実践医療用語の語構成と意味
ー語構成要素語彙試案表の作成に向けてー

相良かおる[1)]　小野正子[1)]　高崎智子[1)]　東条佳奈[2)]　麻子軒[2)]　山崎誠[3)]

1) 西南女学院大学　2)大阪大学　3)国立国語研究所

医療記録には，合成語となる多くの専門用語が含まれるが，その語構成は明らかになってはいない．本研究では医療記録文に含まれる合成語 7,194 語を対象に，語構成解析と意味解析を実施した．その結果，医療の観点から意味的にも統語的にも妥当な語構成要素 5,787 要素を抽出し，これらを意味的に分類するために 93 種類の意味ラベルを設定し，すべての語構成要素に意味ラベルを付与した．また，合成語における語構成要素の出現位置を調べたところ，語頭にくる語構成要素で最も出現頻度が高いものは「先天性」であり，語末では「損傷」であった．意味ラベルの頻度を調べた結果，語頭では「身体部位」の，語末では「病名」の出現頻度が最も高かった．

Construction and meaning of practical medical terms
- Creation of a tentative word construction element lexicon -

Kaoru Sagara[1)]　Masako Ono[1)]　Satoko Takasaki[1)]
Kana Tojo[2)]　Tzu-Hsuan Ma[2)]　Makoto Yamazaki[3)]

1)Seinan Jo Gakuin University　2) Osaka University
3) National Institute for Japanese Language and Linguistics

Medical records contain many synthetic terms, but their word structure is unknown. In the present study, we analyzed the word construction and meaning of 7,194 compound words included in medical records. From the results, we identified 5,787 word construction elements that were valid both semantically and syntactically from a medical standpoint, established 93 different semantic labels to classify the elements semantically, and applied these semantic labels to all word construction elements. An investigation of the position of the word construction elements in compound words revealed that elements that came at the beginning of words most often were [先天性“congenital”], whereas the elements that came at the end of words were [損傷“injury”]. An investigation of the frequency of semantic labels showed that [身体部位“body part”] appeared most commonly at the beginning of words and [病名“disease name”] at the end of words.

1. はじめに

コロナ禍により，オンライン診療が普及することで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る．

日々の電子カルテの入力は時間に追われ，用語の標準化もなされていないことから，電子カルテデータには，略語，施設や診療科特有の業界用語，そして誤字脱字が含まれる．その結果，①データ中のノイズの除去，および②単語分割などの前処理が困難であり，精度の高い自然言語処理は期待できない．

②の単語分割については，形態素解析器 MeCab で利用可能な実践医療用語辞書 ComeJisyo[1]が公開されているが，登録する語の単位が決められていない．そこで筆者らは本辞書を実践医療用語の言語資源と捉え，2018 年度より，ComeJisyoSjis-1 の合成語を対象に語構成の分析に着手している．

本研究の第一目標は，医療記録に含まれる合成語（以下，「実践医療用語」という）を構成する『実践医療用語語彙表』を作成することである．

第二の目標は，実践医療用語の語構成に関する言語学的な知見を得ることである．

第三の目標は，得られた知見を含む成果物を医療実践，医療教育の領域で，出来得れば言語学研究の領域においても利用可能な形で公開することである．

本稿では，合成語 7,194 語を構成する語構成要素とこれらに付与した意味ラベル，そして『語構成要素語彙試案表』について述べる．

『語構成要素語彙試案表』は，医療記録文の分かち書きの支援や合成語の自動抽出，そして意味の推測などの利活用を想定した，第三の目標に関連する研究である．

以降，第 2 章では，本研究で用いる用語を説明し，第 3 章では，関連する研究について述べる．第 4 章では，語構成分析の方法を説明し，第 5 章では，分析結果について述べる．第 6 章では，本

研究で明らかになった課題について述べ，第 7 章で『語構成要素語彙試案表』の作成について説明する．

2. 用語の定義

実践医療用語：

専門分野で使われる「専門用語」には，①学術上の専門用語（学術用語）と，②それ以外の専門用語がある．本研究では，医療施設で使われる医療記録に含まれる①と②を合わせた用語を「実践医療用語」という．

語構成要素：

合成語を構成する要素で，本研究では「医療の観点から意味的にまたは統語的に分割可能なすべての語」と定義する．分割できない合成語については元の合成語を語構成要素とする．

