看護記録文の計量的用語調査

相良かおる* 小野正子* 鈴木隆弘** 嶋田元*** 小作浩美****

* 西南女学院大学 ** 千葉大学医学部附属病院
*** 聖路加国際病院 **** (独) 情報通信研究機構

電子カルテシステムの普及により，テキスト形式の医療情報が蓄積される．我々は医療情報に含まれる用語を収集し，自然言語処理用の辞書を作成している．そして，医療情報に含まれる語 34,142 語を収録した形態素解析器 Mecab 用の辞書 “ComeJisyo V2” を公開している．今回，“ComeJisyoV2” を利用し，2 病院で蓄積された看護記録文の用語調査を行い，語種構成および品種構成を明らかにした．本稿では，この調査結果について述べる．

Quantitative study of nursing records

Kaoru Sagara* Masako Ono* Takahiro Suzuki**
Gen Shimada*** Hiromi itoh Ozaku****

* Seinan Jo Gakuin University ** Chiba University Hospital
*** St. Luke's International Hospital
**** National Institute of Information and Communications Technology

With the spread of the medical record system, medical information has been accumulated in co-medeical machine-readable documents. We are collecting the words in co-medical documents and constructing a dictionary data for natural language processing system. We have released a dictionary "ComeJisyo V2" for a word segmenter program Mecab, which includes about 34,142 words. “ComeJisyo V2” was used for vocabulary research of the nursing records which were accumulated in two hospitals. Then, the structure of the word class and the part of speech were clarified. The paper describes the results of investigation.

1 はじめに

政府主導により電子化が進み，電子カルテシステムを導入する施設が増え，日々テキスト形式の医療情報（以下，「医療情報」という）が蓄積される．そして大量に医療情報が蓄積されると，これらの活用が要望される．実際，“医療情報” & “2 次利用” を検索語として Google scholar で検

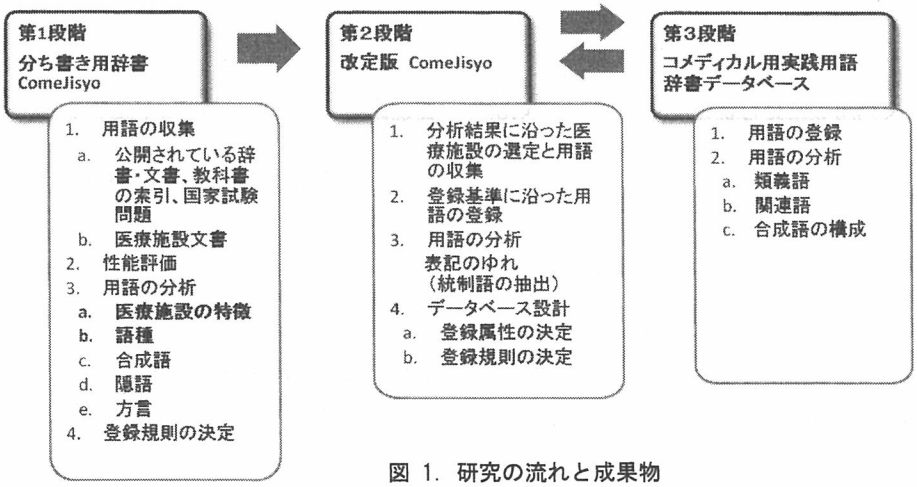

図 1. 研究の流れと成果物

索したところ，検索結果は 94 件，“医療情報” &“テキストマイニング”では 113 件（2010 年 10 月 20 日時点）であった．

テキスト形式で蓄積される医療情報から新たな情報や知識を発見するためには，テキストの構造化が必要となる[1]．

そしてテキストの構造化は，

（1） 文字列を言語単位に切り分ける

（2） 単語ならびに文としての構造を付与する

（3） 構造を意味に結びつける

といった言語処理が必要となる．

しかしながら，日本語の文を意味に結び付けられる言語単位に切り分ける処理は容易ではない．

現在，いくつかの形態素解析プログラムが開発され，公開されており，新聞や論文，書籍等では比較的良い精度で形態素に分割できるが，新語や略語，記号，外字等が多く含まれ，体言止め等，形式性の低い日本語テキストでは上手く解析できない．

