実践医療用語辞書 ComeJisyoSjis-1 の作成

相良かおる、小野正子
西南女学院大学
sagara@seinan-jo.ac.jp

1. はじめに

およそ 20 年前の 2001 年、厚生労働省は、全国 400 床以上の 6 割に電子カルテシステムを導入するという具体的な目標を掲げた。

医療記録が電子化されることで機械処理が可能となり、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などの 2 次利用が要望されると考えた筆者は、電子化された医療記録データの自然言語処理を支援することを目的に、2004 年より看護実践用語の収集を開始し、用語の分析 [1)2)3)]と看護用語の標準化に関する調査研究 [4)5)6)]を行った。加えて看護支援システムの稼働状況を調査するために電子カルテシステムが稼働または一部稼働している施設 50 施設を訪問し、電子カルテシステムに詳しい看護師等に半構成的面接による調査を行い、入力環境を視察した [7)]。

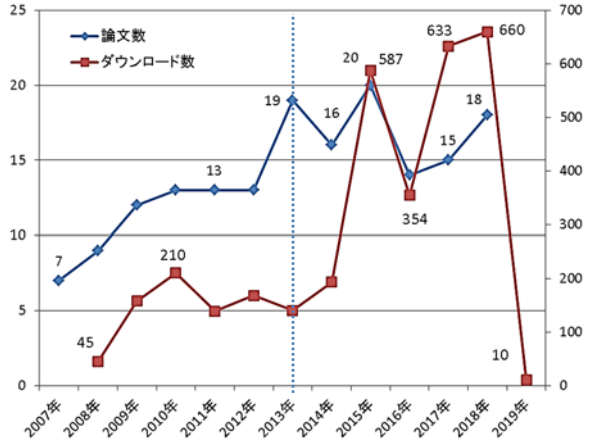

図 1　論文数とダウンロード数(2019 年 1 月 14 日検索)

これらの調査結果を踏まえて、2008 年には形態素解析器 Mecab の辞書として利用可能で、かつ人間可読、すなわち人に有益な情報を付加した実践医療用語辞書 ComeJisyoV1（登録語数 30,146 語）の無償公開を開始し、以後、随時更新を続け ComeJisyoV5-1（登録語数 77,760 語）を 2013 年 11 月に公開した [8)9)10)11)12)]。

図 1 は、CiNii の論文検索において「テキストマイニング」と「医療」の AND 検索の検索結果 191 件と ComeJisyo のダウンロード数の推移である。ComeJisyoV5-1 を公開した 2013 年以降、ダウンロード数が増加していることが分かる。

なお、2001 年に厚生労働省が掲げた目標は達成されなかったものの、2017 年の 400 床以上の医療施設の電子カルテシステム導入率は 76.3%となり [13)]、医療記録データから症状名や診断名等の用語を自動抽出する技術は益々重要となり、機械学習を用いた研究もなされるようになってきた。

機械学習により高い精度で結果を出すためには、大量の学習用データ（注釈付きコーパス）が必要であり、この学習用データを作成する上でも、医療に関する知識や知見を記述するための実践医療用語辞書が必要である。

表 1　OS 別 ComeJisyo のダウンロード数
（2018.1.15～2019.1.14）

OS	ダウンロード数	割合
Windows	425	64.3%
Mac	146	22.1%
Linux	66	10.0%
Other	24	3.6%
計	661	100%

ComeJisyo の作成過程で、①看護記録、医師記録等は、記入する医療従事者により、また内科と歯科等の診療科により、使われる用語に相違があること、②学術論文や契約書等と異なり、表現や用語の標準化がなされておらず、方言や業界用語のみならず、表記のゆれや誤字脱字等が含まれることを確認している。そして、これらの医療記録文書を精度良く解析するための網羅性の高い辞書の作成が困難であることも認識している。

