様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月　 ５ 日現在

機関番号：１４６０３
研究種目：基盤研究(B)（一般）
研究期間：2021～2023
課題番号：２１Ｈ０３７７７

研究課題名（和文）医療記録文に含まれる合成語の語構成解析　－リアルワールドデータの利活用に向けてー

研究課題名（英文）Medical Terminology Analysis: Word Structure and Semantic Classification of Compound Words to Support Linguistic Processing of Real Data

研究代表者
相良　かおる（Sagara, Kaoru）

奈良先端科学技術大学院大学・先端科学技術研究科・客員准教授

研究者番号：００３３０８８７

交付決定額（研究期間全体）：（直接経費）　　9,400,000円

研究成果の概要（和文）：　医療記録に含まれる合成語810語を、用語集などに記載され慣用的に使われるもの、慣用的でないもの、機能語を使わずに連文や節相当の語と捉えられるものの3タイプに分類した。次に合成語810語を医療の観点から語分割し、得られた語構成要素1,264語に50種類の意味ラベルを付与した。また、合成語の語構造を記述する記法を定め、公開を目的にこれらをまとめた一覧表を作成した。
　加えて、蓄積された記録データを施設内で医療従事者自らが処理できるように、正規化などの前処理、語分割、合成語および語構成要素の抽出機能を持つ支援ツールを作成した。

研究成果の学術的意義や社会的意義
　医療記録に含まれる合成語には、辞書の見出し語にはない「食渣停滞」など、単語の一般的な意味から全体の意味の推測が困難なものがある。これらを医療の観点から語分割し、各要素に意味ラベルを付与したことで、言い換えが可能となり、機械学習や医療従事者向けの教育に活用できる。また、表記の揺れや同義語の統制が可能となることから、自然言語処理の精度の向上が期待できる。
　加えて本言語処理支援ツールは、医療従事者が施設内で個人情報を含む医療記録の語分割および用語抽出を可能にする。

研究成果の概要（英文）：The 810 compound words in the medical records were classified into three categories: those listed in the glossary and used idiomatically, those that are not idiomatic and those that are considered to correspond to a series of sentences or phrases without functional words. The 810 compound words were then segmented from a medical perspective and the resulting 1,264 word components were assigned 50 different semantic labels. Furthermore, a method for describing the word structure of compound words was defined. A lexical table consisting of these was then created.
Furthermore, a support tool with pre-processing functions, including normalization, word segmentation and the extraction of compound words and word components, was created to enable healthcare professionals to process the accumulated medical record data themselves in their facilities.

研究分野：医療言語処理

キーワード：医療用語　合成語　語構造　意味分類　電子カルテ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

我が国では、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、日々大量の医療記録データが蓄積されている。またコロナ禍により、オンライン診療が普及し、中小規模病院においても医療記録の電子化が進むと考えられた。

これらの医療記録データには、複合語に加え、略語や隠語、多様な表記・表現や、格助詞が省略された臨時一語に相当する合成語（本研究では、「医療縮約表現」という）が含まれる。しかし、個人情報が含まれ門外不出であることから、その詳細は明らかになっていない。従って、自然言語処理における語分割や語の正規化等の前処理に必要な語単位の認定規則、および表記の揺れや同義語を統一するための情報がない。

一方、非構造データであるこれら医療記録データの二次利用に関する研究は年々増加し、Google Scholar で言語の設定を“English”とし、検索キー「“EMRs” & “electronic medical records” & “Japan”」を検索（2020 年 10 月 10 日）したところ、2000 年～2010 年と 2010 年～2020 年の検索結果は、452 件→2,050 件（4.54 倍）となっていた。

そして大量の医療記録データを対象とした研究は、国民の医療情報を匿名加工して、大学や製薬企業の研究開発などでの活用を可能にする仕組みを定めた「次世代医療基盤法」が 2018 年 5 月に施行されたことにより益々増加すると考えられた。