一般に形態素解析器は単語単位に分割するために形態素解析辞書を用いており，新聞記事に出現するような一般の用語に対する形態素解析辞書は 20 万語規模で整備され公開されている．

しかし，個人情報を含み倫理上公開が困難な医療情報に含まれる実践用語や専門用語を登録した形態素解析辞書の整備と公開は発展途上にある．

また，医療情報を 2 次利用するためには，形態素解析後に，複数の単語を意味のある語にまとめる必要があり，そのための計算機処理用辞書も不可欠である．

そこで我々は，医療情報の自然言語処理を支援するためのコメディカル用実践用語データベースの作成に取り組んでおり，現在 Web 上で公開されている関連文書や辞書を基に形態素解析器 Mecab[2]を利用した分ち書き用辞書 ComeJisyo[3]（登録語 34,142 語）を作成し公開している（図 1 第 1 段階）．

医療情報を適切に分ち書きするには，以下の処理が必要である．

(1) 実際に医療施設で使われる用語を収集し，辞書に登録・拡充すること．

(2) 利用者にとって「適切な分ち書き」とは何かについて検討し，辞書に登録する単位認定の方針を明確にすること．

(3) Mecab は機械学習用の品詞タグ付きコーパスにより，単語を同定する上で必要なコストを求めしていることから，登録語を含む機械学習用の品詞タグ付きコーパスの整備．

本稿では，単位認定の規則に従って作成された改定版 ComeJisyo の作成（図 1 第 2 段階）に向けて行った，2 つの医療施設の看護記録文の用語調査について述べる．

2 ComeJisyo の概要

2.1 日本文の分ち書き

入力文中の単語を同定し，その語形変化を解析する処理を形態素解析（morphological analysis）という．形態素解析の基本的な機能は，単語分割，活用処理，ヨミガナ振り，品詞付与である[4]．形態素解析を行うソフトウェアを形態素解析器と呼び，無償で入手できるものに，Juman, ChaSen, Mecab 等がある．

単語間にスペース等の区切りのない日本語における単語の単位の認定は，形態素解析辞書に登録された単語によって決まり，Mecab 用の形態素解析辞書には，品詞や読み等の情報に加えて，登録された単語を含む品詞タグ付きコーパスを基に機械学習により得られたコストが付与されている．

すなわち，形態素解析辞書に単語ではなく複合語を登録し，その複合語を含む品詞タグ付きコーパスにより機械学習しコストを求めることで，複合語を単位として分割することが可能となる．

例えば，**「時折失見当識見られ，家族不在時にNSコールが押せるか要観察．」**という文を一般的な辞書で形態素解析すると，**「時折｜失｜見当｜識｜見｜られ｜，｜家族｜不在｜時｜に｜N｜S｜コール｜が｜押せる｜か｜要｜観察｜．」**となるが，あらかじめ「失見当識」，「NS コール」，「要観察」を形態素解析辞書に登録し，これら含む機械学習用品詞タグ付きコーパスを作成してコストを求めることで，**「時折｜失見当識｜見られ｜，｜家族｜不在｜時｜に｜NSコール｜が｜押せる｜か｜要観察｜．」**というように分ち書きすることが可能となる．

ただし，形態素解析辞書の登録については，「単位認定の方針を定めずに，場当たり的に語を登録すると，さまざまな粒度の単位で語が切り出され，単位の不均質性が生じ，語彙調査等の研究で形態素解析ソフトを利用する際に，問題が生じる[5]．」との指摘がある．

2.2 ComeJisyo の作成方針

個人情報を含む医療情報は，倫理上医療施設外に持ち出すことが出来ないことから，医療情報の入手が困難であった．また，医療施設で蓄積され

る医療情報には，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だろうとの予測は容易であるが，語種構成や品種構成の実態は不明であった．