合　成　語：脳幹多発性硬化症
語構成要素：脳幹，多発性，硬化症，多発性硬化症

語構成要素列：

合成語を構成する語構成要素の順序組．医療の観点から一つの意味を持つ語を語構成要素とする短い単位の語構成要素列（短）と，更にこれらを慣用的に使われる単位や統語的に妥当とする単位でまとめた長い単位の語構成要素列（長）の 2 種類がある．

合　成　語：　脳幹多発性硬化症
語構成要素列（長）：脳幹｜多発性硬化症
語構成要素列（短）：脳幹｜多発性｜硬化症

意味ラベル：

語構成要素に付与する，意味を表すラベルで，共同研究者等で命名したものである．

「回旋位」のように他の語構成要素と結合することで意味が決まる多義の場合は「・」により「状態・位置」のように複数の意味ラベルを列挙する．また，「胃粘膜下」のように位置と身体部位の両方の意味を持つ場合は，「／」を用いて「位置／身体部位」のように表記する．

意味ラベル列：

語構成要素列の各語構成要素に付与された意味ラベルからなる順序組．

語構成要素列（長）：脳幹｜多発性硬化症
意味ラベル列：　身体部位｜病名
語構成要素列（短）：脳幹｜多発性｜硬化症
意味ラベル列：身体部位｜病態｜病態・状態

短単位：

短単位とは，国立国語研究所が言語の形態的側面に着目して規定した斉一な言語単位である．現代語において意味を持つ最小単位を規定した上で，最小単位を短単位の認定規定に基づき結合させる，または結合させないことにより，認定される．

合　成　語：脳幹多発性硬化症
短 単 位 列：脳幹｜多発｜性｜硬化｜症

右側主要部の規則：

形態的に複雑な語（合成語）の統語的な性質・役割を決定する主要部はその語の右側の要素（語末）にあるという規則である．

3. 関連研究

ここでは医学および看護学の領域における専門用語に関する研究の概要を述べる．

医学用語（医学分野での学術用語）の問題点について開原（2010）は，標準化ができていないこと，一般人からみて難解なことを挙げ，標準化の問題として①概念の標準化が困難なこと，②同義語の問題，③表記法の「ゆれ」の問題を挙げている[2].

実践医療用語においても同様であり，「摂食禁止」のことを西日本では「絶食」，関東では「禁食」を用いるなど地域間較差があること[3], また，看護記録には，「粘稠（ねんちゅう）」を「粘調（ねんちょう）」，「自己抜去」を「事故抜去」というような誤字・誤変換が含まれることが分かっている[4].

医学用語の計量的調査法については斎藤（1967）が詳述している[5]. 暫定的に医学用語の認定単位を①非重要語（助詞，助動詞，接辞など），②合成語，③単位語（漢字 2 字の集合，および意味的には重要な価値を持たず，慣習的に付加する語など），④特殊単位語（「胃」「肝」などの漢字 1 字でも用語として認められるもの，「～性」などの漢字一字で接尾的な機能を果すもの，外国語，「顕微鏡」のような分離することで本来の概念を示すことのできない漢字 2 字以上の語）の 4 種に規定し，論文の標題 1,000 から抽出した単位語 8,037 語（延べ数）の分析と評価を実施している．

接尾的機能を果たす語の中で，最も頻度が高いものは「～的」であり，本来の「らしさ」という性質を示す以外に，中国語の「的」に等しい格助詞「の」働きをすること，性質または属性を表現する機能を持つ「～性」の中には，意味的差異が明らかではないもの（「淋巴肉腫」「淋巴性肉腫」）があること，これらを乱用した“x 化 β 状 α 性 μ 的”などの構造を持つ難解で単位語に分割するのが困難な合成語が多くあることを指摘している．

本稿で扱う合成語 7,194 語の語構成解析においても当初は「肝」や「腎」は語と認定しない方針であったが，現在は一語として扱っており，上記 4 種の認定単位を支持する．なお本合成語の語構成分析では，「～性」の頻度が最も高いことが分かっている[6].

更に斉藤（1967）は，単位語の性格の分析に，独自のカテゴリーⓐ（器官，器具：胃），ⓑ（症

状，現象：腫瘍）,ⓒ（用途，性質：～性などの合成語），ⓓ（対象：マウス），ⓔ（操作，処置：移植）を用いたカテゴリー化による尺度の分析と，合成語の生起頻度を構成する単位語の生起確率の積を用いた統計的アプローチによる尺度を用いた分析を実施している．加えて，日本語と英語の対応上の問題として表記体の相違と，語の示す概念の差違を指摘した上で，単位語 382 語と英語の索引用語集 MeSH(Medical Subject Headings)との対応を行い，完全に適合したものが約 51%あったことを報告している．