このような状況の中、厚生労働省は 2010 年 3 月より「保健医療情報分野における標準規格（厚生労働省標準規格）」を順次定め、2020 年現在、一般社団法人 医療情報システム開発センター（MEDIS－DC）より開発された①病名マスター、②歯科病名マスター、③臨床検査マスター、④医療品 HOT コードマスター、⑤看護実践用語標準マスター、⑥画像検査マスター等が採択され、無償で公開された。しかし、これらは、各学会で作成された用語集や辞書、医学書から選定されており、ある程度語彙化され、医療縮約表現は含まれない。また、これらは医療従事者への応用を念頭にした医療情報学的な解析や言語学的な解析用に開発されたものではなく、自然言語処理の前処理に利用可能なシソーラスの整備も不十分である。

医療ビッグデータ（ＲＷＤ：Real World Data）の機械学習による自然言語処理では、①頻度情報で妥当な結果となる程のデータ量があること、②低頻度語に重要な語が含まれていないことに加えて、③結果に対する説明責任が必要だと考えられる。

医療施設内で蓄積される医療記録データの量が機械学習に十分な量であると仮定した場合においても、患者の症状を表す医療縮約表現などの多様な表現は、低頻度の語を増やし、「患者の症状」という重要な情報の獲得を困難にする。また、得られた結果を説明する上でも、機械学習の結果を確認する上でも、正解データ、すなわち人間可読のデータは必要である。

医療縮約表現を構成する語構成要素とこれらの意味、そして文法的な関係が明らかになれば、言い換えが可能になる。言い換えが可能になれば、表記の揺れや同義語を統制することができ、自然言語処理の精度の向上が期待できる。また、難解な医療縮約表現を分かりやすい言葉に言い換えることで、外国人医療従事者向けの教育に活用することができると考えた。

２．研究の目的

本研究の目的は以下の二つである。

(1) 医療記録に出現する「唾液流出良好」「頭部 CT/MRI」「白血球減少 grade1」「炎症反応高値」などの用語集や医学辞書等に立項されていない医療縮約表現を対象に、これらを構成する医療の観点から有意味な語構成要素を明らかにし、意味を表す意味ラベルを定め、「医療縮約表現語彙試案表（仮称）」を作成し、公開すること

(2) 個人情報を含み門外不出である医療記録データを、施設内で医療従事者がコンピュータ処理をするために以下の機能を持つ支援ツールを作成し公開すること
　① 操作が簡単であり、ネットワークに接続することなく、ボタンのクリックで起動する
　② 前処理の機能として、文字コードの変換、正規化、文字列検索と抽出、n-gram の抽出
　③ 形態素解析器 MeCab 用のユーザ辞書作成支援
　④ 医療縮約表現の解析で得られた知見を利用した合成語の抽出機能

３．研究の方法

医療縮約表現の解析は、以下の手順で行った。

(1) 医療縮約表現の定義、選定方法、語構成要素の分割規則の策定
　選定基準は、石井（2007）の臨時一語の認定基準を参考に品詞情報を用いて策定
　語構成要素の分割では、品詞などを考慮した機能的な単位を用いるのではなく、医療の観点から有意味性のある単位とし、同様に医療の観点に立つ意味分類を行う

(2) 医療縮約表現の選定：方法(1)で定めた抽出基準に則り、医療記録データから抽出した合成語を見出し語に含む実践医療用語辞書 ComeJisyoUtf8-3（登録語数 118,404 語）より、機械的に抽出後、更に語末の単語が異なる語をランダムに抽出

(3) 用語解析：医療縮約表現の語構成解析と意味解析

① まず、医療従事者により、医療縮約表現を、病名や手術名などの複合語、慣用的に使われる造語、臨時的な造語の 3 つのタイプに分類
② 次に、医療従事者、日本語学研究者を含む共同研究者の合議により、医療の観点から、医療縮約表現を語構成要素に分割
③ 得られた語構成要素に医療の観点による意味ラベルの付与

(4) 得られた研究成果からなる『医療縮約表現版 語構成要素試案表(仮称)』の作成
(5) 医療従事者のための言語処理支援ツールの作成

4．研究成果

表 1 は、ComeJisyoUtf8-3 の見出し語から文書頻度が 1 以上の 5,690 語を抽出した後、ランダムに語末の異なる 810 語を対象とし解析した結果である（方法(2)(3)）。