そこで，第1段階として，単位認定の方針や規則を定めずに，①Web で公開されている看護領域の文書と研究目的で利用許可を得た看護領域文書に含まれる名詞 50,805 語，②看護学の教科書の索引より抽出した語 40,833 語，③看護師国家試験問題 2002 年～2007 年から抽出した名詞 9,478 語，④Web で公開されている医療用語辞書 48,875 語を対象に，2 種類以上に出現している語 30,146 語を抽出し，さらに⑤栄養管理，栄養指導分野のテキストから抽出し 3 名の臨床の管理栄養士により精査・検証し選定した 3,996 語を加えた 34,142 語を登録語として Mecab 用の形態素辞書 ComeJisyoV2 を作成した（図1の第1段階）．

現在 Web で公開しているのは，

(1) NAIST-jdic と ComeJisyoV2 を統合した Linux 版システム辞書，
(2) Microsoft Windows XP および Vista で動作確認済みの Windows 版パッケージ，
(3) CSV 形式の ComeJisyoV2

の3種である．

辞書のフォーマットは，形態素解析エンジン Mecab 用 CSV 形式の Seed 辞書と同一の項目に加え，a) 臨床，b) 教育，c) 国家試験，d) その他と，使われる場面（出典）を 4 つに分類し，属性として追加記述している．

3 調査準備

3.1 調査単位の設定

看護記録文には「右眼窩内異物除去術」等の合成語が含まれるが，今回の調査では，これらを更に短い単位に分割するのではなく，合成語を一つの単位として扱う．また，看護記録文に含まれる検査値等の数値については，すべて 9 に置き換えており，数字と mg 等の助数詞は別の単位とする．

例:

分泌物｜多く｜咳嗽反射｜乏しい｜ため｜肺ケア｜必要

99｜時間｜以内｜の｜停止｜が｜9｜回｜発生

3.2 見出し語の抽出基準

(1) 同じ見出し語とする場合：
- 煙草（煙草，たばこ，タバコ）
- %（パーセント，%）

(2) 異なる見出し語とする場合：
- 動詞の活用形
- ＋，足す（意味が同一でない場合）
- Scale，スケール（英字とカタカナ表記）
- 9，⑨，九，Ⅸ（異なる数字の表記）

3.3 語種の種類

本稿で扱う語種は以下のとおりである．

和語：ひらがなと訓読みの言葉
漢語：音読みの漢字で表記された言葉
カタカナ語：カタカナで表記された言葉
英字：アルファベットで表記された言葉
数字：アラビア数字“9”，漢数字，ローマ数字（Ⅰ，Ⅱ，Ⅲ...），点つき数字（ⅰ，ⅱ，ⅲ...）
混種語：和語と漢語の組み合わせを除く，2 種以上の語種を含む言葉

3.4 品詞情報の種類

今回，集計の対象とする品詞は，名詞（体の類），動詞（用の類），形容詞（相の類），助詞・助動詞（機能語）の 4 種で，他は「その他」として纏めている．

4 調査手順

4.1 調査データ

国立大学法人A病院と公益法人B病院の看護記録文から同文を削除した後，文字長の降順に並べ替え，総数を 3,000 で割った値で等間隔に 3,000 行を抽出し，半角文字を全角に変換したものを調査データとする．

4.2 調査手順

（1） 語種・品詞の調査

① 分ち書き：ComeJisyoV2 を含むシステム辞書を用いて，形態素解析器 Mecab0.98 で解析し，人手で見直し，品詞および誤字の修正と，単位語の抽出を行う．

② 見出し語の抽出：①で求めた単位語について，解析用辞書 UniDic（Version1.3.12）[6]と照合し，見出し語を抽出した上で，人手で抽出基準に従って見出し語の抽出を行う．

