医療縮約表現の分析と課題

相良かおる[1]　黒田航[2]　東条佳奈[3]　西嶋佑太郎[4]　麻子軒[5]　山崎誠[6]
[1]奈良先端科学技術大学院大学　[2]杏林大学　[3]大阪大学　[4]京都大学　[5]関西大学　[6]国立国語研究所

概要

医療記録に含まれる合成語には，複合語に加えて，助詞が省略された句や節に相当するものがある．我々はこれらを「医療縮約表現」とし，現在 813 語の医療縮約表現の分析を行い，1,336 種類の語構成要素に 53 種類の意味ラベルを付与し，3 回目の見直しを行っている．分析の過程で，医療縮約表現には，①「文」相当となるもの，②接頭語「未」が語末に来るもの，③「施行」「あり」「術」などの語が省略されるものがあることがわかった．また，医療縮約表現の選定方法，語構成要素の認定，意味分類および意味ラベルについての課題も明らかになった．

1 背景と目的

医療記録データには，複数の語が連結された合成語が多く存在する．カルテの電子化が可能になった 2006 年，第 1 著者である相良は，電子カルテシステムの導入により施設内で蓄積された医療記録データの言語処理支援を目的に，医療記録を含む医療文書から実践医療用語を収集し，2008 年に形態素解析器 Mecab のユーザ辞書として利用可能な実践医療用語辞書 ComeJisyoV1[i]を公開した．以後，随時更新を続け 2021 年公開の ComeJisyoUtf8-3 の見出し語は 118,404 語となった．本辞書は，見出し語の語単位を定めず，医療従事者が医療文書から一語と認識した語を抽出し，登録しており，複数の語が連結された合成語が多く含まれる．これらの合成語には，複合語だけでなく，「気管支肺胞洗浄異常」[ii]のように，助詞が省略され，前後の文脈なしでは，「語」か「文」かの判断が困難な臨時一語[1][2][3]に相当するものも含まれる．なお，本研究では，これらの合成語を「医療縮約表現」と呼ぶ．

また「気管支肺胞洗浄異常」は，一般的な単語の意味から合成語全体の意味を推測することは難しい．これらを理解するためには，言語学的な研究が必要となるが，個人情報が含まれる医療文書は門外不出であり，こうした合成語の言語学的な調査はあまり行われてこなかった．そこで，(1)合成語の語構造を明らかにすること，(2)合成語を構成する語構成要素に意味ラベルを付与すること，(3)医療合成語の日本語学または言語学的知見を得ること，(4)得られた知見を含む成果物を医療実践，医療教育の領域で，出来得れば言語学研究の領域においても利用可能な形で公開することを目的とし，これらの合成語の解析に着手した．

具体的には，2018 年より日本語学研究者および医療従事者の協力を得て，ComeJisyoSjis-1 より選定した複合語 7,192 語の解析を行い，2021 年にその成果を『実践医療用語__語構成要素語彙試案表 Ver.2』として公開した[iii]．併せて 2021 年からは，ComeJisyoUtf8-3 の見出し語より医療縮約表現 5,690 語を選定し，語末の形態素が異なる 813 語の分析を行っている．

本発表では，医療縮約表現の解析結果と課題について述べる．

2 用語の定義

医療文書：医療記録を含め，医療に関する文書の総称．

医療記録：本研究では，2010 年～2015 年に小規模電子カルテシステム（EMR）により蓄積された医師においての診療録，看護師の看護経過記録データなどをいう．なお，厚生労働省より「書面に変えて電磁的記録[iv]により作成，縦覧，または交付を行うことができる」との通知が交付されたのは，2006

[i] ComeJisyo プロジェクト日本語トップページ – OSDN より公開：https://ja.osdn.net/projects/comedic/
但し，2023 年現在，接続困難な状態にある．

[ii] 「気管支肺胞洗浄」は検査法で．気管支から肺胞までを洗浄した排液に異常があることを意味する．

[iii] 言語資源協会より公開
https://www.gsk.or.jp/catalog/gsk2020-g

[iv] 電子化した小規模システム「電子カルテ（Electric Medical Record：EMR）」をいう。近年で

年 6 月である．よって，紙カルテから電子カルテに移行されて間もない時期に蓄積されたものである．

実践医療用語：医療記録で使用される用語をいう．その中には、教育・研究機関で使われる学術医療用語も含まれる。

複合語：本研究では，「手指」「損傷」などの単独で使う事のできる語（語基）が 2 つ以上からなる「手指損傷」などの語と，語基に「非」や「性」などの単独では使えない語（接辞）が連結した「非自己免疫性」などの合成語をいう．