劉（2000）は，同じ医学でも専門領域によって意味の解釈が異なること，「右上葉肺扁平上皮癌」を「概念が失われない最小の単位」で分割する場合，臨床医学の領域では，「右｜上｜葉｜肺｜**扁平上皮癌**」となるが，解剖学の領域では，「右｜上｜葉｜肺｜**扁平上皮**｜**癌**」となることを指摘している[7]. そして医学用語集には意味構造の記述が大切であるとしつつも，実用的でかつ適応対象がある程度広い用語集として，独自の意味構造を構築せず，ICD[8]や SNOMED-CD[9]などの既存の分類との関連を，それらに対応させ，意味構造を持たせない構造化臨床医学用語集を提案している．

表 **1 関連研究と研究対象**

	分　野	
	医　学	看護学
学術用語	・現状と課題[2] ・計量的用語調査法[5] ・構造化用語集[7] ・構造解析[12]	
実践医療用語		・用語の病院間較差[3] ・誤字調査[4] ・既存の分類との照合[10][11]

実践医療用語に関しては，ComeJisyo の登録語 739 語と MeSH のカテゴリーに準拠した「医療用語シソーラス」との照合を行い，表記体とその語の意味が一致したものが 228 語（約 31%）あり，患者の状態を表す語を分類するカテゴリーがないことを明らかにしている[10]．また，日本語の一般用語を分類したシソーラスである『分類語彙表 増補改訂版』との照合において，体内の臓器や管腔に挿入・内在された状態を示す「留置：indwelling」が「捕縛・保釈」に分類されるなど，一般的な語の意味と異なる語があり，既存の分類との関連を対応させることが難しいことが分かっている[11].

合成語の構造解析については，小山（1994）が，約 2,200 の人体部位関連語から，人手で約 1,000 の要素語を暫定的に選び 5 つのカテゴリーに分類し，合成規則を用いた合成語の要素分解の実験を行っている[12].

誤字脱字がない，専門辞書に立項されている専門用語を多く含む学術医学用語を対象とした研究は半世紀前から行われているが，個人情報を含む門外不出の医療記録に含まれる実践医療用語を対象とした研究は少なく，また実態調査に留まっている（表 1）．

4. 方法

本研究に着手した 2018 年から現在までの分析方法を以下に述べる．

4.1 言語資源

医療の知識を持たない共同研究者による意味的な分割を考慮し，分かち書き用実践医療用語辞書 ComeJisyoSjis-1 の登録語 111,664 語より一般的な語（『分類語彙表 増補改訂版』[13]収録の語）を含む合成語 7,194 語を本解析データとする．

4.2 手順

手順 1.　機械的分割（機械的順序列）

A)　形態素解析器 MeCab0.996[14]と見出し語約 87 万語の解析用辞書 UniDic[15]により，合成語を短単位に分割

B)　付与された品詞ラベルが「形状詞」または「記号」の場合，品詞ラベルを「名詞」に変更

C)　「接尾辞」の語を直前の語に連結

D)　「接頭辞」の語を直後の語に連結

E)　連続する「カタカナ」のみの語を連結

短単位列：肝内｜門脈｜下｜大｜静脈｜短絡

手順 2.　意味的分割（意味的要素列）

共同研究者 5 名で手順 1 の語を分担して意味的に妥当な語にまとめた語構成要素の順序組（以下，語構成要素列）を作成し，表 2 の意味カテゴリー[16]を参考に各語構成要素に意味ラベルを付与した．なお，共同研究者の専門領域は，日本語学が 3 名，情報科学が 1 名，看護学が 1 名である．

表 2　意味カテゴリー

自然物	動植物	物品	食品
道具	薬品	力	人間
機械	衣料	部分	家具
資材	地類	容器	建物
空間	形状	数量	動き
状態	時間		

要素列：肝内｜**門脈下**｜大静脈｜短絡

意味列：空間｜身体部位・位置｜身体部位｜状態

手順 3.　医療的観点による分割（医療的要素列）

(1) 看護学および情報科学の共同研究者 2 名により，医療的な観点からの見直しを行い，意味により分割した「短い単位」の語構成要素列（短）から統語的な纏まりを考慮した「長い単位」の語構成要素列（長）を試作した．

(2) 臨床看護の経験者に委託し，医療の観点からみて意味的にも統語的にも妥当な 2 種類の語構成要素列「長い単位」と「短い単位」を作成した．以下は参考にした辞書である．

