医療経過記録における漢字連続複合語の計量的分析

山崎 誠
国立国語研究所 言語資源研究系

相良 かおる
西南女学院大学 保健福祉学部

医療経過記録に現れる漢字連続複合語の語構成的な特徴を計量的に分析した．分析の観点は，語の長さ，品詞，語種およびシソーラスにおける意味分野である．同様の分析を『現代日本語書き言葉均衡コーパス』でも行い，両者を比較することで，医療経過記録の言語的特徴を明らかにした．

Quantitative Analysis of the *Kanji* Compound Words in the Medical Record

Makoto Yamazaki
Dept. of Corpus Studies
National Institute for Japanese Language and Linguistics

Kaoru Sagara
Dept. of Nutritional Sciences
Faculty of Health and Welfare
Seinan Jo Gakuin University

This study is intended as a quantitative investigation of structural analysis of *kanji* compound words in the medical records using word length, part of speech, word origin and semantic category in the thesaurus. We investigated "Balanced Corpus of Contemporary Written Japanese" in similar way and received the characteristics of the medical record.

１．背景

医療従事者により日々作成される医療経過記録は，個人情報が含まれるため，データの提供が難しく，なかなかその実態がつかめていない．

今回我々は，研究利用のために提供された倫理的配慮のなされた小児看護領域のプログレスノート（以後，「医療経過記録」と言う．）のデータ（スペースを含め 1,355,656 文字）について言語的調査を行った．

医療経過記録は，症状や処置などを簡潔に記録するため，短い文が多く，文を圧縮したような表現が頻繁に現れる．樺島（1979）によると，要約的な文章は名詞の比率が大きいという指摘がある．今回用いた医療経過記録のデータの品詞分布は表 1 のようになっている．樺島（1979）の調査では名詞の比率が高いテキストとして，新聞の見出し(名詞比率 74.0%)，新聞記事(同 68.3%)が挙げられているが，表 1 の名詞の比率は新聞見出しよりも高いことが分かる．なお，表 1 は，データを Unidic-mecab 2.1.2 により形態素解析した結果から，樺島の調査結果に合わせるため，当該の品詞のみを抜き出して集計したものである．

表 1　医療経過記録の品詞分布

品詞	語数	割合（%）
名詞	362,166	77.23
動詞	72,491	15.46
形容詞	11,534	2.46
形状詞	8,863	1.89
副詞	10,157	2.17
連体詞	2,238	0.48
接続詞	941	0.20
感動詞	531	0.11
計	468,921	100.00

本稿では，要約的な文章の特徴の一つである複合語を採り上げ，その構造を計量的に分析することにした．要約的な文章に出現する複合語の中には林（1982），石井（1993）らが扱っている臨時一語が多く含まれると予想されることから，その構造分析を通して，医療経過記録の特徴を明らかにしようとするものである．今回の調査では，視覚的にひとまとまりの印象がある漢字連続複合語を対象とした．

２．データの概要

使用したデータは，小児科の医療経過記録の自由記載部分を抽出したものであり，記載者および閲覧者は，医師，看護師，臨床工学士，栄養士，理学・言語・作業療法士，ＭＲＬ（診療情報管理士），ＭＳＷ（ソーシャルワーカー），臨床心理士，内視鏡事務，歯科衛生士・技工士，薬剤師，薬剤部事務，医事課事務員，医事課受付係である．

医療施設での匿名化処理として，数値は“9”に置換し，固有名詞および個人名は“X”に置換されている．

そして利用者が意図的に改行を行った個所および文中の". "の直後で分割したものを１行の文字列とし，同内容の文は削除されている．

従って，“患児が「お腹がすいた．何か食べたい．」と言った．”というデータは，“患児が「お腹がすいた．”，“何か食べたい．”，“と言った．”の 3 行に分割されるため，構文についての分析調査には適さない．その他に以下の制限事項がある．

