語基の係り受けパターンについて調査・分析した。

Analysis to Seven-Kanji Compound Words in Entry Terms of Dictionaries

Sadahiro Umeki, Tomonori Gotoh
Department of Information Science, Kanagawa University

Kanji compound words or noun phrase consisted in them intend to explain key concepts or themes in Japanese texts.. Especially long kanji compound words have these characteristics in academic papers or patent documents. Long kanji compound word, which has five letters more consists of short word bases and have syntactically and semantically structures. It should be much beneficial to study to a large set of long kanji words based on word base sequence patterns.

Our research examines the patterns of the large set of kanji compound words with seven letters which are contained in entry terms of the various kinds of dictionaries. This paper reports the occurrences of kanji compound words and the number of parts of speech sequence per a word base sequence pattern.

1. はじめに

日本語のテキストにおいて、主要な概念・テーマは漢字熟語または漢字熟語を含む名詞句に表現される。特に数文字以上の漢字熟語は、複雑な構造を有し、その構造を分析することは高精度の形態素解析、関連語の選定、未知語の推定などに有効と考えられる。

漢字熟語を対象とした複合語の構造解析の研究は、野村の研究[1][2]を嚆矢とし、大規模な医学用語データベースを対象とした調査[3][4]、語彙概念構造に基づく解析の研究[5][6]などがある。

本研究では、コーパスとしては大規模な漢字熟語集合(10^4)を対象とし、漢字熟語を表層的な特性、具体的には、その長さ(文字数)、構成語基の長さ／構成語基数の観点から、構成語基毎に下記の項目を明らかにしようとするものである。

(1) 出現熟語数
(2) 構成語基パターン毎の品詞列パターン
(3) 構成語基パターン毎の係り受け

5文字熟語および6文字熟語の上記項目の調査分析結果については既に報告した[7][8]。本研究では、7文字漢字熟語を対象とし上記項目の解析結果について報告する。

2. コーパスおよび解析手順

コーパスおよび対象漢字塾語群の特性は、上述5文字、6文字漢字熟語群と同様であり、詳細は省略する。本研究の対象漢字熟語集合、すなわち同一の特性をもつ漢字熟語の総数は6527であった。

解析手順、使用された品詞体系は6文字熟語集合のときと同一であり省略する。

3. 結果

対象熟語は3、4、5、6、7語基のいずれかに分割された。

3.1 語基構成パターン

漢字熟語の構成語基数毎の出現頻度を表3.1に示す。

表3.1 構成語基数毎の熟語数

構成基数	3	4	5	6	7
熟語数	128	5357	1026	13	3
比率(%)	1.96	82.07	15.72	0.20	0.05

3語基に分割するためには、少なくとも1語は3文字以上で分割される必要があるために、調査した7文字漢字熟語中では、2%ほどしか2語基は見られなかった。

表3.2 構成語基数毎の構成パターン数

分割語基長	種類		
3語基 (15)		5語基	
1 [2] 5 [1]	3	1 [3] 2 [2]	10
1 [1] 2 [1] 4 [1]	6	1 [4] 3 [1]	5
1 [1] 3 [2]	3	6語基	
2 [2] 3 [1]	3	1 [5] 2 [1]	6
4語基 (20)		7語基	
1 [3] 4 [1]	4	1 [7]	1
1 [2] 2 [1] 3 [1]	12		
1 [1] 2 [3]	4		

表3.2は、7字漢字熟語の構成語基の全ての組み合わせを表している。

3.2 品詞列パターンの分析

表3.2に挙げた構成語基数毎に出現した品詞列(並び)パターンの数を表3.3に挙げる。

表3.3 構成語基数毎の品詞列パターン数

構成基数	3	4	5	6	7
パターン数	17	225	277	12	1
比率(%)	3.20	42.29	52.07	2.26	0.19

表3.3から、5語基が品詞列パターンで最多となった。

表3.4は、構成語基数単位に品詞列を整列した結果の一端を表している。この表から、語基数(行)および先頭品詞(列)の観点からその特性を明確にするために、それぞれの比率を算出した。

