様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年６月８日現在

機関番号：３７１１９
研究種目：基盤研究（B）
研究期間：2009～2011
課題番号：２１３０００９９
研究課題名（和文）　コメディカル実践用語辞書データベースの作成
研究課題名（英文）　Creating a dictionary database of medical terminology for medical professionals
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・講師
研究者番号：００３３０８８７

研究成果の概要（和文）：
標準化された用語がないまま、電子カルテシステムは普及し、電子医療記録文書が蓄積される中､我々は医療記録文書で使われる用語77,775語を収録した辞書ComeJisyoを作成･公開し、また、語種と字種の分布を明らかにした。
ComeJisyoは､電子医療記録文の単語分割の解析精度を90%以上に向上させ､複数の解析結果の比較（メタ分析）を可能とする。また、ComeJisyo に付加されるヨミガナは、音声への変換や仮名漢字変換等に活用できる。

研究成果の概要（英文）：
The increasingly widespread use of an electronic health record system without a standardized terminology has caused an accumulation of electronic medical records.
In order to perform natural language processing of these records, we created ComeJisyo, a dictionary comprised of 77,775 terms that is currently available to the public. After the latest version was complete, we investigated word class and character type in the electronic medical records provided by two separate hospitals.
Using ComeJisyo improves the accuracy of the text segmentation of electronic medical records to more than 90%. ComeJisyo also makes it possible to compare results of multiple analyses (i.e. meta-analysis) and it includes a pronunciation key that enables more practical kana-kanji and speech conversion.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	3,600,000	1,080,000	4,680,000
２０１０年度	1,700,000	510,000	2,210,000
２０１１年度	2,100,000	630,000	2,730,000
年度			
年度			
総　計	7,400,000	2,220,000	9,620,000

研究分野：総合領域
科研費の分科・細目：情報学・図書館情報学・人文社会情報学
キーワード：　医療用語、医療情報、電子カルテ、自然言語処理、形態素解析

１．研究開始当初の背景

短い時間で記録することが求められる医療記録には、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これらの医療記録は、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、保管される。

一方、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り、また日々蓄積される大量の医療情報は、従来の患者診療用途（一次利用）に加え、統計資料、臨床研究や疫学研究、教育訓練、そして診療情報管理等への２次利用も可能となってきた。

しかし、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、独自の用語が用いられ、医療機関間での電子的情報交換やデータベース作成が困難になっており、我が国全体として非効率な情報化となっている。蓄積されたこれらの医療情報を活用するためには、各医療機関独自の用語間の互換が必要である。

そして互換には、（１）医療情報に含まれる造語、略語、方言、専門用語を収録した用語集に加え、（２）用語間の関係についての情報が必要であるが、共に整備がなされていない状態である。

２．研究の目的

医療従事者により入力・利用されている医療情報の活用を支援する上で有益な自然言語処理用の機械可読実践用語辞書の作成と公開が本研究の目的である。

３．研究の方法

以下に研究の方法を示す。

（１）個人情報を削除し、倫理的配慮のなされた電子医療記録文書（以下、「記録文書」という）を収集する。

（２）収集した記録文書より実践用語を抽出し、分かち書き用の辞書を作成・公開する。

（３）記録文書より品詞、フリガナ、ヨミガナ等のタグを付加したタグ付きコーパスを作成する。

（４）収集した用語の語彙構成の分析および同義または類義によるグループ分けを行う。

４．研究成果

（１）ComeJisyo の作成・公開

品詞、フリガナ、ヨミガナ、利用される場面などの属性を付加した、分ち書き用辞書 ComeJisyo を作成し公開した。ComeJisyo は形態素解析器 Mecab 用のユーザ辞書の形式で作成し、コンパイル済みの Windows 版 Mecab 用ユーザ辞書と csv 形式の２種類を公開している。

新聞等に含まれる一般的な用語からなる辞書と ComeJisyo を併用することで、電子医療記録文の単語分割の解析精度は向上し、複数の解析結果の比較(メタ分析)が可能となる。