- 医学書院　医学大辞典
- 医学英和辞典　第 12 版
- 医学書院　看護大事典 第 2 版
- 南山堂　医学大辞典　第 20 版
- ステッドマン医学大辞典 改訂第 6 版
- ブリタニカ国際大百科事典　2019
- 広辞苑　第 7 版

要素列（短）：肝内門脈｜**下大静脈**｜短絡

意味列（短）：身体部位｜身体部位｜状態

要素列（長）：肝内門脈下大静脈｜短絡

意味列（長）：身体部位｜状態

手順 4.　意味ラベルの見直し

臨床医の経験を持つ共同研究者により，語構成要素に付与された意味ラベルについて医療の観点から見直しを行った．

要素列（長）：肝内門脈下大静脈短絡

意味列（長）：**身体部位・病態**

手順 5.　計量的調査

① 手順 2 と，手順 4 の「長い単位」および「短い単位」の語構成要素，意味ラベル，意味ラベル列の頻度調査を行う．

② 手順 4 の「長い単位」と「短い単位」の語構成要素列と意味ラベル列それぞれを統語し重複を削除した異なり数を求める．

③ ②について語構成要素と意味ラベルの語頭，語末，語中の頻度調査を行う．

5. 結果と考察

5.1 語構成要素

表 3　合成語 1 語あたりの要素数の分布

要素数	手順 2		手順 4（短）		手順 4（長）	
1	480	6.7%	568	7.9%	882	12.3%
2	3,528	49.0%	3,993	55.5%	5,289	73.5%
3	2,477	34.4%	2,159	30.0%	946	13.1%
4	603	8.4%	405	5.6%	64	0.9%
5	88	1.2%	56	0.8%	13	0.2%
6	18	0.3%	13	0.2%	0	0.0%
計	7,194	100.0%	7,194	100%	7,194	100.0%

合成語 1 語に含まれる語構成要素数の分布を表 3 に示す．手順 2 における最大分割数は 6 要素であるが，手順 4 の「長い単位」による分割では 5 要素となっている．表 4 は，全合成語 7,194 語より分割された語構成要素数をまとめたものである．手順 4 の「短い単位」は手順 2 に比べて，延べ数で 0.95 倍，異なり数は 1.1 倍となっている．「長い単位」においては延べ数 0.82 倍，異なり数 1.2 倍となっている．

表 4　語構成要素数の概要

	延べ数	異なり数	異なり数
手順 2	17,928	4,078	
手順 4(長)	14,622	5,079	**5,787**
手順 4(短)	16,993	4,320	

表 5 は，語構成要素列を比較したものである．手順 4「短い単位」と手順 2 の比較では，85.4%が一致し，「長い単位」との比較では 61.1%が一致している．一般的な日本語を含む合成語であっても，「肝内｜門脈｜下｜大｜静脈｜短絡」の「下」の結合位置を判断することは難しく，医療の実践的な知識が必要であることが分かる．

なお，本研究の成果物として作成する『語構成要素語彙試案表』では，手順 4 で得られた語構成要素 5,787 語(表 4)とこれらに付与された意味ラベル 93 種(表 6)を用いる．

表 5　語構成要素列の比較

	一致		不一致		計
手順 4(長)：手順 2	4,398	61.1%	2,796	38.9%	7,194
手順 4(短)：手順 2	6,143	85.4%	1,051	14.6%	7,194
手順 4(長)：(短)	5,073	70.5%	2,121	29.5%	7,194

5.2 意味ラベル

表 6 は，語構成要素に付与した意味ラベルの概要である．手順 2 の段階において，語構成要素 1 要素に付与する意味ラベルは一つと定め，意味ラベルを設置した．しかし，合成語の意味を理解している臨床経験者には，意味ラベルを一つに限定できない語構成要素があった．

例えば「萎縮」や「拡張」は，状態（結果）を表す場合や，変化（過程）を表す場合，増減の多寡を表す場合などがあり，意味ラベルを一つに限定することが困難である．そこでこのような語構成要素には，該当する意味ラベルを「・」で列挙することとし，また，上位・下位関係にある概念については「＞」を用いることとした．その結果，手順 2 では，95 種類の意味ラベルを用い，4,078 要素に付与した意味ラベルの組は 133 種類となり，複数の意味ラベルからなる意味ラベルの組は

40 種類であった.

語構成要素：肝萎縮
意味ラベル：状態・変化＞増減

その後，手順 2 の結果について，多義の語構成要素には，他の要素と結合することで意味が決まるものと，他の要素との結合に関係なく多義のものがある．「・」ではこれらの区別が出来ないこと，また，統括する範囲が曖昧であることを含めて上位・下位の関係を「＞」で表現することの妥当性について意見が出され，複数の意味ラベルを列挙する際には「・」(or)と「／」(and)を用いることとし，5,787 要素に付与した意味ラベル列は 147 種類となり，複数の意味ラベルからなる意味ラベル列は 58 種類となった.