③ 単位語の総数（延べ語数）を求め，見出し語の総数（異なり語数）とそれぞれの見出し語の出現頻度を求める．

④ 見出し語を，語種別および品種別に分類する．

（2） ComeJisyo の有効性の検証

A 2 病院それぞれについて，調査データ 3,000 行に含まれる ComeJisyoV2 の登録語とその

出現頻度を求める．

B　①により分ち書きされた語について，ComeJisyoV2により切り出された登録語とその出現頻度を求める．

C　①により分ち書きされた語について，Mecabの未知語処理により切り出された登録語とその出現頻度を求める．

D　ComeJisyoへの登録候補となる過分割された合成語の種類を求める．

5 結果

表1は語種・品詞の調査結果をまとめたものである．

表1．語種・品詞の分類と2病院に共通の異なり語数

	A病院				B病院				2病院共通異なり語数		
	異なり語数		延べ語数		異なり語数		延べ語数		語	A病院	B病院
和語	991	29.1%	15,497	54.3%	1,270	20.6%	23,163	45.6%	484	49%	38%
漢語	2,127	47.8%	6,763	23.5%	2,959	48.1%	10,261	20.2%	917	43%	31%
カタカナ語	358	8.2%	703	2.4%	529	8.6%	1,247	2.5%	129	36%	24%
英字	158	3.6%	783	2.9%	597	9.7%	2,606	5.1%	86	54%	14%
数字	—	—	1,460	5.1%	—	—	3,211	6.3%	—	—	—
混種	450	10.0%	661	2.3%	731	11.9%	1,165	2.3%	49	11%	7%
記号	51	1.2%	2,167	9.5%	71	1.2%	9,163	18.0%	32	63%	7%
計	4,135	—	28,034	100.0%	6,157	—	50,816	100.0%	1,697	41%	28%
名詞	3,451	77.9%	13,250	46.2%	5,419	88.0%	21,925	43.1%	1,437	42%	27%
動詞	346	13.7%	3,395	15.6%	428	7.0%	5,194	10.2%	154	45%	36%
形容詞	80	2.8%	729	2.7%	82	1.3%	379	0.7%	39	49%	48%
助詞・助動詞	104	2.6%	9,485	28.3%	110	1.8%	13,928	27.4%	92	88%	84%
その他	154	3.1%	1,175	7.2%	118	1.9%	9,390	18.5%	—	—	—
計	4,135	100.0%	28,034	100.0%	6,157	100.0%	50,816	100.0%	1,722	42%	28%

今回，句読点で区切ったものを1行としたことで，体言止めで句読点のない記述の多いB病院の調査データ 3,000 行に含まれる単位数は，延べ語数 50,816 語となり，A病院 28,034 語の約 1.8 倍，異なり語数においては 6,157 語となり，A病院 4,135 語の約 1.5 倍となった．

語種構成をみると，2病院共に異なり語数で漢語が最も多く，次に和語が多くなっている．なお，和語には出現頻度の高い助詞や助動詞が含まれているため，延べ語数では，和語と漢語の割合が逆転している．

文献[7]の表 4.6 によれば，雑誌 90 種に含まれる語の語彙調査（1962）の結果，異なり語数における和語の割合 47.50%，漢語 47.59%，外来語 9.77%，混種語 6.02%，延べ語数における和語の割合は 53.86%，漢語 41.27%，外来語 2.92%，混種語 1.95%とある．

調査単位が本調査と異なるため，単純な比較はできないものの，和語の異なり語数の割合 29.1%と 20.6%は，47.50%よりもかなり少なく，漢語の延べ語数の割合 23.5%と 20.2%も 41.27%より少なくなっている．一方，英字とカタカナ語を外来語とすると，A病院は異なり語数で 11.8%（8.2%＋3.6%），延べ語数で 5.3%（2.4%＋2.9%），B病院は 18.3%（8.6%＋9.7%）と 7.6%（2.5%＋5.1%）となり，雑誌 90 種の外来語の異なり語数の割合 9.77%，延べ語数の 2.92%と比べ多くなっており，雑誌 90 種に含まれる日本文と，医療施設で蓄積される医療情報とでは，語種構成がかなり異なっていることが推測される．