今回、①Windows 環境での利用が約 6 割であり、Linux および MacOS での利用が約 3 割であること（表 1）、②UTF コードを前提とするテキストマイニング等のツールを利用するために、ComeJisyoV5-1 を UTF コードに変換する際、Shift_JIS で扱える半角カタカナや機種依存文字が正しく表示されないこと、③看護師国家試験の冊子体で使われる印刷標準字体の中には、Shift_JIS では扱えない第 3 水準の漢字「剝」「掻」「頬」「嚢」「塡」が使われていること[14]から、(1)教育・研究利用と(2)医療施設の環境(Windows、Shift_JIS)での利用の用途別に 2 種類の辞書を作成・公開することとし、UTF 版の ComeJisyoUtf8-1 を昨年 2018 年 11 月に公開した[15]。

本発表では、登録語数約 10 万語の Shift_JIS 版の ComeJisyoSjis-1 について述べる。

2. ComeJisyo の公開履歴

形態素解析器 MeCab 用の辞書を作成するためには、コストを学習するための大量の注釈付きコーパスが必要であるが、注釈付き医療記録コーパスの作成は困難である。

そこで ComeJisyoV1 と ComeJisyoV2 では 1994 年の毎日新聞 CD-ROM を購入し、奈良先端科学技術大学院大学において改変された RWC (Real World Computing)コーパスを用いてコストを設定し、システム辞書として公開した。

しかしながら新聞記事を基にした学習用コーパスに医療用語が出現する確率は低いため、適切なコスト設定ではなく、登録語が過分割されるなど、語分割の精度は芳しくなかった。

そこで、ComeJisyoV3 以降は、MeCab のシステム辞書として推奨される ipadic のコストを参考に人手でコストを設定し、ipadic と併用するユーザ辞書として公開している[10]。

下記に ComeJisyo の公開履歴をまとめる。

(1) ComeJisyoV1：登録語数 30,146 語
公開日：2008 年 11 月
特　徴：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。改変版 RWC によりコストを設定し、システム辞書として公開。

(2) ComeJisyoV2：登録語数 34,142 語
公開日：2010 年 1 月
特　徴：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。改変版 RWC によりコストを設定し、システム辞書として公開。

(3) ComeJisyoV3：登録語数 41,592 語
公開日：2011 年 3 月
特　徴：医療施設 2 施設より提供された倫理的配慮のなされた看護経過記録および看護模擬経過記録から、臨床看護経験者 3 名が抽出・選定した用語 7,450 語を追加。ipadic のコストを参考に人手でコストを設定し、ユーザ辞書として公開。

(4) ComeJisyoV3-1：登録語数 41,542 語
公開日：2011 年 12 月
特　徴：ComeJisyoV3 の品詞の見直しを行い、助詞を含む用語を削除。過分割される 318 語のコストを調整。

(5) ComeJisyoV4：登録語数 77,647 語
公開日：2012 年 1 月
特　徴：新たにプログレスノートを加えた 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。

(6) ComeJisyoV5：登録語数 77,775 語
公開日：2012 年 3 月
特　徴：一般的な用語ながら「根性（コンセイ）」「呼名（コメイ）」「嫌気（ケンキ）」等、医療分野特有の読み方をする語 128 語を追加。

(7) ComeJisyoV5-1：登録語数 77,760 語
公開日：2013 年 11 月
特　徴：経済連携協定により 2008 年以降受け入れが始まった外国人看護師候補者の教育利用を想定し、ComeJisyoV5 の用語の見直しを行い、看護師および管理栄養士国家試験（2008 年 - 2013 年）での出現情報および、看護師国家試験問題に含まれる英語表記の疾病名を付加。

(8) UTF 版 ComeJisyoUtf8-1：75,831 語
公開日：2018 年 11 月
文字コード：Utf-8（BOM 無し）

特　徴：ComeJisyoV5-1 の登録語を Unicode 正規化形式 NFKC 形式に変換して得られた語を基に、看護師、助産師、管理栄養士国家試験（2013 年 - 2017 年）の出現情報と、看護師及び管理栄養士養成校で使われている教科書 54 冊の索引における出現情報を追加。