表 6 意味ラベルの概要

		要素数	付与ラベル (多義含む)	多義 (内数)	ラベル種類
手順 4	(長)	5,079	132	(56)	84
	(短)	4,320	137	(49)	93
	統合	**5,787**	**142**	**(58)**	**93**
手順 2		**4,078**	**133**	**(40)**	**95**

臨床看護の経験を持つ共著者と研究協力者より，医療を要する状態を表わす意味ラベル「症状」の追加が提案され「症状」を追加した．その後，臨床医の経験を持つ共著者による見直しの中で，「病態」が追加された．「状態」「症状」「病態」の大まかな違いを以下に示す.

状態：医療の要・不要に関わらず，外からみて分かる状態.
症状：本人の訴えによる状態
病態：体の中で起こっている機序

表 7 変更された意味ラベル

追加：	エネルギー，その他，**ヒト**，安全，医薬品，運動，衛生用品，**患者属性**，**患部**，関係（文法），文法用語，軌跡，形態，経過，**検査**，指標，自然，**手技**，**体外物質**，**体内物質**，妊娠，熱，**病因**，**病原体**，**病態**，**病名**，分析，方向，方法，法規
削除：	温度，音，奇形，教育，空間，計算，原因，固有名，使用，治療，疾患，障害，人格，**人名**，数値，性，接触，戦争，知覚，**地名**，超過，通過，電気，認知，波動，分泌物，並列，保健衛生

表 7 は手順 4 の見直しの中で変更となった意味ラベルの一覧である．「病因」や「患部」など，一般的には馴染みのないラベルがある.

「病因」が付与された語構成要素に「寒冷」「欠失」「中毒」「爆発」などがある．これらは語構成要素本来の意味による意味ラベルとはなっておらず，93 種類の意味ラベルには，語構成要素自身の意味を表すものと，合成語の中での関わりを表すものが混在している.

一方，削除された意味ラベルに「地名」と「人名」がある．「ロシア出血熱」「クーリー貧血」などは，地名や人名を除くと語の持つ概念が変わってしまうため分割しないこととした．医療記録に記載される情報を知る上で重要な診療の流れと記載される情報の概要を以下に示す．下線は，臨床看護および臨床医の経験者による見直しの中で新たに付与された意味ラベルである．「症状」「病態」などは，語そのものの意味を表すのではなく，患者または疾患に対する意味的な関わりを表している.

診療の流れと情報

- **受診**
  - 患者属性
  - 症状
  - 患部（身体部位・位置）
  - 時間・頻度
- **検診・診断**
  - 検査名
  - 指標
  - 状態（状態・病態・症状）
  - 患部（身体部位）
  - 病名
  - 病因
- **治療**
  - 医療行為
  - 患部（身体部位・位置）
  - 状態（状態・病態・症状）
  - 手段（機器・手技・薬品）
  - 時間・頻度
- **経過観察**
  - 患部（身体部位・位置）
  - 時間

表 8 意味ラベル列数（異なり）

	意味ラベル列数	計
手順 4 (長)	986	7,194
手順 4 (短)	1,548	7,194
手順 2	1,380	7,194

表 8 は，合成語に付与された意味ラベル列の数をまとめたものである．医療の観点からの見直しにより複数の意味ラベルを持つ語構成要素が増えたことで（表 6），「短い単位」での意味ラベル列の数が多くなっている．一方，複数の語構成要素をまとめて一つの語構成要素とする「長い単位」の語構成要素に付与する意味ラベルは，右側主要部の規則により最右側の語構成要素に付与された意味ラベルとしていることから，意味ラベ

ル列の数は少なくなる[17].

表 9 は, 高頻度の意味ラベル列をまとめたものである. 「身体部位｜病名」の頻度が最も高くなっていることがわかる. なお, 手順 2 では「病名」ではなく「疾患」が付与されている.