二つの病院を比較してみると，和語，英字，記号の延べ語数に占める割合（以下，使用率という）は，A病院の和語の使用率 54.3%，英字 2.9%，記号 9.5%に対し，B病院の和語は 45.6%，英字は 5.1%，記号は 18%と，相違がみられる．そこで 2 病院の記録文の語種構成について，和語，

英字，記号，その他の4項目でχ^2値を求め独立性の検定を行ったところ，χ^2値＝1958.48 となり，χ^2 値 (3, 0.05)＝7.81 よりも大幅に大きく，語種構成に有意な差があることがわかる．

品詞構成においても，品詞5項目の検定においてχ^2値＝3573 となり，χ^2値 (4, 0.05) ＝9.49 を大きく上回っており，有意な差があることがわかる．

実際，A 病院の記録文に比べ，B 病院では英語を交えた体言止めの記述が多くみられる．また，増加，上昇，減少，降下，を記号“↓”，“↑”で記述する傾向がある．

例:

1) **本日FOLEY抜去し，自尿999ML／D帯**

2) **寝ているとSPO2 99前半まで↓だが，体位を変えたりしてSPO2↑**

次に，タイプ・トークン比 (TTR) を見ると，A 病院は 0.147 (4,135 語／28,034 語) で B 病院の 0.121 (6,157 語／50,816 語) より大きく，A 病院の方が多様な語を用いていると考えられる．

名詞の異なり語数をみると B 病院は A 病院の約 1.6 倍 (5,419 語／3,451 語) となり，これは総異なり語数の差約 1.5 倍 (6,157 語／4,135 語) と近い値であり，また，両異なり語数に共通する名詞は 1,437 語で A 病院の 42% (1,437 語／3,451 語)，B 病院の 27% (1,437 語／5,419 語) であることから，両病院で使われる名詞に相違があることが分かる．

表2. 調査データに含まれる ComeJisyo 登録語

	A 病院 語の種類		B 病院 語の種類		2 病院共通 語	A 病院	B 病院
A 行内に含まれる ComeJisyo 登録語	848	21%	1371	22%	517	61%	38%
B ComeJisyo により切り出せた登録語	623	73%	953	69%	335	54%	35%
C 解析器の未知語処理で切り出せた登録語	21	2%	151	11%	—	—	—
D 過分割された単語＆合成語	1,430	—	885	—	277	19%	31%

※ A の割合は総見出し語数に占める割合 (848／4,135, 1,371／6,157)，B と C の割合は A に占める割合

表 2 は，ComeJisyo の有効性を調べるために，調査データに含まれる ComeJisyo の登録語 (34,142 語) について調査した結果である．

なお，表 2 の D の過分割された単語＆合成語は，あらかじめ単位認定の方針を決めて抽出したものではなく，臨床経験を持つ看護教員が一つの意味のある言葉と認識したものである．

形態素解析前の A 病院 3,000 行内に含まれる登録語は，848 語であり，そのうち形態素解析により正しく分割された登録語は 644 語で 76% (644 語／848 語) であったが，ComeJisyo の属性が付加されたもの，すなわち ComeJisyo との照合により切り出されたものは 623 語で 73% (623 語／848 語) であり，21 語 (644 語－623 語) は，Mecab の未知語処理により分割された英字であった．B 病院 3,000 行については，登録語 1,371 語であり，形態素解析で正しく分割された登録語は 1,104 語 81% (1,104 語／1,371 語) で，内 ComeJisyo との照合により切り出されたものは 953 語で 69% (953 語／1,371 語)，未知語処理により正しく分割された語は 151 語 (1,104 語－953 語) でこれらも英字であった．