①検査項目である"Co2"や"HbA1c"など数値を含む固有名詞は"Co9"，"HbA9c"となっている．

②アルファベットによる固有名詞は匿名化の対象外としている．

③カタカナの固有名詞や文字長が 1 文字の固有名詞は匿名化の置換対象から除外されている．

3．調査方法

語彙の抽出および頻度調査には，Perl 言語で実装された佐野（2003）の CLTOOL と，独自で作成した Perl プログラムを使用した．具体的な調査手順は以下の通りである．

（ア）正規表現を用い，Perl プログラムで連続する JIS 第 1 水準，第 2 水準の漢字の末尾にひらがなの“す”，“し”，“さ”が付加している文字列を抽出した．

（イ）目視により“食事少し”などの接辞が副詞であるものを削除しサ変接続可能な名詞・複合名詞を抽出した．語数は延べで 7,936 語，異なりで 1,101 語である．

（ウ）末尾の“す”，“し”，“さ”を削除．

（エ）取得された漢字連続文字列を「茶まめ」（MeCab ver.0.996+unidic-mecab ver.2.1.1）で解析し，短単位で 17,530 語（延べ），733 語（異なり）を得た．誤解析の分析は詳細には行ってないが，未知語となったのは異なりで 5 語，延べで 15 語であり，いずれも 1%未満である．なお，誤解析の人手修正は行っていない．

（オ）得られた短単位に『分類語彙表増補改訂版』の意味情報を付与した．

抽出された漢字文字列には，必ずしも複合語とみなせないものも交じっているが，本稿では便宜上複合語と呼ぶことにする．

4．語構成の分析 1：語彙量

得られた漢字文字列は最大 6 つの短単位に分割された．短単位数ごとの分布を表 2 に示した．3 短単位以内の複合語が異なりで 90.5%，延べで 97.9%を占め，ほとんどが比較的短い複合語であることが分かる．ちなみに最長の漢字複合語は 6 短単位で，「火/曜日/胸部/単純/写真/評価」と「喘息/性/気管/支/炎/指摘」の 2 つであった．

表 2　短単位数ごとの語数

短単位数	複合語数（異なり）	複合語数（延べ）
2	651	6,469
3	345	1,300
4	84	145
5	19	20
6	2	2
計	1,101	7,936

語構成上の位置における語（短単位）の分布を図 1 に示す．語構成上の位置は，末尾を基準に考え，語末に当たる語の位置を 0，その 1 つ前の語を 1，2 つ前の語を 2，というように数えている．語構成上の位置が 2 以上になると，複合語数が延べ，異なりともに減少していることが分かる．

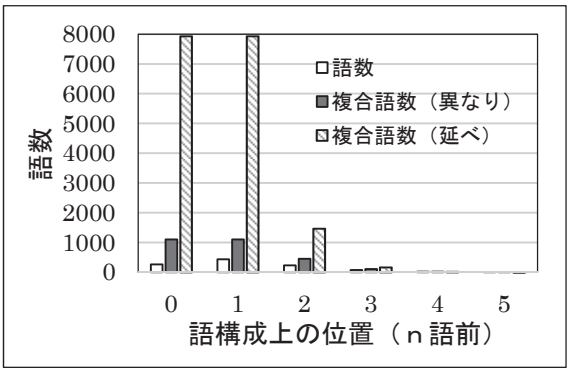

図 1　語構成上の位置と語数

語末での頻度が高く，かつ，語末以外に現れない語として，「終了(285)，安定(158)，施行(155)，改善(152)，確認(78)，導入(71)，低下(65)，評価(63)，注意(60)，悪化(60)，採取(55)，中止(51)」などがある（括弧内の数字は使用頻度（延べ数）である）．いずれも動作性の名詞である．逆に，語末には現れず，語末以外にだけ現れる語として，「経口(382)，骨髄(337)，血液(300)，本日(281)，炎症(264)，体重(247)，酸素(239)，療法(217)，状態(197)，明日(160)，再(157)，光線(145)，軽度(132)，尿(102)，当院(97)，剤(94)，抗生(82)，昨日(80)」などがある．こちらは，動作性の名詞がなく，モノの名前や時を表す名詞があることが特徴的である．