医療縮約表現：医療記録で使われる合成語の内，「治療終了後」や「両下肢感覚低下」のような句および「主語＋述語（SV）」の構造を持つ節に相当する語，ならびに専門用語辞書に立項されていない「微量注入」などの合成語をいう．

語構成要素：医療縮約表現を構成する要素で，言語学や日本語学の形態認定の単位ではなく，医療の観点から有意味性を反映する単位に分割した文字列を語構成要素とする．例えば，「呼吸状態変調」の語構成要素は，「呼吸状態」と「変調」の 2 要素となる．なお，「医療の観点から有意味性を反映する単位」には，臨床医学では「扁平上皮癌」となり，解剖学では，「扁平上皮｜癌」のように専門分野により単位が異なることがわかっている[4].

医療の観点：医療分野を言語使用域に限定し，実践医療用語の利用者の立場・見地をいう．

意味ラベル：各語構成要素に付与した医療の観点からみた意味を表すラベルをいう．例えば，語構成要素「ポケット」には，「患部,状態,病態」と 3 種類の意味ラベルを付与している．

3 方法

3.1 調査対象

臨時一語の認定基準[5]を参考に ComeJisyoUtf8-3 の見出し語 118,404 語より，品詞がサ変名詞，接尾，形容動詞語幹，副詞可能である文字長 4 文字以上のものから，病名および英語訳のあるもの，国家試験問題文に出現する語を除いた 5,690 語について，MeCab 0.996 と Unidic-cwj-2.2.0 を用いて自動形態素解析し，語末の形態素で分類した後，同じ語末のグループから，乱数により一語を抽出し語末の異なる 813 語を分析対象とした．

3.2 研究方法

3.2.1 文書の特徴

文書の特徴を表す方法に文字種の割合がある．しかし，割合を出すために必要な個人情報を含む医療文書の入手は困難である．そこで，医療文書として，模擬診療録テキスト・データ[v]，第 117 回医師国家試験問題[vi]から B,E,F の 3 問題，教育用の看護経過記録の 3 種類のデータを用い，他分野の文書データとして BCCWJ[vii]のジャンル別データから図書(LB)，法律(OL)，会議録(OM)，教科書(OT)，書籍(PB)，新聞(PN)，加えて，令和 5 年司法試験短答式問題文（憲法，民法，刑法）[viii]を用いて，これら 10 種の文書データの文字種割合を求めた．

3.2.2 医療縮約表現の分析方法

最初に形態素解析された医療縮約表現について，「医療の観点からみた有意味性を反映する単位」にまとめあげたものを語構成要素とし，各語構成要素に意味ラベルを付与した．

なお，行為を表す語構成要素には，行為の主体者によって，治療行為，患者行為，支援者行為，非医療者行為の 4 種類のラベルを付与した．現在，語構成要素の単位について 2 回目の見直しと，見直した語構成要素に意味ラベルの付与を終え，3 回目の見直しを行っている．

3.2.3 医療縮約表現に後続する文字

ComeJisyo の見出し語から選定した医療縮約表現は，前後の文脈がないため「語」か「文」の判断ができない．そこで，医療縮約表現に後続する文字が句点か格助詞かにより便宜的に「文」か「語」を判別することにし，ComeJisyo の見出し語を抽出する

は、「電子保健医療記録システム（Electric Health Record System：EHR）」が出現している。

[v] 言語資源協会　GSK2012-D
https://www.gsk.or.jp/catalog/gsk2012-d/

[vi] 厚生労働省
https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp230502-01.html

[vii] BCCWJ_CharacterTable_byRegister.xlsx
https://clrd.ninjal.ac.jp/bccwj/bcc-chu.html

[viii]
https://www.moj.go.jp/jinji/shihoushiken/jinji08_00198.html

目的で取得した匿名加工済みでかつ検査値を削除した医療記録データ（2014 年に入力の約 150MB，以下「抽出用診療録データ」）と，医療縮約表現 813 語を照合し，出現した 296 語について後続する文字が句点「。」，語構成要素となり得る文字列（「語の一部」），文字列「あり」，「なし」，および格助詞「の／を／に／で／と／が／にて／へ／から／より」の語数を調べた．なお，「．」「有り」「無し」などの異表記の調査は行わない．

4 結果

4.1 文書の文字種割合

表 1 は，ジャンル別の文字種割合をまとめたものである．医療分野の文書では，アルファベット，数字の割合が高い．また，医療分野の文書においては，模擬診療録と医師国家試験問題文では，アルファベットの割合が高いのに対し，看護経過記録では，カタカナとひらがなの割合が高い．