ComeJisyo に登録されているにも関わらず，正しく分割されない原因として考えられることは以下の 3 点である．

(1) 単位認定基準を定めずに合成語を登録していること．

(2) Mecab のシステムの中で，区切り文字または特殊記号として扱われる文字が，調査データすなわち医療情報の中に含まれていること．

(3) 機械学習用コーパスの整備が不十分で新聞記事 1 年分の記事からなる学習用コーパスを使っているため，適切なコストが辞書に付加されず，日常語の分割が優先され，過分割が生じること．

登録語が過分割された例:

未治療 ⇒ 未｜治療
訪問看護 ⇒ 訪問｜看護
歩行障害 ⇒ 歩行｜障害
熱発 ⇒ 熱｜発
脳外 ⇒ 脳｜外

また，機械学習コーパスの不備は，文献[4]で指摘されているような，単位の不均質性，すなわ

ち粒度の異なる単位で分割される原因にもなっている．

不均質性の例（2 語共に ComeJisyo 登録語）：

放射線治療　⇒　放射線｜治療

放射線療法　⇒　放射線療法

一方，連続する数字や連続するアルファベットの中には，Mecab の未知語処理により正しく分割されるものがあり，前述したように A 病院 21 語，B 病院 151 語が未知語処理により正しく分割されていた．この Mecab の機能により，英字を正しく連結し英単語として分割できれば，ComeJisyo への英語の登録は不要となるが現状では，登録されている英単語が全て正しく分割される訳ではない．以下に過分割の例を示す．

英単語の過分割例：

AIR	⇒A｜IR	（空気）
JCS	⇒JC｜S	（日本昏睡スケール）
DR	⇒D｜R	（医師）
NS	⇒N｜S	（看護師）

表3．登録候補（表2の D）の合成語の語種構成

A 病院	B 病院	英字		カタカナ		和語※		漢語		数字		記号	
958	546							○					
221	140			○				○					
52	80	○						○					
44	21					○		○					
30	17	○											
28	11			○									
23	14	○		○									
19	17							○		○			
18	6					○							
9	1			○		○							
5	0	○								○			
5	8							○				○	
4	15	○										○	
4	0			○				○		○			
2	0	○		○				○		○			
2	3	○		○								○	
1	3	○		○				○					
1	0	○		○						○			
1	1	○				○							
1	0			○				○				○	
1	0			○						○			
1	0									○			
0	1	○						○				○	
0	1	○						○		○			
計 1,430	計 885	A 121	B 135	A 293	B 172	A 72	B 29	A 1,307	B 817	A 33	B 18	A 12	B 27

注）　○：合成語に含まれる語種

和語：「痺れ消失」などひらがなを含むものを和語とし「昼絶食」等は漢語に含めている．

今回の調査で，2 病院の調査データに含まれる異なり語の約 2 割が ComeJisyo の登録語であることから，ComeJisyo は医療情報の分ち書きに有益ではあるものの，表 2 の D 過分割された単語&合成語の総数 2,038 語（1,430 語+885 語－277 語）は，調査データに含まれる ComeJisyo 登録語数 1,702 語（848 語＋1,371 語－517 語）よりも多く，実践用語の収集と辞書の拡充は必須である．

また，調査データには検査値の単位「mmg／日」「ml／day」や「V／S」「T－PA」「99.9」「（－）」「++」等の区切り文字や特殊記号を含むもの，機種依存文字「㎎」「㏄」「Ⅲ」「ⅳ」等，そしてローマ数字や矢印記号等も含まれ，形態素解析器 Mecab の機能だけでは適切に分割できないことが明らかとなった．

文献[8]によれば，「医学用語の特質としては，用語の使用方法が非常に粗雑で難解なことが挙げられる．たとえば単位語の無制限に近い集合による造語と，さらに補助語の機能を乱用した“χ化β状α性μ的”等の構造式である．次のような例では，語なのか句なのか，さらにはどれを単位語とすべきかが判断しかねるものが非常に多く抽出された」として“悪性型黒色棘細胞増殖症”，“直腸周囲組織異所寄生”，“電子顕微鏡的細胞組織病理学的研究”の 3 例が提示されている．