メタ分析が可能になることで、施設間での医療記録文で用いられる用語や書式の相違が明らかとなり、統制語の整備、あるいは用語の標準化の促進が期待できる。

また、ComeJisyo には、品詞以外にヨミガナが付加され、csv 形式でも公開しており、医療文書の音声変換や、医療文書を入力する際の仮名漢字変換、そして教育の場において利用することができる。

公開履歴の概要を以下に示す。

① ComeJisyoV1：登録語数 30,146 語

登録語は、看護学教科書の索引語（40,833 語)、2002 年～2007 年の看護師国家試験問題に含まれる用語 (9,478 語)、看護領域文書より抽出した用語 (50,805 語)、Web 上で公開されている用語辞書（48,875 語）の４種の言語資源より共通に出現する用語を選定している。

② ComeJisyoV2：登録語数 34,142 語

臨床管理栄養士３名により選定した栄養管理・栄養指導分野で使われる用語 3,996 語を追加し公開している。

③ ComeJisyoV3：登録語数 41,592 語

医療施設２施設より提供された倫理的配慮のなされた記録文書および看護領域の文書から、臨床看護経験者３名が抽出・選定した用語 7,450 語を追加している。

④ ComeJisyoV4：登録語数 77,647 語

新たに１施設を加えた３施設の記録文書より、臨床看護経験者３名が抽出・選定した 36,055 語を追加し公開している。

⑤ ComeJisyoV5：登録語数 77,775 語

新聞雑誌などに含まれる一般的な用語の中に、「根性（コンセイ)」、「呼名（コメイ)」、「嫌気（ケンキ)」等、医療分野特有の音読みをする語があることが判明したため、これらの語 128 語を追加している。

尚、2012 年６月７日現在での ComeJisyo のダウンロード数は、623 件となっている。

（２）タグ付きコーパスの作成

医療記録文に、品詞、フリガナ、ヨミガナを付加したタグ付きコーパス 869,712 行を作成した。

（３）語彙調査

記録文書に含まれる用語について、語彙調査を行い、以下のことが明らかになった。

① 施設間での相違

先ず初めに２施設の記録文書各 3,000 行に含まれる語の語種構成と品詞構成を調べた。

その結果、A 病院に比べ、B 病院において

は、増加、上昇、減少、降下、を記号“↑”、“↓”で記述する傾向にあり、記号の割合が多く、品詞の構成にも相違がみられること、B 病院の記録文書の方が英文字の割合が若干高くなっていることが明らかとなり、施設間によって使われる用語の語種構成および品詞に相違があることが示唆された（図１、図２）。

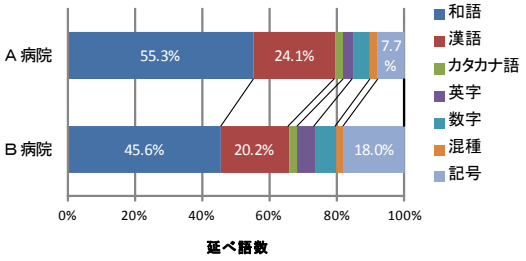

図1. 語種構成の比較

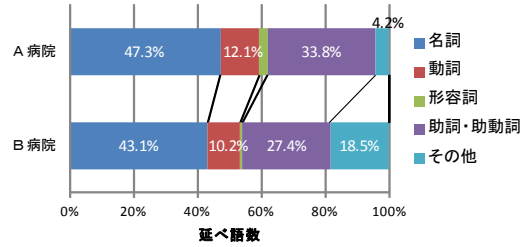

図2. 品詞構成の比較

② ComeJisyo の見出し語

複合語の多い ComeJisyoV3 の見出し語を、一般テキスト向け形態素解析辞書 UniDic を用いて「短単位」に分解し、語種割合を調べた（図３）。

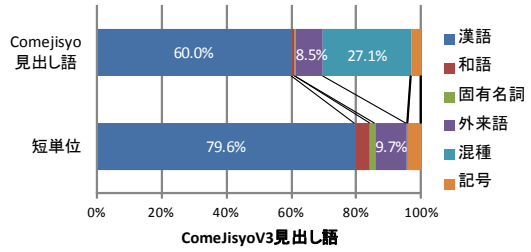

図3. 語種の割合

複合語を多く含む ComeJisyoV3 の見出し語の語種をみると、必然的に混種語が多くなっている。しかしながら、短単位に分割すると、一般の辞書の見出し語、和語 33.2%、漢語 49.4%、外来語 9.0%、混種語 8.4%と比較して、漢語が約 80%と圧倒的に多く、和語が 5%と少ないことが明らかとなった。