表 9　高頻度の意味ラベル列

順位	手順 4（長）	度数	相対度数	手順 4（短）	度数	相対度数	手順 2	度数	相対度数
1	身体部位｜病名	1092	15.2%	身体部位｜病名	910	12.6%	身体部位｜疾患	972	13.5%
2	身体部位｜状態	347	4.8%	身体部位｜状態	285	4.0%	状態｜疾患	633	8.8%
3	病因｜病名	308	4.3%	身体部位｜身体部位｜病名	168	2.3%	状態｜身体部位｜疾患	293	4.1%
4	病名	262	3.6%	病名	154	2.1%	身体部位｜身体部位｜疾患	196	2.7%
5	状態	169	2.3%	病因｜病名	154	2.1%	身体部位｜状態｜疾患	168	2.3%
	総計	7,194		総計	7,194	100%	総計	7,194	100%

表 10　合成語における語構成要素の位置

	語頭		語末		語頭＆語末		語中		計
	要素数	割合	要素数	割合	要素数	割合	要素数	割合	
（長）	3,515	69.2%	2,366	46.6%	1,008	19.8%	206	4.1%	5,079
（短）	2,867	66.3%	1,658	38.4%	687	15.9%	485	11.2%	4,323
統合	**3,672**	**63.5%**	**2,818**	**48.7%**	**1,436**	**24.8%**	**733**	**12.7%**	**5,787**

表 11　手順 4 の意味ラベル列における位置別の意味ラベルの頻度（上位 5 位）

順位	語頭	度数	相対度数	累積相対度数	語末	度数	相対度数	累積相対度数	語中	度数	相対度数	累積相対度数
1	身体部位	3,207	38.1%	38.1%	病名	3,339	35.9%	35.9%	身体部位	1,230	36.3%	36.3%
2	病因	1201	14.3%	52.4%	状態	1,468	15.8%	51.7%	状態	429	12.6%	48.9%
3	状態	663	7.9%	60.3%	医療行為	820	8.8%	60.5%	病態	288	8.5%	57.4%
4	病態	512	6.1%	66.4%	症状	731	7.9%	68.4%	病因	228	6.7%	64.1%
5	経過	354	4.2%	70.6%	病名・病態	689	7.4%	75.8%	病名	127	3.7%	67.8%
総計		8,414	100.0%		総計	9,295	100.0%		総計	3,393	100.0%	
付与ラベル数		107				100				81		

表 12　手順 4 の語構成要素列における位置別の語構成要素の頻度（上位 5 位）

順位	語頭	度数	相対度数	累積相対度数	語末	度数	相対度数	累積相対度数	語中	度数	相対度数	累積相対度数
1	先天性	280	3.2%	3.2%	損傷	557	6.0%	6.0%	良性	155	3.6%	3.6%
2	急性	157	1.8%	4.9%	腫瘍	431	4.6%	10.6%	悪性	120	2.8%	6.4%
3	結核性	119	1.3%	6.3%	手術	270	2.9%	13.5%	多発	103	2.4%	8.8%
4	新生児	111	1.2%	7.5%	出血	199	2.1%	15.6%	多発性	67	1.6%	10.4%
5	外傷性	106	1.2%	8.7%	障害	168	1.8%	17.4%	先天性	48	1.1%	11.5%
6	遺伝性	78	0.9%	9.6%	挫傷	155	1.7%	19.1%	神経	46	1.1%	12.5%
7	一過性	77	0.9%	10.4%	麻痺	153	1.6%	20.8%	欠乏性	40	0.9%	13.5%
8	耳介	74	0.8%	11.3%	中毒	148	1.6%	22.3%	動脈	34	0.8%	14.3%
9	後天性	72	0.8%	12.1%	後遺症	118	1.3%	23.6%	脳	33	0.8%	15.0%
10	多発性	56	0.6%	12.7%	狭窄症	116	1.2%	24.9%	腫瘍	32	0.7%	15.8%
総計		8,881			総計	9,313			総計	4,297		
付与ラベル数		2,490				2,537				1,150		

5.3 合成語における語構成要素の位置

合成語 7,194 語における語構成要素 5,787 語の位置をまとめたものが表 10 である.

表 3 より, 合成語 1 語当たりの語構成要素数の最頻値が 2 であることから, 語頭および語末にくる

語構成要素の割合が高くなっている.

表 11 は，手順 4 における「長い単位」と「短い単位」の意味ラベル列の異なり数 9,295 列において，「語頭」「語中」「語末」ごとに意味ラベルの頻度を調べた結果である. 語頭および語中で最も頻度の高いものは「身体部位」で相対度数は 38.1%と 36.3%，語末では「病名」で 35.9%である.

一方表 12 は手順 4 における「長い単位」と「短い単位」の語構成要素列の異なり数 9,313 列における語構成要素の頻度を「語頭」「語中」「語末」ごとに調べた結果である. 最も頻度が高い語構成要素は, 語頭では「先天性」で, 語末では「損傷」, 語中では「良性」であった.