今回の調査データの中にも，“高度変動一過性除脈”，“両側腎後性腎不全”“自宅血圧測定施行”などの合成語がみられる．

表 3 は，登録候補となる単語&合成語（表 2 の D）の語種構成をまとめたものである．

両病院共に漢語の割合が多く，A 病院は 958 語で全体 1,430 語に占める割合は 67%，B 病院は 546 語で 62%となっている．第 2 番目に多いのは，「大腸ポリープ切除術」等のカタカナ語と漢語からなる合成語で，A 病院 221 語で 16%，B 病院 140 語で 16%となっている．そして 3 番目に多いのは英字と漢語からなる合成語で，A 病院 52 語で 4%，B 病院 80 語で 9%となっている．

表 1 の語種構成の分析と同じく，B 病院は，A 病院に比べて英字の割合が高く，A 病院 8%（121 語／1,430 語）に対して，B 病院は 15%（135 語／885 語）であり．また，記号が含まれる割合も多く，A 病院が 0.8 %（12 語／1,430 語）に対して B 病院は 3%（27 語／885 語）となっている．

混種語の結合形式については，文献[9]に総合雑誌延べ語数 116,000 語，異なり語数 15,000 語を標本とした調査結果があり，「混種語の延べ語数は 600 語足らず，異なり語数は 300 語余りで，10 回以上繰り返して用いられたものは延べ語数 106 語，異なり語数 6 語」と記載されている．

すなわち，混種語の割合は延べ語数で 0.5%（600 語／116,000 語），異なり語数で 2%（300 語／15,000 語）であり，本調査表 1 での混種の割合，延べ語数で約 2%，異なり語数で約 10%との相違は明らかである．

以下は，文献[9]記載の頻度による順位についての記載部分を抜粋したものである．

① **漢語＋和語（座敷，絵描き等の 10 種）**
② **和語＋漢語（場所，踏み台等の 8 種）**
③ **漢語＋外来語（急カーブ，全スト等の 5 種）**
④ **外来語＋漢語（ガス弾，チフス菌等 5 種）**
⑤ **和語＋外来語（赤ランプ，やみドル等 4 種）**
⑥ **外来語＋和語（インキつぼ，ガラス戸等 4 語）**

表 3 の単語&合成語候補 2,038 語（1,430 語+885 語－277 語）語の内，漢語の合成語と頻度の高い混種語について以下に一例を示す．

① **漢語（958 語&546 語）：**
全量摂取（頻度 21 回），胸部症状（8 回）
② **カタカナ＋漢語（221 語&140 語）：**
ネブライイザー施行（4 回），体動コール（11 回）
③ **英字＋漢語（53 語&80 語）：**
FOLEY挿入（5 回），造影 CT 施行（1 回）
④ **漢語＋和語（44 語&21 語）：**
表情穏やか（5 回），創部滲み出し（2 回）
⑤ **カタカナ（28 語&11 語）：**
ライントラブル（4 回），ペインスケール（3 回）
⑥ **英字＋カタカナ（23 語&14 語）：**
ENBDチューブ（3 回），ヒューマリンR（5 回）

6 まとめ

国立大学法人 A 病院と公益法人 B 病院の記録文を対象に用語調査を行った．

本調査に際し，他の研究成果との比較を容易にするために，文献[10]にある調査単位を検討したが，合成語を単語に分割する作業は，専門用語の知識を持つ臨床で働いた経験のある者にとっても困難であった．また，和語と漢語の分類についても，医療情報にはなじみのない漢字が多く含まれ，厳密に分類することは困難で，正しく分類できてはいない．なお，合成語を単語単位に分割することの難しさについては文献[11]に言及されており，本調査では 3.1 節に記載の通り，合成語を一つの単位としている．

本調査で明らかになったことは次の通りである．

(1) 2 病院の調査データにおける語種および品詞の構造には大きな相違がある．
(2) 雑誌に含まれる日本語に比べて，看護記録文に含まれる語は，英字，カタカナ，数字，記号を含む混種語の割合が高く，合成語が多く含まれる．