表3.4 構成語基数の先頭品詞毎の品詞列パターン数

語基数	サ変	形動	形容	接頭	動詞	名詞	数詞
3語基	3	0	1	1	1	9	0
4語基	51	0	13	18	10	82	8
5語基	37	5	17	67	4	105	37
6語基	1	0	2	3	0	4	2
7語基	0	0	0	0	0	0	1
計	92	5	33	89	15	200	48

表3.5は特定の先頭品詞で開始する品詞列パターンの総数で除しその比率を表したものである。

表3.5 構成語基数毎の品詞列パターン数比率

語基数	サ変	形動	形容	接頭	動詞	名詞	数詞
3語基	3.3	0	3.0	1.1	6.7	4.5	0
4語基	55.4	0	39.4	20.2	66.7	41	16.7
5語基	40.2	100	51.5	75.3	26.7	52.5	77.1
6語基	1.1	0	6.1	3.4	0	2	4.2
7語基	0	0	0	0	0	0	2.1

表3.5から、名詞、サ変以外を見ると、4語基では、動詞を先頭品詞に持つものが多いのに対して、5語基になると、接頭辞、数詞の割合がかなり増えていることがわかる。

表3.6 構成語基の先頭品詞毎の品詞列パターン数比率

語基数	サ変	形動	形容	接頭	動詞	名詞	数詞
3語基	20	0	6.7	6.7	6.7	60	0
4語基	28	0	7.1	9.9	5.5	45.1	4.4
5語基	13.6	1.8	6.3	24.6	1.5	38.6	13.6
6語基	8.3	0	16.7	25	0	33.3	16.7
7語基	0	0	0	0	0	0	100

表3.6は語基数単位で、特定品詞で開始する品詞列パターンの全体の比率を表している。この表に基づいて、以下に語基数毎にその特性について記述する。

(1) 3語基

表3.6から60%が名詞ではじまっていることがわかる。

(2) 4 語基

表3.6から、73%が名詞またはサ変で始まっていることがわかる。

(3) 5 語基

表3.6から、52%が名詞またはサ変で始まっていることがわかる。一方で、38%が数詞または接頭で始まっていることがわかる。この分、形動を除き他の先頭品詞の比率は減少している。

(4) 6 語基

サ変、名詞の比率は減少し、数詞、特に形容詞で始まる品詞列パターンが5語基までと比較して増加していることがわかる。

(5) 7 語基

熟語総数が3語(表3.1)で、数詞が変化するのみであり、先頭品詞の特性がコーパス外の6語基からなる6文字漢字熟語に対して一致するかどうかは定かではない。

(例)

四	腕	二	脚	二	頭	体
三	腕	二	脚	二	頭	体
二	腕	二	脚	二	頭	体
数詞	名詞	数詞	名詞	数詞	名詞	名詞

以上から、全体として語基数の増加により、下記の傾向が判明する。

減少: サ変、形容動詞
増加: 接頭辞、数詞

表 3.7 係り受けパターン比率

3 語基	3A	3B
パターン数	13	11
比率	54	46

4 語基	4A	4B	4C	4D	4E	4F
パターン数	149	57	39	60	45	1
比率	42	16	11	17	13	0.3

5 語基	5A	5B	5C	5D	5E	5F	
パターン数	118	71	25	70	5	10	
比率	31	19	7	19	1	3	
	5G	5H	5I	5J	5K	5L	5M
	1	42	4	7	5	5	13
	0.3	11	1	2	1	1	3

6 語基	6A	6B	6C	6D	6E	6F	6G	6H	6I
パターン数	1	1	1	1	1	4	1	1	1
比率	8.3	8.3	8.3	8.3	8.3	33	8.3	8.3	8.3

3.3 係り受けパターン

本節では、構成語基数毎の係り受けパターンと品詞列パターンとの出現傾向についての結果を述べる。

表 3.7 に語基数毎の出現した係り受けパターンの種類を挙げる。この表から、出現頻度の極端に低い係り受けパターンを除けば、語基数が増加すると特定係り受けパターンの出現の偏差が減少することがわかる。