表 3 は，複合語数（異なり）の多い順に語を示したものである．上位は語末に位置する語が多いことが分かる．表中の「平均使用度数」は複合語数（延べ）を複合語数（異なり）で割った値であるが，この値が高いと，よく使われる複合語が多く，逆にこの値が低いと，よく使われる複合語が少ないことになる．表 3 からは，「開始」「受診」の平均使用度数が高いことが分かる．「投与開始（92），内服開始（74），当院受診（76），外来受診（56）」などの頻度の高い複合語が用いられているためである．一方，「注意」は平均使用度数が低い．これは，特定のことがらにではなく，さまざまなことへの注意が記述されていることを示唆する．

表 3　複合語数（異なり）の多い語

位置	語	複合語数（異なり）	複合語数（延べ）	平均使用度数
0	開始	48	549	11.44
0	継続	46	257	5.59
0	改善	37	152	4.11
0	注意	34	60	1.76
0	終了	31	285	9.19
0	安定	30	158	5.27
0	受診	27	335	12.41
0	施行	27	155	5.74
1	本日	24	239	9.96
0	確認	24	78	3.25
1	状態	22	197	8.95
1	症状	21	116	5.52
1	昨日	21	57	2.71

5．語構成の分析２：品詞と語種

表 4，表 5 は語構成上の位置と品詞，表 6，表 7 は，語構成上の位置と語種との関係である．もともと漢字連続であることから，名詞及び漢語が多いことは当然であるが，副詞や接頭辞，接尾辞，また，和語もある程度出現している．具体的な語と頻度は右の通りである．() 内の数字は，異なり/延べの順である．なお語構成上の位置 5 に出現する名詞は「火曜」と「喘息」であった．

表 4　語構成上の位置と品詞（異なり）

品詞／位置	0	1	2	3	4	5
名詞	1096	931	405	93	19	2
動詞	-	-	-	1	-	-
形状詞	-	4	2	-	-	-
副詞	-	38	2	1	-	-
接頭辞	-	20	17	4	1	-
接尾辞	3	104	23	6	1	-
記号	2	-	-	-	-	-
未知語	-	4	1	-	-	-

表 5　語構成上の位置と品詞（延べ）

品詞／位置	0	1	2	3	4	5
名詞	7894	7232	1375	155	20	2
動詞	-	-	-	1	-	-
形状詞	-	12	3	-	-	-
副詞	-	104	3	1	-	-
接頭辞	-	160	49	4	1	-
接尾辞	40	414	36	6	1	-
記号	2	-	-	-	-	-
未知語	-	14	1	-	-	-

表 6　語構成上の位置と語種（異なり）

語種/位置	0	1	2	3	4	5
和語	6	89	38	16	2	-
漢語	1093	1004	410	89	19	2
混種語	-	4	1	-	-	-
記号	2	-	-	-	-	-
（未知語）	4	4	1	-	-	-

表 7　語構成上の位置と語種（延べ）

語種/位置	0	1	2	3	4	5
和語	8	424	86	19	2	-
漢語	7926	7487	1379	148	20	2
混種語	-	11	1	-	-	-
記号	2	-	-	-	-	-
（未知語）	14	1	-	-	-	-

副詞：
依然(3/3)，一旦(8/40)，極[ごく](1/1)，極力(1/1)，十分(4/13)，順次(2/2)，随分(2/3)，大分(1/2)，大変(1/1)，直接(3/3)，適宜(11/35)，時折(2/2)，突然(2/2)

接頭辞：
片[かた](1/1)，肝(3/9)，血(2/21)，現(1/2)，再(17/157)，消(1/1)，大[だい](1/1)，短(1/1)，長(1/3)，低(2/2)，微(1/1)，鼻[び](2/2)，副(6/10)，毎(1/1)，両(2/2)