また，本調査において ComeJisyoV2 を利用し，有効性を調べたところ，辞書に登録されている語の 7 割は正しく分割されるものの，残り 3 割は正しく分割されず，人手による見直し作業に多くの時間を要した．

厚生労働省で 3 年ごとに実施される医療施設調査では，医療施設を，(i) 病院，一般診療所，歯科診療所などの施設による分類，(ii) 公的医療機関，医療法人，公益法人などの開設者による分類，(iii) 病床規模による分類，(iv) 診療科目別に分類している．そしてその他に，(v) 医療法による，公的医療機関，地域医療支援病院，特定機能病院などの医療制度上の分類もある．平成 21 年 2 月末の医療施設動態調査[13]によれば，医療施設の総数は 176,326 施設，その内，病院は 8,781 施設であった．

あらゆる医療情報に対応できる形態素解析辞書の構築は理想ではあるが，非現実的であり，どのような領域の医療情報を対象にするのか，対象とする医療情報を限定する必要がある．しかし，医療施設別に医療情報を分類すべきなのか，診療科別に分類すべきなのか，そもそも現実的に医療情報の分類は可能なのかすらも不明である．

そこで，厚生労働省が 2001 年に全国 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ている[12]ことから，我々は，400 床以上の一般総合病院を対象にしている．

当面は，分類法 (ii) を用い，様々な開設者による 400 床以上の病院から医療情報を収集し，(1) 単位認定の方法を定めずに，ComeJisyo の登録語の拡充を行い，併行して，(2) 機械学習用のコーパスを整備し，登録語を正しく分割できるよう，解析精度の向上を優先させる．また，単位認定の方針を定める上で重要な合成語の構造を知る上で，(3) 合成語の最後部の単語，例えば「施行」(CT 施行，クーリング施行，介助マッサージ施行，創部全抜糸施行) 等を抽出して頻度情報とともに一覧表として蓄積し，直前に隣接する単語とその共起頻度情報を求めることの 3 項目を優先させる．

他方，用語調査の結果および ComeJisyo の解析精度の情報を医療従事者に開示し，施設内で用いる用語や記述方法の標準化と，医療情報の解析方法の標準化を促していきたいと考えている．

謝辞

本研究は，科学研究補助金　基盤研究 (B)「コメディカル実践用語辞書データベースの作成」(課題番号　21300099) の支援を受けています．

参考文献

[1] 辻井潤一監訳：テキストマイニングハンドブック，東京電機大学出版局，2010
[2] Mecab, http://mecab.sourceforge.net/
[3] ComeJisyo http://sourceforge.jp/projects/comedic/
[4] 松本裕治，影山太郎，永田昌明，齋藤洋典：言語の科学 3　単語と辞書，岩波書店，2004, p.54-92.
[5] 言語処理学会編：言語処理学事典，共立出版，2009, p.142.
[6] UniDic, http://www.tokuteicorpus.jp/dist/
[7] 長尾真，黒橋禎夫，佐藤理史，池原悟，中野洋：言語の科学 9　言語情報処理，岩波書店，2004, p.181.
[8] 斎藤孝：索引作業のための自然語処理の研究－医学用語の計量的調査－，Library Science No.5，1967, p.51-72.
[9] 斎藤倫明，石井正彦編：語構成　日本語研究資料集　第 1 期第 13 巻，ひつじ書房，1997, p.37.
[10] 長尾真，黒橋禎夫，佐藤理史，池原悟，中野洋：言語の科学 9　言語情報処理，岩波書店，2004, p.170 -172.
[11] 文献[9], p.250-267.
[12] 厚生労働省：保健医療分野の情報化にむけてのグランドデザインの策定について http://www.mhlw.go.jp/shingi/0112/s1226-1.html
[13] 医療施設動態調査 (平成２１年２月末概数) http://www.mhlw.go.jp/toukei/saikin/hw/iryosd/m09/is0902.html