以下、語基数毎に出現した係り受けパターンを挙げる。

(1) 3 語基

図 3.1 に示す 2 種類の係り受けパターンが見られた。

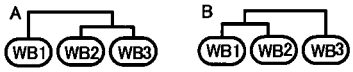


図 3.1 3 語基の係り受けパターン

(WB*は出現する順序での語基を表している様)

(2) 4 語基

図 3.2 に示す 6 種類の係り受けパターンが見られた。パターン D は並列構造を持っていることがわかり表 3.7 からもこの構造による品詞列パターンは 1 パターンしかないことがわかる。

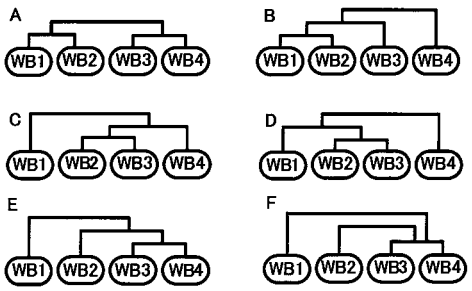


図 3.2 4 語基の係り受けパターン

(3) 5 語基

図 3.3 に示す 13 種類の係り受けパターンが見られた。

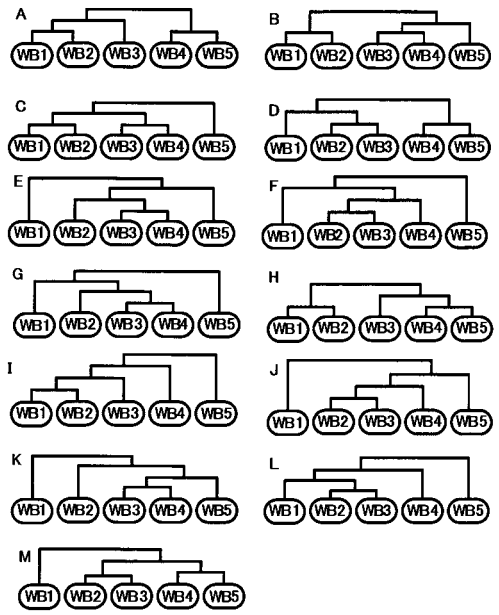


図 3.3　5 語基の係り受けパターン

(4) 6 語基構成熟語

図 3.4 に示す 9 種類の係り受けパターンが見られた。

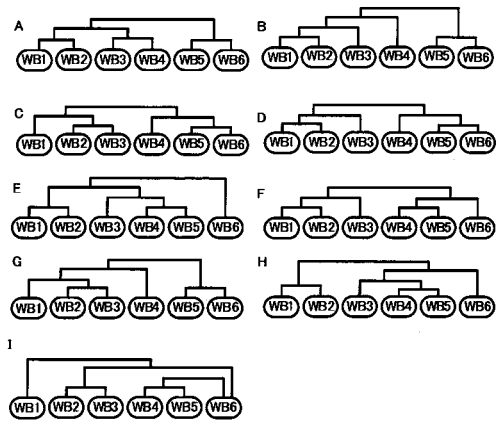


図3.4　6語基の係り受けパターン

(5) 7 語基構成熟語

同様に7語基構成熟語について調査した結果、図3.5に示すように、この係り受けに限定される形のみであった。品詞列パターンからも分かるように、7語基構成幹事熟語は、今回このパターンしか出現しなかった。

