医療経過記録における名詞連続の計量的特徴

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）‡

Metric Characteristics of Noun Sequences in Medical Progress Notes

Makoto Yamazaki (Dept. Corpus Studies, NINJAL)
Sagara Kaoru (Faculty of Health and Welfare, Seinan Jo Gakuin University)

要旨

医療経過記録は医療の現場において作成されるメモ的な性格の強い文章である。品詞的な特徴としては名詞の比率が高く、助詞・助動詞の省略が多い。このような文章によく見られる現象として機能語を用いず、名詞等を連続して用いる臨時一語的な用法がある。本発表では､小児科における医療経過記録約 90 万短単位から抽出した名詞連続の構造を品詞、語種、意味を中心に分析したものである。比較の対象として BCCWJ を用いて、医療経過記録の特徴を明らかにした。

１．はじめに

医療記録には、専門用語に加え、略語や隠語が、そして独特な表現が含まれる。紙媒体に記録される医療記録は､限られた場所で限られた医療従事者により記録され､閲覧され､保管されてきたが、近年の電子カルテシステムの普及により、施設内での情報の共有が可能となった。

しかし、医療用語の標準化がなされないまま、電子カルテシステムが導入されていることから、医療記録データには、表記のゆれや誤字を含む同義語、類義語が含まれている。これらの自然言語処理には、機械可読のコーパスや用語辞書が必要であるが、個人情報を含む医療記録は門外不出であり、言語学的調査は容易ではない。

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、短単位で 906,504）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。

山崎・相良（2014）では要約的な文章に出現する複合語の中に林（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、

† yamazaki@ninjal.ac.jp
‡ sagara@seinan-jo.ac.jp

医療経過記録の特徴を明らかにしようとした。同稿ではサ変動詞になる漢字連続に限って分析したが、本稿では、名詞連続の特徴を『現代日本語書き言葉均衡コーパス』（以下、BCCWJ と略す）との比較を通して観察する。

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

２．データ

２．１　医療経過記録

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。

また、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 行の文字列とした。同内容の文があった場合は、一方は削除されている。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。

２．２　BCCWJ

対照するデータとして BCCWJ（ver.1.0）を選んだ。BCCWJ 全体、および、医学系のサンプルがまとまって存在し、ジャンルとして抽出可能な LB（図書館書籍）、PB（出版書籍）、OC（Yahoo!知恵袋）、OY（Yahoo!ブログ）、PM（出版雑誌）から該当するサンプルを選んだ。

３．方法

３．１　形態素解析

医療経過記録のデータは MeCab ver.0.996+unidic-mecab ver.2.1.1 で解析し、品詞の大分類が「名詞」（品詞の中分類が「接尾辞-名詞的」を含む）の連続を抽出した。抽出された名詞連続数は延べで 63,916 個、異なりで 21,874 個である。

BCCWJ のデータは既に形態素解析が施されているので、それを利用した。

３．２　医学系サンプルの抽出

BCCWJ 全体のほかに、医学系のサンプルをまとまって抽出できるレジスターとして LB、PB、OC、OY、PM がある。それぞれ、以下の方法でサンプルを抽出した。

LB、PB：NDC が 49（医学）ではじまる 1,137 サンプル（内訳：LB、346 サンプル、PB、791 サンプル）。

OC：ジャンル 3 が「健康、病気、ダイエット」および「病気、症状、ヘルスケア」である 3,705 サンプル。

OY：ジャンル 2 が「病気、症状」である 694 サンプル。

PM：ジャンル 3 が「医学」である 31 サンプル。

４．結果

４．１　BCCWJ 全体との比較

医療経過記録と BCCWJ 全体の品詞、語種の割合を比較する。この場合の品詞、語種は名詞連続を構成している各短単位をすべて数えたものである。図 1 は品詞の構成比、図 2 は語種の構成比である。品詞は名詞を中分類まで細分して示した。名詞の中分類には、普通名詞、固有名詞、数詞の 3 つがあるが、医療経過記録（図で「MD」と示した）は BCCWJ