表 1　ジャンル別文書の文字種割合

	記号	数字	ひらがな	カタカナ	漢字	英字	計
模擬診療	13.10%	6.30%	23.50%	6.60%	36.80%	13.70%	100%
医師国試	9.80%	7.90%	21.20%	4.30%	34.10%	22.60%	100%
看護経過	2.80%	7.40%	36.80%	8.10%	40.40%	4.40%	100%
司法短答	9.60%	3.10%	43.70%	3.50%	38.00%	2.10%	100%
法律(OL)	7.11%	0.39%	37.56%	0.40%	54.54%	0.00%	100%
図書(LB)	10.16%	0.53%	52.96%	6.50%	29.17%	0.67%	100%
会議録(OM)	5.79%	0.01%	60.42%	1.75%	31.91%	0.12%	100%
教科書(OT)	11.05%	3.10%	45.59%	6.56%	31.64%	2.06%	100%
書籍(PB)	10.08%	1.23%	48.66%	7.15%	31.36%	1.52%	100%
新聞(PN)	10.56%	2.38%	36.48%	8.44%	41.41%	0.74%	100%

4.2　医療縮約表現の語構成

医療縮約表現の平均値および中央値は 5 文字，最大値は 14 文字，最小値は 4 文字であった．見直し 2 回目を終えた時点の語構成要素数は，1,336 種類であり，意味ラベルは 53 種類である（表 2）．

表 3 は，医療縮約表現に含まれる語構成要素数をまとめたものである．語構成要素 2 要素からなる医療縮約表現が 611 種と最も多く，その内，文字数が 4 文字のものが 291 種ある．

表 2　医療縮約表現の概要

医療縮約表現	語構成要素	意味ラベル
813	1,336	53

表 3　語構成要素数分割数

要素数	1	2	3	4	5	計
縮約表現	130	611	60	10	2	813

表 4 は，語構成要素に付与した意味ラベル数をまとめたものである．最も多くの意味ラベルを付与した語構成要素「呼吸状態」には，[身体機能][生理][状態][指標]の 4 種類の意味ラベルを付与した．なお，「介助」には，[治療行為][支援者行為][非医療行為]の 3 種のラベルを付与し，「圧迫」には，[状態][治療行為][患者行為][非医療行為]の 4 種のラベルを付与しているが，表 4 では，行為ラベルを 1 ラベルとしてまとめ「介助」はラベル数 1 に，「圧迫」はラベル数 2 としている．

表 4　語構成要素に付与した意味ラベル数

ラベル数	1	2	3	4	計
要素数	1,048	258	29	1	1,336

4.3　後続する文字

表 5 は，抽出用診療録データに出現する医療縮約表現 296 語に後続する句点「。」と文字列「あり」「なし」，そして「あり」「なし」以外の意味を持つ文字列についてまとめたものである．句点が後続するものが 45%，「異常なし」「終了」「済み」「開始」「上」「後」などの意味を持つ文字列が後続するものも 46%あり，その中には接頭辞「未」があった．句点と意味を持つ文字列が共に後続するものは，67 語（23%）であった．

表 5　医療縮約表現に後続する文字

後続の文字列	縮約表現	割合（全 296 語）
句点「。」	133	45%
「あり」	41	14%
「なし」	23	8%
意味を持つ文字列	137	46%

表 6　医療縮約表現に後続する格助詞

格助詞	縮約表現	割合（全 296 語）
の	90	30%
を	90	30%
に	74	25%
で	57	19%
と	51	17%
が	50	17%
にて	23	8%
へ	17	6%
から	8	3%
より	6	2%

表 6 は，後続文字列が格助詞である医療縮約表現の異なり語数をまとめたものである．「右眼瞼下垂（の）」「視力回復（の）」など「の」が後続するものと，「を」が後続する「右眼瞼下垂（を）」

「食道透視（を）」などの医療縮約表現が 90 語（30%）あった．

「から」と「より」以外の全ての格助詞が後続するものに「入院継続」と「学校生活」があった．

5 考察と課題

医師が作成した模擬診療録と医師国家試験問題文にはアルファベットが多く，教育用看護経過記録ではカタカナが多い傾向（表 1）は，「フォローアップ」「follow-up」「F/U」「FU」などの表記の揺れとして，医療記録にもみられる．

語構成要素数が 2 種類で文字長が 4 文字の医療縮約表現 291 語において，「患者誤認」は「医療の観点」から 1 要素とも考えられる．これは，臨時一語の認定基準[5]に倣い，①専門用語辞書に立項されていない医療縮約表現について，②格助詞の挿入の可否をもとに語分割したためである．そこで，今後，語分割を見直す際，②の前に，公開されている症例報告での出現の有無なども確認する．