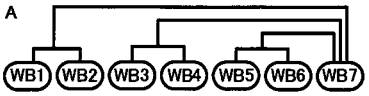


図3.5　7語基の係り受けパターン

今回の7文字コーパス中では、全て同様の分割がなされ、品詞列も同一なこともあり、係り受けパターンも図3.5のみに限定された。

4. 考察

6文字漢字熟語までと比べ、係り受けの解析が格段に複雑さを増し、また解析困難とされる係り受けも7文字からは割合が増えた。原因としては、6文字までと比べて、多い語基で構成される漢字熟語が増えたために、解析に、より複雑なパターンを想定する必要があった。実際7語基構成のような同じ語基への係り受けのパターンを除いても、3語基では2種類、4語基では5種類に比べ、5語基で15パターンあまりある。尚、1〜2文字で構成されている語基が多いため、7文字からは表4.1からも分かるように、4語基以上で構成されている事が多いので、係り受けのパターン数は多くなる。

表 4.1　語基構成パターン

1-1-5	0	0	3-1-2-1	8	4
1-5-1	0	0	3-2-1-1	2	2
5-1-1	0	0	1-2-2-2	501	59
1-2-4	0	0	2-1-2-2	1586	73
1-4-2	6	3	2-2-1-2	886	68
2-1-4	0	0	2-2-2-1	2330	69
2-4-1	1	1	1-1-1-2-2	28	16
4-1-2	4	2	1-1-2-1-2	40	27
4-2-1	8	2	1-1-2-2-1	115	48
1-3-3	0	0	1-2-1-1-2	47	32
3-1-3	0	0	1-2-1-2-1	235	71
3-3-1	1	1	1-2-2-1-1	16	11
2-2-3	46	6	2-1-1-1-2	48	32
2-3-2	30	7	2-1-1-2-1	399	80
3-2-2	32	5	2-1-2-1-1	77	27
1-1-1-4	0	0	2-2-1-1-1	21	13
1-1-4-1	0	0	1-1-1-1-3	0	0
1-4-1-1	1	1	1-1-1-3-1	0	0
4-1-1-1	0	0	1-1-3-1-1	0	0
1-1-2-3	0	0	1-3-1-1-1	0	0
1-1-3-2	1	1	3-1-1-1-1	0	0
1-2-1-3	1	1	1-1-1-1-1-2	0	0
1-2-3-1	8	5	1-1-1-1-2-1	5	4
1-3-1-2	3	3	1-1-1-2-1-1	3	3
1-3-2-1	3	3	1-1-2-1-1-1	1	1
2-1-1-3	8	6	1-2-1-1-1-1	2	2
2-1-3-1	5	2	2-1-1-1-1-1	2	2
2-3-1-1	3	3	1-1-1-1-1-1-1	3	1
3-1-1-2	1	1			

分割パターンごとの出現頻度とその分割による品詞列パターン数を表 4.1 に示す。この表から分割手法による特徴がわかる。ここで 4 語基では、2-2-2-1 として分割されたものが最も多く、続いて 2-1-2-2 が多いということがわかる。5 語基になると、4 語基と比べて出現頻度は少ないものの 2-1-1-2-1 に分割されるものが最も多く、続いて 1-2-1-2-1 に分割されるものが多いことがわかる。