文字コードを Utf-8 にしたことで、全角英数字と半角カタカナ、そして丸数字・機種依存文字を含む登録語は削除された。但し「Ⅱ型糖尿病」等の複合語内のローマ数字は残している。また、JIS 第 3 水準の漢字「剝(剥)」「頬(頬)」「嚢(嚢)」「塡(填)」「掻(掻)」を含む語は登録している。

3. ComeJisyoSjis-1 の概要

3.1 登録語候補

前述の ComeJisyoV1 から ComeJisyoV5-1 は、看護領域の文書から抽出した語を登録語としている。ComeJisyoSjis-1 には、新たに医師経過記録に含まれる語を登録している。

具体的には、下記の言語資源から抽出した 102,893 語を ComeJisyoV5-1 の登録語 77,760 語に追加した 180,653 語より重複する語を除いた 112,728 語を登録語候補とする。

(1) 医師経過記録 1 年分
　抽出語：43,105 語
(2) 看護師国家試験（2012 年 - 2016 年）
　抽出語：6,422 語
(3) 助産師国家試験（2013 年 - 2015 年）
　抽出語：1,714 語
(4) 管理栄養士国家試験（2012 年 - 2016 年）
　抽出語：6,427 語
(5) 看護師＆管理栄養士養成用教科書 54 冊
　索引見出し語：45,225 語

3.2 登録語の単位

辞書を作成する際、登録する語の単位を定めることが重要である。

実践医療用語辞書の作成に着手した 2004 年、医療用語は標準化されておらず、医療記録データには、専門用語をはじめ、略語、隠語、外来語、そして合成語が含まれることは分かっていたが、これらの語構成や語種構成の実態は不明であった。

そこで、単位認定の方針や規則を定めずに，臨床看護の経験者が「一つのまとまった語」と判断したものを 1 単位とした。従って、ComeJisyo の登録語には、助詞・助動詞を省略した臨時一語を含む複合語が多くある。

表 2 は、ComeJisyoV5-1 の登録語 77,760 語を国立国語研究所が規定した短単位に語分割した結果である。EOS（end of sentence）を除いて 203,562 語（約 2.6 倍）に分割される。

表 2 登録語に含まれる unidic 短単位数

	語数	延べ語数	異なり語数
ComeJisyoV5-1	77,760		
Unidic-cwj-2.3.0	281,322	203,562	24,476

ComeJiysoSjis-1 の登録語においても、医療従事者が「一つの語」と判断する複合語（長単位）の語を登録する。

3.3 付加情報

本辞書独自の付加情報を以下に示す。

1) ①看護経過記録、②プログレスノート、③看護教育用模擬経過記録、④模擬診療記録 [16]、⑤医師経過記録における文書頻度。最大は“5”。
2) 医師経過記録での出現の有無。出現する場合は“D”。
3) 登録語の識別番号。ComeJisyoV1 からの通し番号を付加。

3.4 コストの設定

ComeJisyoV5-1 および ipadic のコストを基に人手で設定する。

4. ComeJisyo の持つ問題

ComeJisyo を作成する過程で気付いた問題には以下のものがある。

(1) 網羅性（語彙の拡充）

医療記録データの入手が困難なこともあり、多様な医療記録を適切な精度で解析できる辞書の作成は不可能である。しかしながら、近年、処置名や薬品名等の標準化が進められ公開されている [17]。

そこで、利用者の用途に合わせて各自で辞書の作成ができるように、人間可読な csv 形式のファイルも併せて公開する。

また、文書頻度に加えて、医師経過記録に出現する語には“D”を付加する。

(2) 読み仮名

医療記録データに出現する語は書き言葉であり、読み方の分からない語も多い。教育・研修利用を考え、看護師国家試験問題冊子に記載のフリガナと市販の用語辞書・事典に記載のフリガナを付加し、これらに記載のないものは、臨床看護および医師の判断により付加している。複数ある読み仮名（頭蓋：「トウガイ」「ズガイ」）については、臨床看護経験者の意見を基に決めている。