現在，行為を表す語構成要素には行為主体別の意味ラベルを付与している．例えば，「洗浄」には，患者行為，治療行為，支援者行為，非医療者行為の 4 種類のラベルを付与した．しかし，目的別に「衛生」「治療」「疾病予防」「健康維持」のような分類も考えられる．医療実践，医療教育に役立つ意味分類については今後の検討課題である．

後続文字に，「同意書取得未」「画像表示未」と，接頭辞「未」が見つかった．「同意書取得」では，「同意書（未）取得」の「未」が後置され，「画像表示」では，「画像表示（未施行）」の「施行」が省略されたと考えられる．

また，全 813 語中，語末の語が欠落した不完全な医療縮約表現と判断した 19 語の中に後続文字が句点「排便毎日。」と格助詞「肩甲帯離断が」があった．これらは，「排便毎日（あり）。」の「あり」が，「肩甲帯離断（術）が」の「術」が省略されたものであり，「施行」に加えて「あり」や「術」が省略されることが分かった．

林（1982）の臨時一語には，「参議院全国区制改革草案」「‥‥被災者救援対策本部」のような語末が一般名詞となる固有名詞化されたものが含まれるが，本研究では，固有名詞は対象外とし，一般名詞が語末にくるものは除外している．しかし，後続文字列が一般名詞「距離」で，かつ，固有名詞ではない「最大開口距離」があった．

「語末に接頭辞はこない」「語末が一般名詞である医療縮約表現は，固有名詞である」というのは，我々の思い込み，バイアスとも考えられる．

本研究の限界は，このバイアスの他にもある．

先ず(1)研究者の制限がある．共同研究者 6 名の内，医療従事者は 1 名であり，2 回の見直し後も未だ「医療の観点」からみた語分割，意味分類とは言い難い．そして，現実的な解決方法は見当たらない．また，医師と看護師などの職種ごと（表 1），ならびに専門分野ごとに[4]使われる語や意味に違いがあることが分かっており，語構成要素の認定基準や意味に唯一の正解はない．

次に(2)データへのアクセスの制限として，個人情報を含む医療文書の入手の困難さがある．

実践医療用語を分析する上で文脈は必要であり，例文の提示は重要である．実際，後続文字列より，医療縮約表現には①「文」相当のものがあること，②接頭語「未」が語末に来ること，③「施行」「あり」「術」など省略される語があることが分かった．しかし，抽出用診療録データの使用目的の制限から，これらの例文の提示はできない．

現在，入手可能な匿名加工済みの医療情報は「SS-MIX2 標準化ストレージ[ix]」のデータ項目であり，診療録・退院サマリ・看護経過記録などは，含まれない．今後入手可能となる仮名加工医療情報においてもこれらを容易に入手できるとは考え難い．

以上の限界に加えて，実践医療用語は変化している．新型コロナウイルス感染症により新たな医療用語が生まれ，「糖尿病」が「ダイアベティス」になる可能性がある．そして EMR と EHR の用語マスタに同じ用語が登録される訳ではない．

しかし「医療の観点」による語分割と意味分類の方法を提案し，意味ラベルの定義を言語化することは，医療実践や医療教育での利活用において重要である．また，自然言語を単なる記号とみなし，意味を捨象して学習する深層学習を用いた医師の思考の再現や支援の妥当性を判断する上でも，重要だと考えている．

これら限界や課題を踏まえ，解決に努めながら，今後も変化する実践医療用語に対応可能な「医療の観点」から妥当な語分割と意味分類を目指したい．

[ix] http://www.ss-mix.org/cons/ssmix2_about.html

謝辞

本研究は JSPS 科研費 JP21300099，JP18H03400，JP21H0377 の助成を受けたものです.

参考文献

[1]林四郎，臨時一語の構造，日本語学 131 集 p.15-25,1982.
[2]石井正彦，文章における「臨時一語化」と「脱臨時一語化」－脱臨時一語化の形式を中心に－，日本語研究，19p.1-15，東京都立大学，1999.
[3]石井正彦, 臨時一語と文章の凝縮, 国語学, 173 集 p.104-91，1993.
[4] 劉亜斌，里村洋一，佐々木哲明，木村通男，廣瀬康行，山﨑俊司，「構造化臨床医学用語集の構築に関する研究」，医療情報学 20 巻 6 号，p.513-522，2000.
[5] 石井正彦, 現代日本語の複合語形成論, ひつじ書房，2007.
[6] 相良かおる，黒田航，東条佳奈，西嶋佑太郎，麻子軒，山崎誠，「実践医療用語_語構成要素語彙試案表 Ver.3 の公開にむけて」言語資源ワークショップ p.44-53，2023.