6. 当面の課題

ここでは現在明らかになっている課題について述べる.

6.1 語構成要素の抽出と認定単位

用語の意味が理解できなければ, 医療記録文から用語を抽出することは困難である. しかし日本語文法の知識を持ち，かつ，実践医療用語の意味を理解する研究者を見つけることは困難である. そこで, 認定単位を定めずに医療的な意味により合成語の分割を行ってきた. 実践医療用語の分析においても斉藤（1967）の 4 種の認定単位が利用可能である. 問題は，複数の単位語および特殊単位語からなる合成語のどこまでを一つの語構成要素とするかということである.

複数の同じ意味ラベルを持つ合成語の場合, 合成語の意味を決定するためには, 同じ意味ラベルを持つ語構成要素間の関係が必要になる. 例えば,「肋間｜動静脈｜損傷」では，連続する「肋間」と「動静脈」の関係は等位ではなく，「肋間動脈と肋間静脈」を表している. そこで現在は統語関係を記述するのではなく, これらを一つの語構成要素（中間的合成語）として意味ラベルを付与し, 長い単位の語構成要素列と意味ラベル列を作成している.

例：肋間動静脈損傷
要素列（長）：肋間動静脈｜損傷
ラベル列（長）：身体部位｜病名
要素列（短）：肋間｜動静脈｜損傷
ラベル列（短）：身体部位｜身体部位｜病名

このように, ①一般的な意味からなる単位で分割した語構成要素列を, ②医療の観点から見直して「短い単位」の語構成要素列を作成し，③更に統語的に妥当な単位，または，慣用的に一つの用語として使われる単位でまとめた「長い単位」の語構成要素列を作成している.

従って, まとめ上げは分析者の専門的な知識に依存し，「短い単位」に分割する客観的規則も，「短い単位」から「長い単位」にまとめ上げる客観的規則も明確ではない.

例えば人名を含む合成語「フリードライヒ運動失調症」では，フリードライヒ（人名）を切り離すことができないため 1 語となり，「長い単位」と「短い単位」の語構成要素列は一致する.

要素列（短）：フリードライヒ運動失調症
要素列（長）：フリードライヒ運動失調症

しかし，「毛細血管拡張性運動失調症」では,

要素列（短）：毛細血管｜拡張性｜運動失調症
要素列（長）：毛細血管拡張性｜運動失調症

となる.

合成語 7,194 語において，「長い単位」と「短い単位」の語構成要素列が一致する語は 5,073 語（70.5%）となっている（表 5）.

そこで長短 2 種類の語構成要素列より, 単位の認定規則を明確にする予定である.

6.2 意味分類と意味ラベル

「病態」や「病因」など，一般には馴染みのない意味ラベルが導入され, 語構成の分析および意味分類において, より臨床経験と専門的な知識が必要となった.

例えば, 手順 2 で使っていた意味ラベル「物質」を「体内物質」と「体外物質」に分け，「下血」に意味ラベル「体外物質」を付与したところ，共同研究者より，「血液は体内物質であり，下血も体内物質となるのではないか」との疑問が出た. 一方, 臨床経験のある共同研究者からは,「下血」そのものよりも「下血」したことに意味があるので「病態」または「状態」に意味ラベルを変更したいとの意見が出た.

治療の観点からは「病態」であっても，検査の観点からは「検体」となり，意味ラベルは「体外物質」となる．劉（2000）の指摘にあるように同じ医学でも意味の解釈が異なる場合があり, 結合する他の語構成要素によっても意味が変化すると考えられる．その結果，意味ラベルの付与により，重要な情報が失われることも否めない.

また，前述のように，93 種類の意味ラベルには，語構成要素自身の意味を表すものと，合成語の中での関わりを表すものが混在しており, これらの体系化は困難である.

7. 語構成要素語彙試案表

今回，合成語 7,194 語から語構成要素 5,787 要素を抽出し，それぞれに意味ラベルを付与した. これらは, 合成語の抽出および意味の類推に活用することができる．そこで，語構成要素と，品詞情報，ヨミガナ，意味ラベル，そして，合成語の中での位置情報をまとめ，本年度末に『語構成要

素語彙試案表』として公開する予定である.

語構成要素もまた斉藤（1967）が規定した単位語と特殊単位語からなる合成語であり, 単位語および特殊単位語は手順 2 の結果を利活用する.

そして, 語構成要素を単位語および特殊単位語で分割した語構成要素列を作成し, これらを順次『語構成要素語彙試案表』に追加・更新して, 本研究の第一の目標『実践医療用語語彙表』の作成に繋げたいと考えている.