(3) 品詞

最右側の単語の品詞を複合語の品詞として付加している。「慢性虚血性変化（文書頻度4）」の品詞は「名詞 サ変接続」となる。

(4) 登録語の単位

名詞連続語等の長単位の語の登録により、システム辞書ipadicとの語単位に乖離が生じ、利用者の想定する解析結果が得られない場合もある。

例えば、「防護用具」は、ipadicのみの解析では「防護｜用具」となるが、ComeJisyoV5-1には「防護用」が登録されており、併用すると「防護用｜具」となる。テキストマイニングの前処理等に利用する際には、留意する必要がある。

5. おわりに

ComeJisyoの登録語数は2008年から10年を経て、約3万語からComeJisyoSjis-1では約10万語となった。これらには助詞を省略した臨時一語を含む複合語が多く、語彙研究の言語資源としての利用も可能となった。

現在、筆者らは登録語（複合語）の語構成要素の解析とこれらの意味分類に着手しており、得られた知見も、随時公開する予定である。

なお、ComeJisyoSjis-1は、準備が整い次第公開する。

謝辞

本研究は、科学研究費補助金（18H03499）および西南女学院大学共同研究費による助成を受けています。

参考文献

1） 相良かおる: 看護記録に含まれる文書の統語構造, 日本医療情報学会　第5回看護情報研究会論文集,pp.85-88. 2004
2） 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, 人文科学とコンピュータシンポジウム, p.103-110, 2010
3） 小木曽智信，相良かおる: 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発表論文集, p.158-161, 2012
4） 相良かおる,小作浩美,小暮潔: 標準看護実践用語の特徴，第6回看護情報研究会論文集, P.73-75. 2005
5） Kaoru Sagara, Akinori Abe, Hiromi itoh Ozaku, Noriaki Kuwahara, and Kiyoshi Kogure: Features of Standardized Nursing Terminology Sets in Japan, In Proceedings of the 9th on Nursing Informatics (NI2006) , p.471-475,2006
6） 相良かおる,小作浩美,小暮潔,納谷太,桑教則彰: 看護文書の意味解析用辞書の構築における ICNP®と「分類語彙表」の活用可能性, 医療情報学　第24巻　第6号, p.657-665, 2005
7） 相良かおる,黒田裕子,小田正枝,岡崎寿美子,山勢博彰,城戸茂里,平尾百合子,棚橋泰之,林みよ子,脇坂浩,中木高夫:看護支援システムの稼動状況 予備的研究としての半構成的面接調査報告,看護診断学会第11巻第1号　pp.18-28　2006
8） 相良かおる,浅原正幸,小野正子,小作浩美: 形態素解析器 MeCab 用看護用語ユーザ辞書の作成と公開, 第28 回医療情報学連合大会論文集,p.938-939, 2008
9） 相良かおる,浅原正幸,小野正子,外山健二: 形態素エンジン MeCab 用辞書 ComeJisyoV２および看護教育支援用かな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,p.983-984,2009
10） 相良かおる,小野正子,小木曽智信,小作浩美: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 発表論文集,p. 621-624 ,2012
11） 相良かおる,小野正子,小作浩美,鈴木隆弘,高崎光浩,嶋田元:分かち書き用辞書 ComeJisyo の評価,医療情報学　第32巻　第6号,p.301-307.　2012
12） 相良かおる,小野正子:実践医療用語辞書 ComeJisyo の紹介,第33 回医療情報学連合大会論文集, p.828-830,2013
13） 一般社団法人 保健医療福祉情報システム工業会：https://www.jahis.jp/action/id=57?contents_type=23
14） 相良かおる,橋本直幸,小野正子:看護師・助産師・管理栄養士国家試験に含まれる漢字調査,第19回日本医療情報学会看護学術大会論文集,P.137-140, 2018
15） 相良かおる,小野正子,山崎誠. UTF 版実践医療用語辞書 ComeJisyo1.0 の作成:第38回医療情報学連合大会,P.508-511.2018
16） GSK2012-D 模擬診療録テキスト・データ: https://www.gsk.or.jp/catalog/gsk2012-d
17） 一般財団法人 医療情報システム開発センター (MEDIS-DC): https://www.medis.or.jp/4_hyojyun/medis-master/