外来語の高頻度語を見てみると、オランダ語・ドイツ語からの借用語が多く用いられている。以下に出現頻度 50 以上の外来語を示す。なお、原語認定は一般的な国語辞典の記載に従っている。

リンパ（独）、ウィルス（独）、ケア（英）、ヘルニア（ラテン語）、アルコール（蘭）、エネルギー（独）、ビタミン（独）、カテーテル（蘭）、ショック（英）、アレルギー（独）、チューブ(英)、ポリープ(独)、リューマチ(蘭)、ガス（蘭）、ドレーン（英）、ヘルペス（独）、グロブリン（英）、ホルモン（独）

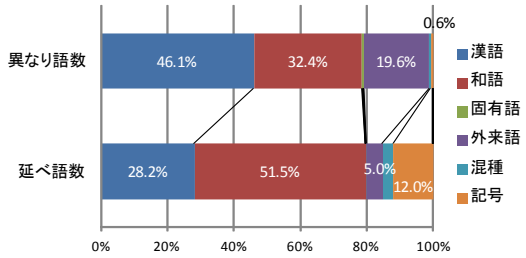

図4. 看護経過記録文中の語種の割合

低頻度のものには化学物質や薬品名が多く、これらは国際的な命名法に従ったものである。この他に固有名詞と普通名詞の区別が難しいものなど、原語を決定できない外来語見出し語も少なくない。

③ 看護経過記録文に含まれる用語の特徴

看護経過記録には、患者本人から得た主観的な情報、医療者が観察した客観的情報、これらから導きだされた結果や判断、そして判断に応じた看護の計画が時系列に記録されている。従って、看護経過記録には、患者本人との会話の内容や日常の様子に加え、医学的な専門用語や表現、そして検査・測定結果等が自然言語で記載されている。

新聞記事など一般的な日本文との相違をみるために、看護経過記録文 10,000 行（418,876 文字）について語彙調査を行った。

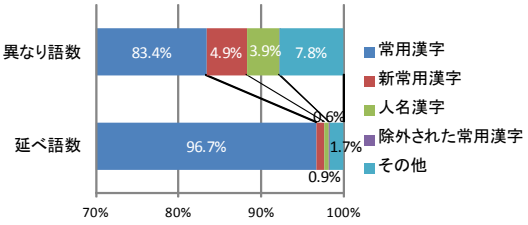

図5. 漢字制限による分類

新聞 70 誌を対象に 1994 年に実施された語彙調査での異なり語数の語種比率、漢語 35.5%、和語 27.7%、外来語 30.7%、混種語 6.1%と比べて、看護経過記録文の漢語の割合は 46.1%と非常に高く、和語も 32.4%と高くなっている、一方外来語は 19.6%と低くなっていることがわかる。

また、漢字制限による分類では、常用漢字の占める割合が 83.4%、新常用漢字が 4.9%と高いものの、一般に使われることの少ない漢字が 11.7%含まれていることが分かる。

例えば、出現頻度の高い人名漢字には、「梢(124)」「臥(114)」「腔(96)」「倦(69)」「坐(64)」「膏(54)」「粥(46)」「穿(37)」「娩(27)」「窄(25)」が、その他の漢字には、「疼(338)」「嘔(314)」「痰(194)」「嗽(111)」「咳(100)」「癌(73)」「痺(69)」「喀(59)」「痙(51)」「攣(51)」「瘡(12)」などがある。