接尾辞：
影[えい](1/1)，園(1/2)，炎(1/1)，温(1/2)，化(6/42)，科(4/60)，華(1/5)，器(7/43)，限(1/1)，個(1/2)，後[ご](15/16)，剤(17/94)，師(1/1)，支(2/65)，肢(1/1)，児(2/7)，日[じつ](2/2)，所(1/1)，書(1/1)，状(7/11)，色[しょく](2/2)，性(5/5)，線(2/10)，素[そ](1/1)，隊(1/6)，値(2/2)，中[ちゅう](1/1)，調(1/1)，的(3/6)，等[とう](15/15)，内(1/2)，斑(1/1)，病(1/1)，品[ひん](1/1)，物[ぶつ](2/2)，胞(1/2)，目(3/3)，薬(20/77)，力(1/1)

和語：
朝(9/12)，朝方(2/2)，足首(2/2)，明日(24/160)，値(1/3)，痂[かさぶた](3/24)，風邪(1/1)，片(1/1)，肩(3/14)，皮(2/22)，昨日(35/80)，切(1/1)，薬(1/18)，口(1/6)，今朝(4/6)，声(1/1)，極[ごく](1/1)，仕事(1/3)，筋(2/9)，咳(1/1)，血 (2/9)，乳(2/4)，父(1/2)，出(1/1)，時折(2/2)，飲水(5/11)，母(12/29)，胎[はら](1/2)，針(1/1)，日(7/9)，左(2/2)，左耳(1/1)，部屋(1/7)，前(4/77)，枕(1/1)，右(2/2)，水(1/3)，目(3/3)，夕方(1/1)，指(2/2)，夜(3/3)

６．語構成の分析３：意味

図 2，図 3 は語構成上の位置と分類語彙表での意味分類（ただし，小数点以下 1 桁の「部門」で整理した）別に割合を調べたものである．1 つの語（短単位）に対して複数の分類番号が対応しているものについては，適切などれか 1 つに決めず，それぞれをカウントしている．図 2・図 3 から，以下のようなことが分かる．