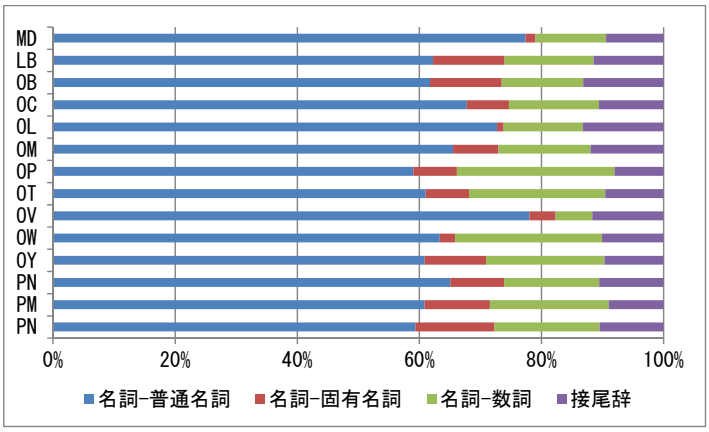

図 1　BCCWJ 全体との比較：品詞の構成比

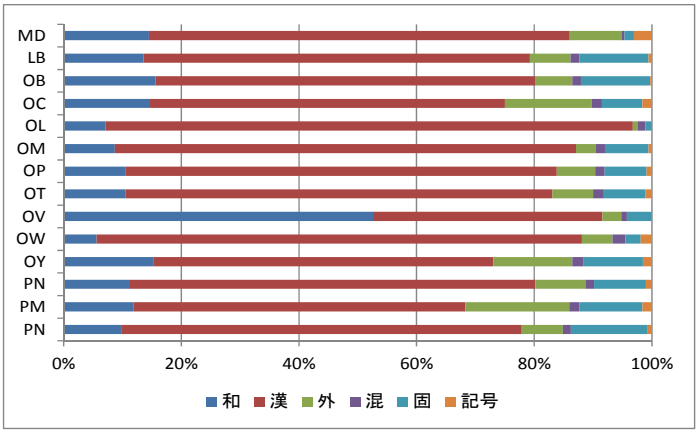

図 2　BCCWJ 全体との比較：語種の構成比

における各レジスターと比べて相対的に普通名詞の割合が高く、固有名詞が少ないことが分かる。固有名詞が少ないのはデータの匿名化のためと思われる。語種では品詞と同様に

固有名詞[1]が少ない。

４．２　医学系サンプルとの比較

図 3、図 4 は BCCWJ から医学系のサンプルを抜き出し、前節と同じ方法で比較したものである。品詞の割合では普通名詞の比率は BCCWJ 全体の場合と大きな差は見られない。