4.2 品詞列パターン

前節の最後に述べた事象を明らかにするために、表 3.4 を単なる語基数から、語基構成パターン単位に粒度を下げた、これを表 4.2 に示す。

表 4.2 構成語基パターン毎の品詞列パターン数

	名詞	動詞	接頭	数詞	形容	形動	サ変
3語基							
1-4-2	1	0	1	0	1	0	0
2-2-3	3	1	0	0	0	1	2
2-4-1	1	0	0	0	0	0	0
3-2-2	5	0	0	0	0	0	0
3-3-1	1	0	0	0	0	0	0
4-1-2	2	0	0	0	0	0	0
4-2-1	2	0	0	0	0	0	0
4語基							
1-1-3-2	1	0	0	0	0	0	0
1-2-1-3	0	0	1	0	0	0	0
1-2-2-2	16	4	17	8	10	0	0
1-2-3-1	1	0	1	0	2	0	0
1-3-1-2	1	1	1	0	0	0	0
1-3-2-1	0	0	1	2	0	0	0
1-4-1-1	1	0	0	0	0	0	0
2-1-1-3	5	0	0	0	0	0	1
2-1-2-2	34	4	0	0	0	13	22
2-1-3-1	1	0	0	0	0	0	1
2-2-1-2	34	1	0	0	1	15	17
2-2-2-1	30	2	0	0	0	18	19
2-3-1-1	1	0	0	0	0	1	1
3-1-1-2	1	0	0	0	0	0	0
3-1-2-1	4	0	0	0	0	0	0
3-2-1-1	2	0	0	0	0	0	0
5語基							
1-1-1-2-2	4	0	4	8	0	0	0
1-1-2-1-2	9	0	6	8	4	0	0
1-1-2-2-1	12	0	16	16	2	0	1
1-2-1-1-2	7	0	17	3	4	0	0
1-2-1-2-1	18	4	29	7	10	0	0
1-2-2-1-1	3	0	5	1	2	0	0
2-1-1-1-2	24	0	0	0	0	0	8
2-1-1-2-1	51	0	0	0	0	5	24
2-1-2-1-1	19	0	0	0	0	0	8
2-2-1-1-1	9	0	0	0	0	1	3
6語基							
1-1-1-1-2-1	2	0	1	0	1	0	0
1-1-1-2-1-1	1	0	0	2	0	0	0
1-1-2-1-1-1	0	0	1	0	0	0	0
1-2-1-1-1-1	0	0	1	0	1	0	0
2-1-1-1-1-1	1	0	0	0	0	0	1
7語基							
1-1-1-1-1-1-1	0	0	0	1	0	0	0

4.3 係り受けパターン

表 4.3 は先頭語基の品詞によってどの係り受けパターンとなるかを示している。

3 語基においては、A と B の出現頻度は同等であるが、係り受けパターン B は A に比べると先頭品詞に多義性がない。3.2 で述べたが、3 語基に分割するためには 3 文字以上の分割が必要となるために、3 文字以上では割合の多い、名詞の出現が多くなり、それが品詞列にも現れている。

表 4.3 3 語基構成熟語の係り受け構造 毎の先頭語基品詞による品詞列数

	サ変	形動	形容	接頭	動詞	名詞	数詞
A	2	1	1	1	1	7	0
B	3	2	0	0	0	6	0

表 4.4 4 語基構成熟語の係り受け構造 毎の先頭語基品詞による品詞列数

	サ変	形動	形容	接頭	動詞	名詞	数詞
A	32	21	11	16	8	54	7
B	16	4	4	7	4	19	3
C	7	10	0	1	0	21	0
D	22	14	1	2	0	20	1
E	13	10	0	1	0	21	0
F	0	0	0	0	0	0	1

7 文字漢字熟語の中では最も係り受けによる先頭品詞が分散しており、先頭品詞から係り受け構造を予測するということが困難である。また、係り受けパターンが変わったことによる、先頭品詞の変化は A の接頭を除いて見られなかった。

表4.5 5語基構成熟語の係り受けパターン 毎の先頭語基品詞による品詞列数

	サ変	形動	形容	接頭	動詞	名詞	数詞
A	14	4	6	18	2	47	27
B	14	0	2	8	0	38	9
C	3	0	3	6	0	12	1
D	0	0	7	40	2	17	4
E	0	0	0	0	0	5	0
F	2	0	1	4	0	3	0
G	0	0	0	1	0	0	0
H	13	1	0	7	0	21	0
I	0	0	0	1	0	2	1
J	0	0	1	0	0	6	0
K	0	0	1	0	0	4	0
L	0	0	2	2	0	0	1
M	2	1	0	0	0	10	0

表 4.5 から分かるとおり、5 語基になると係り受けパターンも増える。一方で、係り受けパターンから先頭名詞に傾向が出ている。J のように形容詞を除けば名詞とほとんど決まるようなパターンもある。