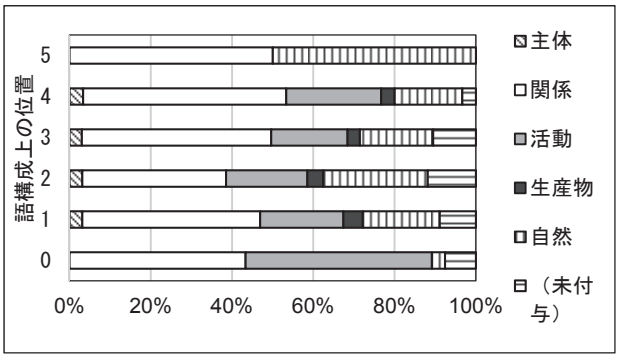

図 2　語構成の位置と意味（異なり）

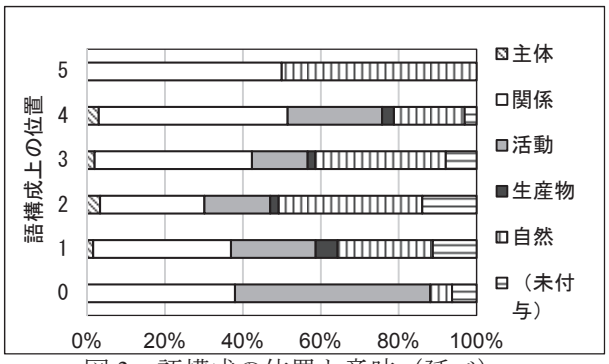

図 3　語構成の位置と意味（延べ）

(1)末尾の語（語構成上の位置が 0）には活動を表す語が相対的に多い．また，末尾の位置には，主体と生産物を表す語がないなど，意味的な種類が少ない．

(2)1 語前～4 語前における分布には大きな違いはない．

図 2，図 3 から異なりと延べの分布がほぼ似ていることが分かる．李（2014）によると，漢語サ変動詞の内部構成の安定性が指摘されている．すなわち，構成要素間の結合関係パタン及び品詞的結合パタンが，「タイプベースとトークンベースを比較しても，全体の傾向は大きく変化していない」のである．図 2，図 3 で示したのは語の内部構成ではなく，複合語の場合の構成であるが，やはり意味結合パタンが延べでも異なりでも似ており，安定しているということが言える．

７．コーパスとの比較

医療経過記録のデータが通常の書き言葉とどのくらい違っているのか，『現代日本語書き言葉均衡コーパス』(BCCWJ) と比較した．

BCCWJ は形態素解析が施されているため，その情報を利用して漢字連続を抽出した．具体的には以下の条件である．

（ア）perl の正規表現¥p{Han}で，原文文字列での一致により漢字表記語を特定する．

（イ）ただし，以下のものは除外した．

・品詞が「名詞-数詞」となっているもの．

・品詞が「名詞-助数詞」及び「名詞－助数詞可能」となっているもの．

・語種が「固」（固有名詞）となっているもの．

（ウ）以上の条件を満たす漢字連続に語彙素「為る」（サ変動詞）が付いているものを対象とする．

この時点で 81,367 語（延べ）29,910 語（異なり）が抽出できた．

（エ）さらに，医療経過記録で出現した，末尾に位置する語 264 語に BCCWJ で抽出した文字列の末尾が一致するものを抽出した．

厳密な意味で医療経過記録からの抽出と同じではないが，上の条件で抽出された漢字複合語は 17,132 語（延べ），7,558 語（異なり）である．医療経過記録では 7,936 語（延べ），1,101 語（異なり）であったから，延べ語数では BCCWJ は医療経過記録の約 2 倍強である．BCCWJ は医療経過記録のデータの 100 倍以上の規模があることを考慮すると，医療経過記録には漢字複合語が出現する割合が高いと言えよう．また，異なりが BCCWJ の約 7 分の 1 程度ということから，医療経過記録は同じ語が繰り返し使われていることが推測できる．

BCCWJ から抽出された複合語は最大 8 つの短単位に分割された．構成要素数と複合語数との関係を表 8 に示した．医療経過記録のデータと同様に語構成上の位置が 2 以上になると，複合語数が延べ，異なりともに減少していることが分かる．

表 8　複合語数（異なり）の多い語

位置	語数	複合語数（異なり）	複合語数（延べ）
0	230	7,558	33,920
1	2,919	7,558	33,920
2	799	1,677	3,517
3	204	282	330
4	40	41	42
5	10	11	11
6	3	3	3
7	2	2	2

表 9 は，複合語数（異なり）の多い順に語を示したものである．医療記録の場合（表 3）と同様に上位は語末に位置する語が多いことが分かる．平均使用度数は表 3 に較べると比較的ちらばりが少ないようである．表 3 と表 9 とで重なってい

る語は「確認」の 1 語だけであり，両者の語彙的な類似性が低いことを示唆する．

表 9　語構成上の位置と語数

位置	語	複合語数（異なり）	複合語数（延べ）	平均使用度数
0	化	2075	18658	8.99
0	使用	135	471	3.49
0	確認	132	536	4.06
0	利用	124	475	3.83
0	紹介	116	434	3.74
0	参加	110	249	2.26
0	実施	105	286	2.72
1	上	97	140	1.44
0	変更	95	272	2.86
1	以上	93	375	4.03
0	決定	89	391	4.39
0	増加	88	563	6.40
0	評価	88	395	4.49
0	検討	83	576	6.94
0	用意	81	196	2.42
1	分	81	844	10.42
1	程度	81	242	2.99