縮 約 表 現 ：腸蠕動音減弱
語構成要素列 ：腸蠕動音｜減弱
意味ラベル列 ：指標 ｜現象，所見，状態

「減弱」のように、複数の意味を持つ語構成要素には複数の意味ラベルを付与した。

これらの成果をまとめ『医療縮約表現版 語構成要素試案表（仮称)』を作成した（方法(4)）。

医療縮約表現には、「便｜尿｜失禁」のように語構成要素の並列および不連続な構造を持つものや、「内服管理｜能力」「内服｜管理能力」のように分割位置が複数あるものがあった。そこでこれらの記述を可能にする複層化形態素解析（Multi-Layered Morphological analysis：MLMA）を提案した。

表 1 タイプ

	語数
タイプ Ⅰ	233
タイプ Ⅱ	200
タイプ Ⅲ	387
計	810

表 2 解析結果

	語数
縮約表現	810
語構成要素	1,264
意味ラベル	50

【MLMA の記述例】

縮約表現	MLMA 記述	語構成要素	意味ラベル
便尿失禁	<便～><尿～><～失禁>	便 尿 失禁 便失禁 尿失禁	生体物質，排泄物 生体物質，排泄物 症状，状態，病態 症状，状態，病態 症状，状態，病態
内服管理能力	<[内服][管理><能力]>	内服管理 能力 内服 管理能力 管理	治療行為~患者行為~支援行為~非医療行為 指標 患者行為，医薬品 指標 治療行為~患者行為~支援行為~非医療行為

個人情報を含み門外不出の医療記録データを医療施設内で医療従事者がコンピュータで処理できるように、前処理機能と MLMA により記述された合成語（文字列）から語構成要素を抽出する機能を持つ言語処理支援ツールを作成した（方法(5)）。

なお、『医療縮約表現版 語構成要素試案表(仮称)』および言語処理支援ツールは、公開に向けて利用マニュアルの作成および最終確認を行っている。これらの準備ができ次第『医療縮約表現版 語構成要素試案表(仮称)』は、言語資源協会（GSK）より、言語処理支援ツールは、ComeJisyo Project の Web ページより、公開する。

今回の解析で医療縮約表現の中には、語末に接頭辞「未」のある「画像表示未」、「あり」「なし」が省略されたと思われる「骨密度低下」、語末に「術」の省略が考えられる「肩甲帯離脱」があった。省略の有無を確認するためには、これらを含む医療記録文を入手し、文脈からの意味および後続する文字列を調べる必要があり、今後の課題である。

図 1 言語処理支援ツールの画面

<参考文献>

① 石井正彦、現代日本語の複合語形成論、ひつじ書房、2007、p. 251-260
② ComeJisyoUtf8-3 https://comejisyo.com/
③ 特定非営利活動法人 言語資源協会(GSK) https://www.gsk.or.jp/

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 相良 かおる、高崎 智子、東条 佳奈、西嶋 佑太郎、山崎 誠	４．巻 1
２．論文標題 「急性」を含む病名の語構成	５．発行年 2023年
３．雑誌名 言語資源ワークショップ発表論文集 = Proceedings of Language Resources Workshop	６．最初と最後の頁 43～51
掲載論文のDOI（デジタルオブジェクト識別子） 10.15084/00003722	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 相良 かおる、黒田 航、東条 佳奈、西嶋 佑太郎、麻 子軒、山崎 誠	４．巻 1
２．論文標題 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	５．発行年 2023年
３．雑誌名 言語資源ワークショップ発表論文集	６．最初と最後の頁 309～318
掲載論文のDOI（デジタルオブジェクト識別子） 10.15084/0002000139	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 東条 佳奈、黒田 航、相良 かおる、高崎 智子、西嶋 佑太郎、麻 子軒、山崎 誠	４．巻 1
２．論文標題 実践医療用語 語構成要素語彙試案表 Ver.2.0 の構築	５．発行年 2023年
３．雑誌名 言語資源ワークショップ発表論文集 = Proceedings of Language Resources Workshop	６．最初と最後の頁 109～116
掲載論文のDOI（デジタルオブジェクト識別子） 10.15084/00003730	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 東条，佳奈，相良，かおる，西嶋，佑太郎，麻，子軒，山崎，誠	４．巻 1
２．論文標題 医療用語に含まれる序数詞について	５．発行年 2021年
３．雑誌名 じんもんこん2021論文集 2021	６．最初と最後の頁 194～199
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 相良 かおる	4．巻 22
2．論文標題 人間可読と機会可読の看護記録を目指して ： 看護と看語	5．発行年 2021年
3．雑誌名 日本医療情報学会看護学術大会論文集	6．最初と最後の頁 77-80
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計12件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 黒田航，相良かおる，東条佳奈，麻子軒，西嶋佑太郎，山崎誠
2．発表標題 要素の重複と不連続性を扱える抽出型の語構成要素解析:並列分散型形態素解析の提案
3．学会等名 言語処理学会第29回年次大会
4．発表年 2023年