ただし、6語基で構成される漢字は出現数が少ないため、係り受けに関しても、殆どその漢字熟語固有の係り受け構造を持っている。

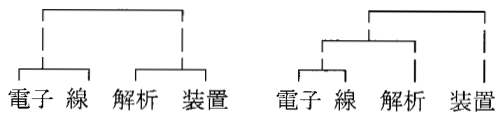
表4.6　6語基構成熟語の係り受けパターン毎の先頭語基品詞による品詞列数

	サ変	形動	形容	接頭	動詞	名詞	数詞
A	0	0	0	0	0	1	0
B	0	0	0	0	0	1	0
C	0	0	0	1	0	0	0
D	0	0	1	0	0	0	0
E	0	0	0	0	0	1	0
F	0	0	1	1	0	0	2
G	0	0	0	1	0	0	0
H	0	0	0	0	0	1	0
I	1	0	0	0	0	0	0

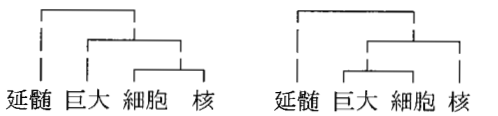
5. 終わりに

6文字漢字熟語と同様に係り受け構造の同定段階で、同定困難な漢字熟語が少数であるが存在した[8]。下記にその実例を一例として挙げる。

実例1(4語基)電子 線 回折 装置
品詞パターン：<名詞><名詞><サ変><名詞>



実例2(4語基) 延髄 巨大 細胞 核
品詞パターン：<名詞><名詞><名詞><名詞>



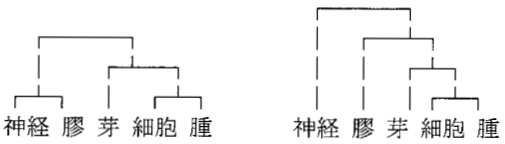
実例1では、「解析装置」、「電子線解析」いずれも、単独で専門用語として用いられる。また、実例2において、「細胞核」、「巨大細胞」のいずれも名詞として単独で用いられる。

以上が4語基の同定困難な例である。続いて5語基の実例を挙げる。

実例3においても、上述の実例と同様に、次の2つの用語が単独で用いられる。すなわち、「神経膠」、「膠芽細胞腫」である。

これらの実例に共通する現象は2つある。1つは、ある語基が2つの異なる単独で用いられる用語に共通して出現することである。他の1つは、特定の品詞パターンに見られる、ということである。すなわち、2つ以上の<名詞>または<サ変>が連続していることである。

実例3(5語基)神経 膠 芽 細胞 腫
品詞パターン：<名詞><名詞><名詞><名詞><名詞>



上記のような事例は6字漢字熟語では希であった[8]。一方、7字漢字熟語では、数十の熟語に見られた。

註・参考文献

[1] 野村雅昭. 三字漢字の構造. 秀英出版. 国立国語研究所報告. No.51. pp.37-62(1973).
[2] 野村雅昭. 四字漢字の構造. 秀英出版. 国立国語研究所報告. No.54. pp.36-80(1974).
[3] 小山照夫, 大江和彦. 医学専門用語の構造解析. 学術情報センター紀要. No.6, pp.115-124(1994).
[4]小山照夫, 大江和彦. 日本語医学専門用語の構造解析. 情報知識学会第2回研究報告会講演論文集. pp.17-20(1994).
[5] 竹内孔一, 内山清子, 吉岡真治, 影浦峡, 小山照夫. 語彙の制約を考慮した複合語解析モデルの構築. 情報処理学会. 情報学基礎研究会報告. No.57, pp.71-78(2000).
[6] 竹内孔一, 内山清子, 吉岡真治, 影浦峡, 小山照夫. 語彙概念構造を利用した複合名詞内の係り関係の解析. 情報処理学会論文誌. Vol.43, No.5, pp.1446-1456(2002).
[7] 郭恩東,森本貴之,後藤智範. 辞書見出し語の5文字漢字熟語を対象とした語基構成の解析.言語処理学会第13回年次大会. pp.348-351(2007).
[8] 梅木定博, 森本貴之, 後藤智範. 辞書見出し語の6文字漢字熟語を対象とした語基構成の解析. 情報処理学会. NL no.184(2008).