ポスター | 知識工学／医療データ解析
ポスター1
知識工学／医療データ解析
2019年11月22日(金) 14:50 〜 16:20 ポスター会場2 (国際展示場 展示ホール8)

[2-P2-1-08] 実践医療用語の語構造に関する考察 ー医療記録に含まれる合成語の妥当な細分割を目指してー

○相良 かおる[1]、小野 正子[1]、山崎 誠[2] （1. 西南女学院大学, 2. 国立国語研究所）

キーワード：Medical compound words, Word structure analysis, Word component, Medical machine-readable documents

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目標を掲げた。

今後、医療記録データの自然言語処理を支援する辞書が必要になると考えた筆者等は、2004年より看護実践用語の収集を開始し、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、2018年11月には登録語数75,831語の UTF版の ComeJisyoUtf8-1を、2019年4月には、医師経過記録から抽出した用語を含む登録語数111,664語の Shift_JIS版の ComeJisyoSjis-1を公開している。

当初、実践医療用語の実態が不明であり、語の単位認定が困難なことから、臨床経験を持つ看護師、管理栄養士、医師等が一つのまとまった語とした語を登録している。

その結果、本辞書には「末梢性神経血管性機能障害リスク状態」等の合成語が多く登録されている。これを国立国語研究所の形態素解析辞書 UniDicにより短単位（＝形態素に相当）に分割すると、「末梢｜性｜神経｜血管｜性｜機能｜障害｜リスク｜状態」と分割される。

このように、本辞書の約11万の登録語には、多様な語種、そして複合語や臨時一語等の合成語が多く含まれる。

筆者等は医療記録文書を解析するための網羅性の高い辞書の構築が困難であることから、登録語の語構造の解析に着手し、その結果を踏まえ、長単位の合成語「末梢性神経血管性機能障害リスク状態」を「末梢性｜神経血管性｜機能障害｜リスク状態」のように意味的にまとまりのある小さな単位（中単位）に分ち書きする辞書の構築を目指している。

本発表では、一般的な単語を含む合成語2000語の語構造と、中単位に分割するための辞書の構築について述べる。

実践医療用語の語構造に関する考察

―医療記録に含まれる合成語の妥当な細分割を目指して―

相良かおる*1、小野正子*1、
山崎誠*2
*1 西南女学院大学、*2 国立国語研究所、

Word component decomposition and semantic composition analysis of medical compound words

Kaoru Sagara*1 , Masako Ono*1 , Makoto Yamazaki*2
*1 Seinan Jo Gakuin University , *2 National Institute for Japanese Language and Linguistics,

With the aim of creating a dictionary to divide medical machine-readable documents into single words that are semantically appropriate, we decomposed 2,771 compound words contained in medical documents into 7,046 word components. These were then classified into 102 categories according to their meaning and were assigned semantic labels. The average number of word components per compound word was 2.54, and the maximum number of word components was 6. The percentage of defined labels with word components was 27% for "condition", 21% for "disease", and 20% for "body part". In order to confirm whether the compound word and the divided word component were semantically appropriate for use as word units in medical documents, the results were compared with 114,883 headwords of four Medical Information Center Development Center (MEDIS-DC) standard master types. The results showed that 94% of the medical compound words and 73% of the word components matched the headwords. The standard masters used were (1) disease name, (2) surgery/treatment, (3) nursing observation, and (4) symptom/finding.

Keywords: Medical compound words, Word structure analysis, Word component

1.はじめに

医療記録に含まれる語（以下、「実践医療用語」という）を登録語とし、2008 年より無償公開を開始した実践医療用語辞書 ComeJisyo は、随時更新を続け、2018 年 11 月には研究・教育分野での利用を目的とした登録語数 75,831 語の UTF 版の ComeJisyoUtf8-1 を公開し、2019 年 4 月には、医師経過記録から抽出した語を含む登録語数 111,664 語の Shift_JIS 版の ComeJisyoSjis-1 を公開している [1]。

本辞書の作成に着手した 2004 年当時、標準化された日本語の医療用語は公開されておらず、また実践医療用語の語構成や語種構成の実態は不明であった。そこで、語の単位認定の規則を定めず、看護師、管理栄養士、医師としての臨床経験を持つ者が一つのまとまった語とした語を登録した結果、本辞書には「末梢性神経血管性機能障害リスク状態」等の合成語が多くある。