1．発表者名 相良かおる，黒田航，東条佳奈，西嶋佑太郎，麻子軒，山崎誠
2．発表標題 医療縮約表現―医療記録に含まれる句や節に相当する合成語―
3．学会等名 言語処理学会第29回年次大会
4．発表年 2023年

1．発表者名 東条佳奈，黒田航，相良かおる，西嶋佑太郎，麻子軒，山崎誠
2．発表標題 医療記録における縮約表現の分析
3．学会等名 言語資源ワークショップ2022
4．発表年 2022年

１．発表者名
麻 子軒，黒田 航，相良 かおる，東条 佳奈，西嶋 佑太郎，山崎 誠

２．発表標題
実践医療用語における語構成要素の結合順序に関する量的調査

３．学会等名
計量国語学会 第66回大会

４．発表年
2022年

１．発表者名
山崎 誠，黒田 航，東条 佳奈，西嶋 佑太郎，麻 子軒，相良 かおる

２．発表標題
医療記録における縮約表現の量的構造―医療用語との比較―

３．学会等名
言語資源ワークショップ2022

４．発表年
2022年

１．発表者名
相良かおる，西嶋佑太郎，東条佳奈，高崎智子，山崎誠

２．発表標題
「急性」を含む病名の語構成

３．学会等名
言語資源ワークショップ2022

４．発表年
2022年

１．発表者名
相良かおる

２．発表標題
合成語の語構造と意味分類

３．学会等名
第23回日本医療情報学会学術大会

４．発表年
2022年

１．発表者名
東条佳奈,相良かおる，西嶋佑太朗，麻子軒，山崎誠

２．発表標題
医療用語に含まれる助数詞について

３．学会等名
人文科学とコンピュータシンポジウム2021

４．発表年
2021年

１．発表者名
相良かおる

２．発表標題
人間可読と機械可読の看護記録を目指して －看護と看語－

３．学会等名
日本医療情報学会看護学術大会論文集 : JAMI-NI,2021

４．発表年
2021年

１．発表者名
相良かおる

２．発表標題
実践医療用語の合成語生成ツールの作成と公開

３．学会等名
第25回日本医療情報学会春季学術大会 25th JAMI(Nov.2021)

４．発表年
2021年

１．発表者名
相良かおる

２．発表標題
ComeJisyoUtf8-3の誤解析調査 －看護師・助産師・管理栄養士国家試験問題文の語分割－

３．学会等名
言語処理学会　第28回年次大会

４．発表年
2022年

１．発表者名
黒田航,相良かおる

２．発表標題
医療用語の is-a オントロジー構築の FCA を使った効率化

３．学会等名
言語処理学会　第28回年次大会

４．発表年
2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
【成果物の公開について】
◆ 医療縮約表現版 語構成要素試案表（仮称）：言語資源協会(GSK)より公開予定
https://www.gsk.or.jp
◆ 言語処理支援ツール：ComeJsisyo Projectより公開予定
https://comejisyo.com/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	東条　佳奈 (Tojo Kana) (20782220)	大阪大学・大学院人文学研究科（人文学専攻、芸術学専攻、日本学専攻）・講師 (14401)	
研究分担者	山崎　誠 (Yamazaki Makoto) (30182489)	大学共同利用機関法人人間文化研究機構国立国語研究所・研究系・客員教授 (62618)	
研究分担者	黒田　航 (Kuroda Kow) (30425764)	杏林大学・医学部・准教授 (32610)	

6．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	麻　子軒 (MA Tzuhsuan) (30880249)	関西大学・国際教育センター・留学生別科特任常勤講師 (34416)	
研究分担者	高崎　智子 (Takasaki Satoko) (30882865)	西南女学院大学・保健福祉学部・教授 (37119)	

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	西嶋　佑太朗 (Nishijima Yutaro)		

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関