表 10 は，医療経過記録の語構成位置 0（末尾）の語が末尾の語全体（延べ語数）の中でどれくらいの割合を占めるかという使用率と BCCWJ で同様に末尾の語の使用率を比べて，その差が医療経過記録の方が多かったものを示した．表 10 では見やすくするために使用率の差の値を 100 倍している．

表 10　医療経過記録に偏る末尾語

語	医.使用率(延べ)	B.使用率(延べ)	使用率の差
開始	0.069	0.005	6.42
摂取	0.056	0.002	5.39
観察	0.053	0.002	5.01
検査	0.051	0.002	4.86
受診	0.042	0.002	4.07
抑制	0.038	0.001	3.70
投与	0.032	0.003	2.93
退院	0.030	0.001	2.89
継続	0.032	0.004	2.81
終了	0.036	0.008	2.78
反応	0.026	0.002	2.41
施行	0.020	0.001	1.81
加療	0.018	0.000	1.74
改善	0.019	0.003	1.64
安定	0.020	0.004	1.62
症状	0.014	0.000	1.45
哺乳	0.012	0.000	1.24
発熱	0.012	0.000	1.17
輸血	0.010	0.000	1.04
内服	0.010	0.000	0.97
採血	0.009	0.000	0.94
接種	0.009	0.001	0.87
発赤	0.008	0.000	0.82
喘息	0.008	0.000	0.81
処方	0.008	0.001	0.77
持続	0.009	0.001	0.74
採取	0.007	0.001	0.64
覚醒	0.007	0.000	0.64
導入	0.009	0.003	0.63
塗布	0.006	0.000	0.62
悪化	0.008	0.002	0.60
解熱	0.006	0.000	0.60
上昇	0.013	0.008	0.56
増悪	0.006	0.000	0.56
増加	0.024	0.018	0.52
中止	0.006	0.001	0.50
予防	0.005	0.000	0.50

表 10 には医療用語に混じって「開始，継続，終了，持続，導入，上昇，増加，中止」などのような一般的な意味の語も入っている．例えば，「開始」は，「化学療法開始(28)，吸入開始(21)，経口摂取開始(22)」など，治療あるいは患者の状況の変化を記録するために用いられている．これらは医療経過記録の特徴を表している語と見ることができる．

表 11　語構成上の位置と品詞（異なり）

品詞／位置	0	1	2	3	4	5	6	7
名詞	5480	6362	1346	235	39	10	3	1
代名詞	-	14	5	2	-	-	-	-
動詞	-	3	1	-	-	-	-	-
形容詞	-	1	3	-	-	-	-	1
形状詞	-	230	27	6	1	-	-	-
副詞	-	580	77	12	1	-	-	-
連体詞	-	1		-	-	-	-	-
接続詞	-	8	1	-	-	-	-	-
接頭辞	-	209	177	22	-	1	-	-
接尾辞	2077	111	30	5	-	-	-	-
助詞	-	30	6	-	-	-	-	-
記号	1	5	2	-	-	-	-	-
未知語	-	2	-	-	-	-	-	-
補助記号	-	2	3	-	-		-	-

表 12　語構成上の位置と品詞（延べ）

品詞／位置	0	1	2	3	4	5	6	7
名詞	15259	27283	2914	273	40	10	3	1
代名詞	-	14	5	2	-	-	-	-
動詞	-	4	1	-	-	-	-	-
形容詞	-	1	3	-	-	-	-	1
形状詞	-	2859	46	10	1	-	-	-
副詞	-	1336	109	12	1	-	-	-
連体詞	-	1	-	-	-	-	-	-
接続詞	-	9	1	-	-	-	-	-
接頭辞	-	1957	397	28	-	1		-
接尾辞	18660	248	31	5	-	-	-	-
助詞	-	38	6	-	-	-	-	-
記号	1	166	1	-	-	-	-	-
未知語	-	2		-	-	-	-	-
補助記号	-	2	3	-	-	-	-	-

表 11, 表 12 は品詞の語構成上の位置における品詞の分布である. 医療経過記録と比べると形状詞が延べ・異なりともに多いこと，また，末尾の語で接尾辞が多いことが特徴的である.