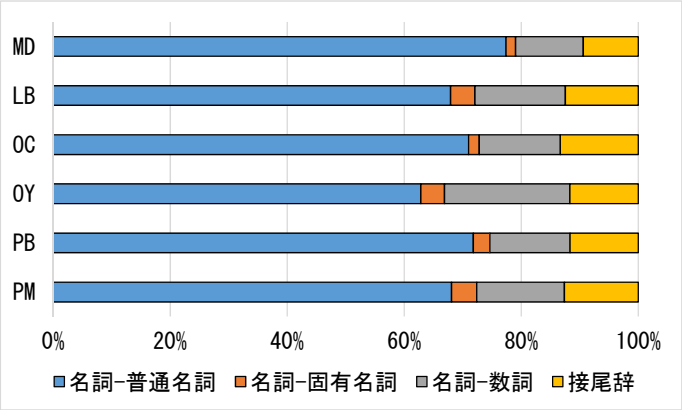

図 3　医学系サンプルとの比較：品詞の構成比

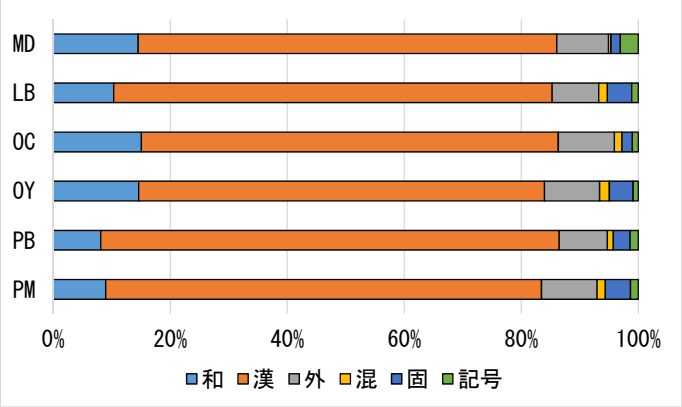

図 4　医学系サンプルとの比較：語種の構成比

次にカバー率の推移を比較する。図 5、図 6 は医療経過記録とデータの大きさが近い LB（図書館書籍の医学系サンプル）を比較したものである。LB の医学系サンプルの名詞連続数は延べで 62,028 個、異なりで 28,201 個であり、医療経過記録の延べ 63,916 個、異なり 21,874 個に近い。

図の横軸は異なり語数、縦軸は延べ語数であり、頻度の降順に並べた場合、上位何語[2]までで延べ語数の何%がカバーできるかを示している。比較のために、横軸、縦軸ともに 0%～100%を範囲とした標準化した数値で示した。

カバー率の推移は、医療経過記録（MD）の立ち上がりが早く、異なり語数の上位 20%で延べ語数の 70%、上位 40%で 80%に達する。図書館書籍（LB）の方はそれよりも 10 ポイントほど低い値となっている。

---

[1] 語種に固有名詞というカテゴリーがあるのは、UniDic の仕様による。

[2] 同順位があった場合、その順位を構成する n 語の異なりに対して、直前の順位を r とすると、r+1、r+2,…, r+n の順位の値を与えている。これはグラフのカーブを円滑化するための便宜的なものである。

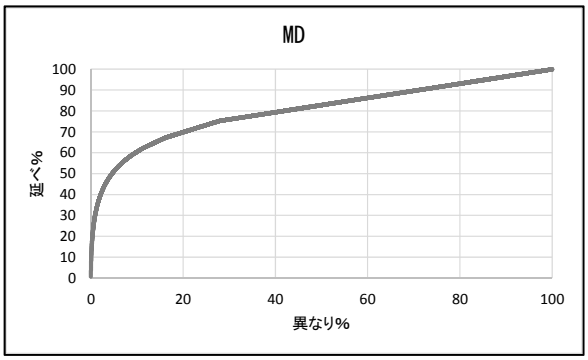

図 5 カバー率（MD)

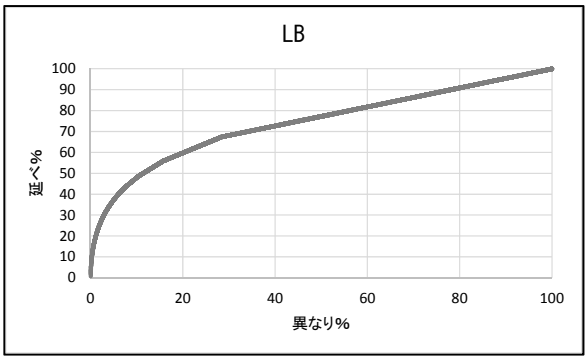

図 6　カバー率（LB)

次に使用頻度の多い語を個別に見てみよう.表 2 はそれぞれのレジスターの上位 20 語を示したものである。医療経過記録（MD）のデータは上記に数字が多く来ていることが特徴的である。これは体温などの検査の値や日付など、記録をとる上で重要な要素として記述されているものと思われる。個別の語としては、LB、OY、PB、PM に「患者さん」という

表 2　使用頻度上位 20 語[3]