8. おわりに

本稿では, 合成語 7,194 語を対象に実施した語構成解析について述べた.

近年, 統計的機械学習を用いた自然言語処理の実用化が進んでいる. 深層学習では, 文字をコード化して処理されるものもある. 一方, 深層学習には, 大量のデータを要し, 低頻度のデータが上手く学習されないという問題がある. また, 統計処理で使われる平均値も標準偏差も, 集団の特徴を表す代表値であり, 個々の特徴を示すものではない.

表 13 低頻度の語構成要素の割合

度数	手順 4(長)		手順 4(短)		手順 2	
1	3,460	23.7%	2,635	15.5%	2,337	13.0%
2	661	4.5%	569	3.3%	598	3.3%
3	287	2.0%	280	1.6%	259	1.4%
計	14,622	100.0%	17,009	100.0%	17,928	100.0%

通常, 小説などの言語作品では, 頻度 1 の語数が最も多くなる. 本研究においても, 頻度 1 の語構成要素の相対度数は, 「長い単位」で 23.7%, 「短い単位」で 15.5%と低頻度の割合は低くはない（表 12）.

本研究では, 語構成要素と付与された意味ラベルを基にした言い換えの研究も行っており, これらのデータは, 分散表現などでの利用が考えられる[18].

我々は, 本研究成果が人の健康に関わる個々の患者の診療に利活用されるということを心に留め, 病状説明の根拠となる情報を失わないように, 文字ではなく意味を持つ語に拘り, 医療の観点から妥当な意味ラベルを付与した『語構成要素語彙試案表』の作成と公開に努めたい.

謝辞

本研究は, 科学研究費補助金「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」（JP18H03499）の助成を受けています.

参考文献

[1] 相良かおる,小野正子.実践医療用語辞書 ComeJisyoSjis-1 の作成.言語処理学会第 25 回年次大会発表論文集,2019,p.1491-1494.
[2] 開原成允. 医学用語の現状と課題. 日本語学,2010,vol.29-15,p.14-24.
[3] 平陽子,植田啓子,有竹由紀子,山田一朗.医療用語・書式の病院間較差に関する研究 第 2 報:－全国調査の集計結果－.日本看護研究学会雑誌.2009,32(3),p.375.
[4] 相良かおる. ComeJisyo の紹介と医療情報に含まれる誤字調査. 情報知識学会誌.2014, 24(2), p.204-209.
[5] 斉藤孝. 索引作業のための自然言語処理の研究―医学用語の計量的調査―. Library Science.1967, No.5, p.51-71.
[6] 東条佳奈, 麻子軒, 相良かおる, 高崎智子, 山崎誠.病名における「-性」の分析―一般書籍との比較から―, 言語資源活用ワークショップ 2020.
[7] 劉亜斌, 里村洋一, 佐々木哲明, 木村通男, 廣瀬康行, 山﨑俊司. 構造化臨床医学用語集の構築に関する研究.2000,20(6),p.513-522.
[8] 「疾病, 傷害及び死因の統計分類」 https://www.mhlw.go.jp/toukei/sippei/ (参照 2020-10-10)
[9] “SNOMED International” http://www.snomed.org/ (参照 2020-10-10)
[10] 相良 かおる,小野 正子,上野 恵子.医療用語のシソーラス作成にむけた予備的調査. 西南女学院大学紀要, 2015,Vol.19,p.109-118.
[11] 相良 かおる,小野 正子,山崎 誠.末尾語にサ変接続名詞を持つ実践医療用語の語彙分類. 人文科学とコンピュータシンポジウム論文集.2016,No.2,p.183-190.
[12] 小山照夫,大江和彦.医学専門用語の構造解析.学術情報センター紀要.1994,(6), p.115-124.
[13] 国立国語研究所.分類語彙表 増補改訂版. 大日本図書, 2004.
[14] MeCab： https://taku910.github.io/mecab/ (参照 2020-10-16)
[15] UniDic：https://unidic.ninjal.ac.jp/ (参照 2020-10-16)
[16] 石井正彦. 現代日本語の複合語形成論. ひつじ書房,2007.
[17] 相良 かおる, 高崎 智子, 東条 佳奈, 麻 子, 山崎 誠.病名を表す合成語の語末調査. 言語資源活用ワークショップ 2020.
[18] 統計的自然言語処理―ことばを扱う機械. 岩波データサイエンス刊行委員会編.岩波データサイエンス.2016,Vol.2