④ 看護経過記録文の表現の特徴

電子カルテシステムでは、短い時間に、小さな入力画面を見ながら、入力されることから、助詞の省略や、記号（上昇：↑、変化：→、変化なし：±）を用いる等の文の短縮化が見られ、誤入力や誤変換も散見される。また、個人が自由に略語を作り、使用することから、「睡眠良好」、「睡眠＋」等、多様な表現が存在する。

カタカナ語の表現では、「クリニカルパス」と「クリティカルパス」のような意図的な表記のゆれに加えて「インスリン」、「インシュリン」等の表記の揺れもみられる。

助詞の省略は、「本日夕方抗生剤後抜針予定」のような漢字の文字列を生み出し、機械的な単語分割を困難にする要因となっている。

また、看護経過記録文の中には、特殊な読み方をする言葉が含まれている。例えば、「鼻汁」は、「はなじる」ではなく、「びじゅう」と音読される。従って、「咳嗽鼻汁なし」は、「がいそうびじゅうなし」と音読される。

意味的な特徴としては、「介入（intervention：何らかの影響を及ぼしたり、あるいは病的過程を変えることを意図する行為や援助）」や「アウトカム（outcome：病気の経過の帰着するところ。これには治癒、軽快、不変、死亡等がある）」等、一般的な意味ではなく、医療従事者特有の意味付けがなされ、使われている用語も少なくない。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）

① 相良かおる、小野正子、鈴木隆弘、嶋田元、小作浩美、看護記録文の計量的用語調査、査読有、情報処理学会シンポジウム論文集、2010、(15) pp.103-110.

②相良かおる、浅原正幸、小野正子、外山健二、形態素解析エンジンMeCab用辞書ComeJisyoV.2 および看護教育支援用仮名漢字変換辞書の作成と公開、第 29 回医療情報学連合大会論文集、査読有、2009、p.983 -984.

〔学会発表〕（計６件）

①相良かおる、電子医療記録の分ち書き用ユーザ辞書ComeJisyoの紹介と単語生起コスト、言語処理学会第 18 回年次大会、2012 年 3 月 15 日、広島市立大学

②小木曽智信、医療分野で使われる複合語の語種構成、第 29 回社会言語科学会研究大会、2012 年 3 月 11 日、桜美林大学

③鈴木隆弘、相良かおる、ComeJisyoの退院時サマリへの適用、第 38 回Mテクノロジー学会大会、2011 年 8 月 6 日、南知多町役場

④相良かおる、分かち書き用辞書ComeJisyoの評価、第 15 回日本医療情報学会春季学術大会、2011 年 6 月 17 日、幕張メッセ国際会議場

⑤相良かおる、小野正子、鈴木隆弘、嶋田元、小作浩美、看護記録文の計量的用語調査、情報処理学会　人文科学とコンピュータ研究会、2010 年 12 月 11 日-12 日、東京工業大学大岡山キャンパス

⑥相良かおる、形態素解析エンジンMecab用辞書ComeJisyoV.2 および看護教育支援用仮名漢字変換辞書の作成と公開、第 29 回医療情報学連合大会、2009 年 11 月 25 日、広島国際会議場

〔その他〕

ComeJisyo の公開ダウンロードサイト
http://sourceforge.jp/projects/comedic/

６．研究組織

(1)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・講師
研究者番号：00330887

(2)研究分担者
（H21 年度～H23 年度）
小野　正子（ONO　MASAKO）
西南女学院大学・保健福祉学部・准教授
研究者番号：50255957
鈴木　隆弘（SUZUKI　TAKAHIRO）
千葉大学・医学部附属病院・准教授
研究者番号：40323422

（H23 年度）
小木曽　智信（OGISO　TOSHINOBU）
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・言語資源研究系・准教授
研究者番号：20337489

（H22 年度）
高崎　光浩（TAKASAKI　MITSUHIRO）
佐賀大学・医学部・准教授
研究者番号：70236206

（H21 年度）
浅原　正幸（ASAHARA　MASAYUKI）
奈良先端科学技術大学院大学・情報科学研究科・助教
研究者番号：80379528
外山　健二（TOYAMA　KENJI）
西南女学院大学・保健福祉学部・教授
研究者番号：60249620