	MD		LB		OC		OY		PB		PM	
順位	頻度	語彙素	頻度	語彙素	頻度	語彙素	頻度	語彙素	頻度	語彙素	頻度	語彙素
1	480	九. 九	512	一つ	170	一度	52	一つ	981	一つ	81	腹膜透析
2	471	九九日	431	患者さん	141	花粉症	38	日本ブログ村	822	患者さん	64	患者さん
3	455	九九	159	日本人	134	皮膚科	36	皆さん	375	高齢者	51	脂肪肝
4	398	九九九	158	人達	122	歯医者	28	花粉症	343	二つ	47	糖尿病
5	389	問題無し	155	糖尿病	119	皆さん	28	午前中	337	蛋白質	44	血液透析
6	369	九九九九グラム	142	母さん	112	整形外科	27	患者さん	314	図一	37	一つ
7	369	九. 九九	142	二つ	109	婦人科	26	一度	283	図二	26	合併症
8	328	九九. 九	139	遺伝子	96	血液検査	25	零零	266	看護師	21	コレステロール値
9	314	九回	137	癌細胞	96	耳鼻科	23	兄ちゃん	262	幾つ	21	大根番茶
10	314	九月	130	蛋白質	91	一週間	23	一日	262	図三	21	水キムチ
11	313	経過観察	126	幾つ	85	一日	23	何度	252	活性化	19	粗鬆症
12	306	圧痛無し	120	一度	75	生理痛	19	三十分	251	糖尿病	18	助中
13	300	九九九九	113	一日	75	一つ	18	子供達	221	人達	17	血液中
14	279	改善傾向	103	動脈硬化	73	健康診断	18	人達	215	遺伝子	16	高齢者
15	262	保育園	102	子供達	71	一箇月	16	一年	197	抗生物質	16	新生血管
16	240	九日	95	厚生省	67	一回	16	艶ちゃん	197	日本人	16	生活習慣病
17	214	九九時	91	治療法	66	歯医者さん	16	ヘルスブログ	195	図四	16	掌握握
18	210	九九. 九度	87	図一	64	医者さん	15	体重増加	191	表二	16	葡萄糖
19	207	全身状態	86	合併症	59	口内炎	15	皆様	190	図五		
20	196	九九度	86	神経細胞	57	産婦人科			188	十二		

[3] OY（Yahoo!ブログ）、PM（出版雑誌）は、同順位の語が複数あり、20 語を超えてしまうため内輪の範囲を挙げた。

語が現れているが、医療経過記録には見られないことである[4]。頻度 1 まで見ると「患者さん皆」が 1 例あるだけである。医療経過記録は患者の様態を記録するものなので、患者の存在が前提となっているため、「患者」という言葉を使う必要がないものと推察される。同様に「医者」「医師」の使用頻度も他のデータと比べると低い。

また、表 2 からは分からないが、症状を現す語として「～無し」が多用されていることも観察される。上位 100 位内に、「問題無し」(389)、「圧痛無し」(306)、「嘔吐無し」(129)、「異常無し」(120)、「下痢無し」(102)、「発熱無し」(99)、「変化無し」(75)、「著変無し」(69)、「咽頭発赤無し」(68)、「左右差無し」(68)、「発赤無し」(65) の 11 語が現れている[5]。逆に「～有り」は少なく、上位 100 位内には「必要有り」(64) の 1 語にとどまっている。

５．まとめと今後の課題

本稿では、医療経過記録の名詞連続を BCCWJ と比較しながら計量語彙論的な観点から概観した。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、名詞連続の示す特徴の一端を具体的に示すことができた。本稿はケーススタディ的な考察であり、今後他の観点も交えた考察が必要である。例えば、語構成的な観点での分析や構成要素間の意味的な関係についての分析が今後の課題である。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「コーパス日本語学の創成」による研究成果の一部である。データとして利用した BCCWJ は、国立国語研究所のプロジェクト及び文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（平成 18～22 年度、領域代表者：前川喜久雄）による補助を得て構築したものである。

参考文献

石井正彦(1993) 臨時一語と文章の凝縮，『国語学』173，pp.104-91.
樺島忠夫(1979)『日本語のスタイルブック』大修館書店.
林四郎(1982) 臨時一語の構造，『国語学』131，pp.15-26.
山崎誠・相良かおる(2014) 医療経過記録における漢字連続複合語の計量的分析，人文科学とコンピュータシンポジウム論文集，pp.221-226.

[4] OC（Yahoo!知恵袋）には順位 29 位（頻度 47）で「患者さん」が登場する。
[5] 括弧内は頻度数。