表 13, 表 14 は語構成上の位置における語種の頻度である. 語種の分布については医療経過記録と BCCWJ とはさほど変わらない.

表 13 語構成上の位置と語種（異なり）

語種/位置	0	1	2	3	4	5	6	7
和語	3	741	155	25	7	1	-	-
漢語	7554	6652	1492	256	34	10	3	2
外	-	1	-	-	-	-	-	-
混種語	-	123	19	1	-	-	-	-
記号	1	7	4	-	-	-	-	-
固	-	32	7	-	-	-	-	-
（未知語）	-	2	-	-	-	-	-	-

表 14　構成上の位置と語種（延べ）

語種/位置	0	1	2	3	4	5	6	7
和語	3	1562	218	26	7	1	-	-
漢語	33916	31842	3260	303	35	10	3	2
外	-	1	-	-	-	-	-	-
混種語	-	302	26	1		-	-	-
記号	1	168	4	-	-	-	-	-
固	-	43	9	-	-	-	-	-
（未知語）	2		-	-	-	-	-	-

図 4, 図 5 は語構成上の位置と意味の関係である. 意味的な分布は医療経過記録とは異なっている. 主体を表す語が多く, 自然を表す語が少ない.

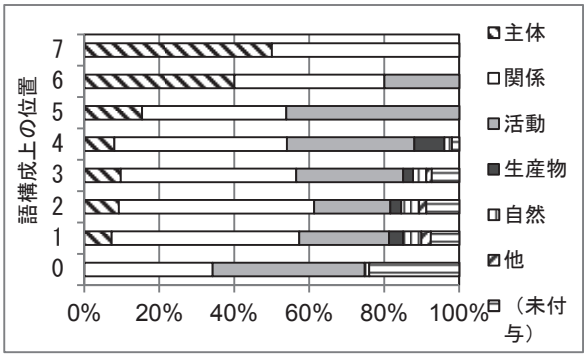

図 4　語構成の位置と意味（異なり）

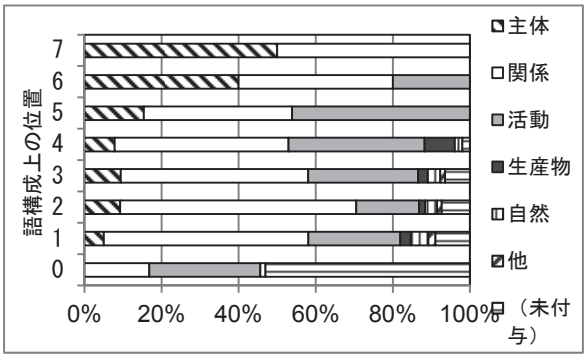

図 5　語構成の位置と意味（延べ）

8．まとめと今後の課題

医療経過記録はデータ的に制限が多く, 分析に限界があるが, 漢字連続複合語の語構成の観察により，その特徴の一旦を把握することができた. 今回はサ変動詞になるものを抽出したが, 今後はすべての漢字連続についての分析を行い, BCCWJ との比較も実施したい.

参考文献

1)　石井正彦：臨時一語と文章の凝縮，国語学, 173, pp.104-91（1993）.
2)　樺島忠夫：日本語のスタイルブック，大修館書店（1979）.
3)　国立国語研究所：分類語彙表増補改訂版，大日本図書（2004）.
4)　佐野洋：WindowsPC による日本語研究法, 共立出版（2003）.
5)　林四郎：臨時一語の構造，国語学，131, pp.15-26（1982）.
6)　李楓：現代日本語における汎用的漢語サ変動詞の抽出とその内部構成の検討, 第 4 回コーパス日本語学ワークショップ予稿集, pp.101-110（2013）.