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　３ 年　 ６ 月 ２５ 日現在

機関番号：３７１１９
研究種目：基盤研究(B)（一般）
研究期間：2018～2020
課題番号：１８Ｈ０３４９９

研究課題名（和文）語形成および意味的情報を付加した実践医療用語辞書の構築

研究課題名（英文）Construction of a medical terms dictionary with information on word formations and meanings

研究代表者
相良　かおる（Sagara, Kaoru）

西南女学院大学・保健福祉学部・准教授

研究者番号：００３３０８８７

交付決定額（研究期間全体）：（直接経費）　　10,300,000円

研究成果の概要（和文）：医療記録に含まれる合成語7,192語の語構造の解析を行い、語構成要素6,380語を選定した。次に得られた語構成要素について意味分類を行い、80種類の意味ラベルを付与した。併せて各語構成要素が、合成語7,192語において、語頭、語中、語末に出現する頻度を調べた。そして、語構成要素、意味ラベル、合成語中の位置頻度と読み仮名を一覧にした『語構成要素語彙試案表』を作成し、言語資源協会（GSK）より公開した。
加えて、医療記録データを語分割し、分割した語の品詞情報を基に合成語を生成するツールGoMusubiを作成し公開した。

研究成果の学術的意義や社会的意義
個人情報を含み医療施設外に持ち出されることのない医療記録に含まれる合成語（7,192語）を構成する語構成要素（6,380語）に意味ラベルを付与した『語構成要素語彙試案表』の無償公開は、医療用語（合成語）の施設外での言語学的研究を可能とする。また、医療記録データから合成語を抽出するツールGoMusubiの無償公開は、各医療施設内での語分割と合成語の抽出を可能とし、そして『語構成要素語彙試案表』を用いることで抽出した合成語の意味の推測ができる。
加えて、語構成要素6,380語に付与したヨミガナや意味ラベルは、医療の専門用語教育にも利活用することができ、社会的にも意義のあるものとなった。

研究成果の概要（英文）：The word structure of 7,192 compound words in medical records was analyzed and 6,380 word components were extracted. The frequency of occurrence of each word component at the beginning, middle, and end of the 7,192 compound words was examined. We then created a “Word Component Lexical Trial Table” that lists the word components, semantic labels, frequency of occurrence in compound words, and pronunciation, and made it available on the GSK (literally “Language Resources Association”) website.
We also created a tool called “GoMusubi” that splits medical record data into words and generates compound words from the part-of-speech information assigned to the split words, and made this tool available on the GSK website.

研究分野：総合領域

キーワード：医療用語　語構成要素　意味分類　合成語

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

非構造データである医療記録データをコンピュータで解析する際、最初に行われる処理は、語分割と品詞の同定である。日本語は文の表記において単語が空白や句読点などで句切られておらず、分かち書き（語分割）に曖昧性が生じる。さらに日本文には複数の語が連なって意味を持つ複合語や臨時一語（以下、「合成語」という）が含まれ、どの部分迄を一つの用語とするのかは、利用目的によって異なる。

従って、医療記録データの解析においても、その目的によって、まとめ上げや分かち書きの単位は異なり、これらを指定するためには、合成語の語構造に関する情報が必要であるが、これらを得るための人間可読または機械可読の辞書はない。

申請者は方言や業界特有の略語や隠語を含む医療記録データの自然言語処理の支援が急務と考え、明確な語単位の基準は定めず、臨床経験を持つ看護職者数名で語の選定を行い、2008 年より形態素解析器 MeCab のユーザ辞書 ComeJisyo の無償公開を開始し、2013 年 11 月に ComeJisyoV5-1（登録語数 77,760 語）を公開している。ComeJisyoV5-1 の 2017 年 1 月 1 日から 10 月 15 日迄のダウンロード数は 499 件であり、これらの中には大学等での研究目的の他、外国人看護師を対象とした日本語教育の教材用としての利用もある。

語単位の基準が定まっていないことから、本辞書の更新には、医療従事者の知識が必要であり、登録語の拡充は容易ではない。

一方、本辞書には多くの合成語が登録されており、合成語の語構造解析の言語資源としての活用が可能である。

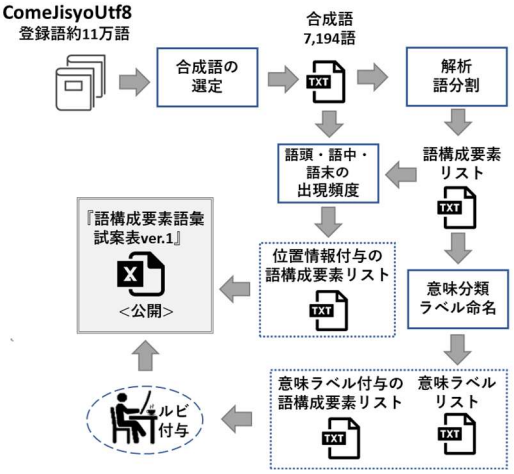
２．研究の目的

上記のように、医療記録に含まれる合成語の語構造は明らかになっていないものの、言語資源の合成語は、ComeJisyo の登録語に含まれている。

そこで、利用目的に沿った語単位の辞書や用語集の作成を支援すること、言語学的な研究の言語資源を提供することを目的とし、合成語の語構造解析を行う。

具体的には、本辞書から合成語を選定し語構造の解析を行い、語構成要素を選定した後、意味分類を行い、各語構成要素に意味ラベルを付与した『語構成要素語彙試案表』を作成し公開する。

また、個人情報を含む医療記録の自然言語処理を施設外で実施することは困難であることから、施設内で語分割し、合成語の抽出が出来るように支援ツールを作成し公開する。



３．研究の方法

ComeJisyo の登録語から合成語を選定し、(1)機械的に形態素（短単位）に分割した後、(2)分割した形態素を意味的な結合の強さでまとめ上げて語構成要素とし、各要素に意味ラベルを付与した。(3)次に臨床看護経験者、次いで臨床医経験者により医療的観点から語構成要素と付与された意味ラベルの見直しを行い、『語構成要素語彙試案表』を作成し公開する。

医療施設内で医療記録データの語分割ならびに合成語の抽出ができるように ComeJisyo を用いた語分割結果の品詞情報を基に合成語を生成するツールを作成する。

４．研究成果

[語構成要素語彙試案表]

医療記録に含まれる合成語 7,192 語を構成する語構成要素 6,380 要素に医療の観点からみた 80 種類の意味ラベルと、各語構成要素の語頭、語中、語末に出現する頻度（合成語 7,192 語中）、そしてヨミガナを付与し、言語資源協会（GSK）より『語構成要素語彙試案表 ver.1』を公開した。

言語資源協会（GSK）https://www.gsk.or.jp/

[UTF 版 ComeJisyo]

多くの医療施設での電子カルテシステムで扱う文字コードは Shift_JIS であることから、医療記録データの分かち書き用の辞書 ComeJisyo の文字コードを Shift_JIS として公開していた。

ところが近年、Unicode を前提とする研究および教育目的での利用者が増えたことから、2018 年 11 月に Unicode 版の ComeJisyoUtf8（登録語数 45,861 語）を作成公開した。その後、合成語の語構造解析で得た知見を反映し本辞書を更新し、公開した。

2018.11	ComeJisyoUtf8-1（登録語数：75,861 語）	新規公開
2019.4	ComeJisyoSjis-1（登録語数：111,664 語）	更新
2020.4	ComeJisyoSjis-2（登録語数：113,553 語）	更新
2020.5	ComeJisyoUtf8-2（登録語数：114,957 語）	更新
2020.7	ComeJisyoUtf8-2r1（登録語数：114,957 語）	英語の誤表記を修正
2021.3	ComeJisyoUtf8-3（登録語数：118,404 語）	更新

[合成語生成ツール GoMusubi]

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語分割し、合成語の抽出ができるように合成語生成ツール GoMusubi を作成し公開した。また、本ツールのプログラムの内、品詞情報から合成語を生成する Python プログラムのソースコード（Wcompounder）を公開した。

2021.3	GoMusbi_Ver.1.0　Pyinstaller により exe 化した実行プログラム　新規公開
2021.4	GoMusbi_Ver.2.0　仕様変更による更新
2021.4	Wcompounder_Ver.1.0 GoMusbi_Ver.2.0 の内、合成語を生成するプログラムのソースコード（Python3.8）

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　7件）

１．著者名 東条 佳奈 , 麻 子軒 , 相良 かおる , 高崎 智子 , 山崎 誠	４．巻 5
２．論文標題 病名における「-性」の分析 : 一般書籍との比較から	５．発行年 2020年
３．雑誌名 言語資源活用ワークショップ発表論文集	６．最初と最後の頁 357 - 364
掲載論文のDOI（デジタルオブジェクト識別子） 10.15084/00003175	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 相良 かおる , 高崎 智子 , 東条 佳奈 , 麻 子軒 , 山崎 誠	４．巻 5
２．論文標題 病名を表す合成語の語末調査	５．発行年 2020年
３．雑誌名 言語資源活用ワークショップ発表論文集	６．最初と最後の頁 151-156
掲載論文のDOI（デジタルオブジェクト識別子） 10.15084/00003154	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 山崎誠	４．巻 5
２．論文標題 実践医療用語を構成する語の計量的分析	５．発行年 2020年
３．雑誌名 言語資源活用ワークショップ発表論文集	６．最初と最後の頁 164 - 173
掲載論文のDOI（デジタルオブジェクト識別子） 10.15084/00003156	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 相良 かおる , 小野 正子 , 高崎 智子 , 東条 佳奈 , 麻 子軒 , 山崎 誠	４．巻 2020
２．論文標題 実践医療用語の語構成と意味　－語構成要素語彙表試案表の作成にむけて－	５．発行年 2020年
３．雑誌名 じんもんこん2020論文集	６．最初と最後の頁 289-296
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 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 麻 子軒，相良 かおる，高崎 智子，東条 佳奈，山崎 誠	４．巻 2020
２．論文標題 意味ラベルを用いた「-性」を含む病名の言い換え	５．発行年 2020年
３．雑誌名 じんもんこん2020論文集	６．最初と最後の頁 283 - 288
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 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 東条佳奈，相良かおる，小野正子，山崎誠	４．巻 11
２．論文標題 実践医療用語における語構成要素の意味分類試案：「先天性」を例に	５．発行年 2019年
３．雑誌名 現代日本語研究	６．最初と最後の頁 40-58
掲載論文のDOI（デジタルオブジェクト識別子） 10.18910/73339	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 山崎誠，相良かおる，小野正子，東条佳奈，麻子軒	４．巻 4
２．論文標題 実践医療用語の語構成要素への分割と意味ラベル付与の試み	５．発行年 2019年
３．雑誌名 言語資源活用ワークショップ発表論文集	６．最初と最後の頁 161-168
掲載論文のDOI（デジタルオブジェクト識別子） 10.15084/00002565	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計14件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 相良 かおる，高崎 智子，東条 佳奈，麻 子軒，山崎 誠
２．発表標題 実践医療用語の語構成解析
３．学会等名 第24回日本医療情報学会春季学術大会
４．発表年 2020年

１．発表者名
相良 かおる，高崎 智子，東条 佳奈，麻 子軒，山崎 誠

２．発表標題
病名を表す合成語の語末調査

３．学会等名
言語資源活用ワークショップ

４．発表年
2020年

１．発表者名
東条 佳奈，相良かおる，高崎 智子，麻 子軒，山崎 誠

２．発表標題
病名における「-性」の分析―一般書籍との比較から―

３．学会等名
言語資源活用ワークショップ

４．発表年
2020年

１．発表者名
山崎 誠

２．発表標題
実践医療用語を構成する語の計量的分析

３．学会等名
言語資源活用ワークショップ

４．発表年
2020年

１．発表者名
相良 かおる

２．発表標題
病名を表す合成語の語構成解析

３．学会等名
第40回医療情報学連合大会

４．発表年
2020年

１．発表者名
相良 かおる

２．発表標題
実践医療用語における語構成要素の意味ラベルについて

３．学会等名
言語処理学会 第27回年次大会

４．発表年
2021年

１．発表者名
相良 かおる

２．発表標題
実践医療用語の語構成と意味　－語構成要素語彙表試案表の作成にむけて－

３．学会等名
じんもんこん2020

４．発表年
2020年

１．発表者名
麻 子軒

２．発表標題
意味ラベルを用いた「-性」を含む病名の言い換え

３．学会等名
じんもんこん2020

４．発表年
2020年

１．発表者名
相良かおる，小野正子，山崎誠

２．発表標題
実践医療用語の語構造に関する考察

３．学会等名
第39回医療情報学連合大会

４．発表年
2019年

１．発表者名
相良かおる,山崎誠,麻子軒,東条佳奈,小野正子,内山清子

２．発表標題
実践医療用語の語構成要素－意味を基準とした分割－

３．学会等名
人文科学とコンピュータシンポジウム

４．発表年
2019年

１．発表者名
内山清子,岡照晃,東条佳奈,小野正子,山崎誠,相良かおる

２．発表標題
実践医療用語の語構成要素抽出の試み

３．学会等名
言語資源活用ワークショップ発表論文集

４．発表年
2018年

１．発表者名
東条佳奈,内山清子,岡照晃,小野正子,相良かおる

２．発表標題
実践医療用語に現れる語構成要素の辞書構築にむけて

３．学会等名
第62回計量国語学会

４．発表年
2018年

１．発表者名
相良かおる,小野正子,山崎誠

２．発表標題
UTF版実践医療用語辞書ComeJisyo1.0の作成

３．学会等名
第38回医療情報学連合大会

４．発表年
2018年

１．発表者名
相良かおる,小野正子

２．発表標題
実践医療用語辞書ComeJisyoSjis-1の作成

３．学会等名
言語処理学会第24回年次大会(NLP2018)

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

◆ComeJisyoおよびGoMusubiのダウンロードサイト
OSDN : https://ja.osdn.net/projects/comedic/

◆『語構成要素語彙試案表』のダウンロードサイト
言語資源協会（GSK） : https://www.gsk.or.jp

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	小野　正子 (Ono Masako) (50255957)	西南女学院大学・保健福祉学部・准教授 (37119)	
研究分担者	東条　佳奈 (Tojo Kana) (20782220)	大阪大学・文学研究科・助教 (14401)	
研究分担者	麻　子軒 (Ma Tzu-Hsuan) (30880249)	大阪大学・文学研究科・招へい研究員 (14401)	

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	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・言語変化研究領域・教授 (62618)	
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)	
研究分担者	内山　清子 (Uchiyama Kiyoko) (20458970)	湘南工科大学・工学部・教授 (32706)	
研究分担者	岡　照晃 (Oka Teruaki) (50782942)	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・コーパス開発センター・特任助教 (62618)	

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