従って、多様な語種、複合語や助詞を省略した臨時一語等の合成語を含む本辞書は、実践医療用語の語彙研究の言語資源となった。

そこで、筆者らは、本辞書の合成語を対象に語構成の分析に着手している [2)3)]。具体的には、これらの合成語を構成する意味的な最少の単位（以下、「語構成要素」という）を抽出し、人間可読の語彙表を作成し、次いで医療記録を語構成要素の単位で分かち書きする機械可読の辞書を作成する予定である。

本発表では、2,771 語の合成語より分割された語構成要素とこれらに付与された意味ラベルの分析結果について述べる。

2.先行研究

専門用語の語構成に関する先行研究として野村・石井(1988)の『学術用語語基連接表』（以下、『連接表』という）[4)]がある。

『連接表』には、1984 年時点における文部省編『学術用語集』23 分野の用語がおさめられ、医療分野の専門用語は含まれていない。なお、2018 年時点で制定された『学術用語集』は 32 分野であり、医学編は 2003 年に初版が刊行され、2014 年に絶版となっている [5)]。

『連接表』の解説には、合成語を語構成要素に分解し、その造語機能を分析するための単位として、語において実質的な意味を担う「語基」を採用していること、その結果、語基認定の問題点として、語基の認定に不統一があること、その原因の一つとして、分析者の専門外の「学術用語」を分析対象としていることが述べられている。

そこで、筆者らを含む本語構成分析に携わる分析者 5 名の専門が、情報科学（1名）、日本語学（3 名）、看護学（1 名）であることから、分析の単位として「語基」を採用せず、医療用語の「意味」を基準とした最少単位を「語構成要素」と定義した。この「意味」の基準については石井（2007：182）の複合名詞の語構造把握のための意味分類 22 種を参考にした [2)3)6)]。

3.方法

語構成の分析対象は、ComeJisyoSjis-1 の登録語 111,664 語より抽出した 7,139 語とした。以下に簡単な抽出手順を示す。

① 方言や医療施設特有の語の排除（汎用性の確保）：Web 上で公開されている辞書等，研究用に収集した医療用語データと本辞書の登録語を照合し、一致した 31,162 語を抽出

② 専門外の分析者による意味による分割の不統一の低減：①の内、『分類語彙表増補改訂版』（以下，『分類

語彙表』)に収録されている語(一般的な語)を含む合成語約 7,139 語(以下、「対象合成語」)を抽出

3.1 語構成要素への分割と意味ラベル

(1) 予備的調査として先ず、対象合成語 7,139 語の内、分析者らが興味を持った『分類語彙表』と一致する語、「先天性(396 語)」「多発(244 語)」「～炎(197 語)」「移植(66 語)」「呼吸(33 語)」「依存(25 語)」等を含む 1,100 語(重複除く)について表 1 に示す石井(2007:182)の 22 種の意味分類(以下、「参考意味分類」という)を基準に語構成要素に分割し、該当する意味分類がない場合は、各自で新たな意味ラベルを付与した。

表 1 石井(2007:182)での意味分類 [6)]

意味ラベル	語 例
自然物	石, 土, 水, ガス, 電気, …
動植物	芋, い, わら, 果実, …
物品	物, 荷, 本, 新聞, 図書, …
食品	パン, 菓子, チーズ, 茶, …
道具	器, のこ, きり, ハンマ, …
薬品	剤, くすり, ワニス, …
力	水圧, 水力, …
人間	人, 者, 工, …
機械	機, 機械, 装置, プレス, …
衣料	布, 絹, 服, 糸, 織物, …
部分	手, 足, 歯, つば, 羽根, …
家具	家具, いす, 洗面台, …
資材	板, 管, 金, 紙, 油, ねじ, …
地類	水路, 橋, 堤, 港, …
容器	コップ, なべ, びん, 皿, …
建物	屋根, むね, はり, 柱, …
空間	場, 場所, 上, 内, 奥, …
形状	みぞ, つや, 曲線, 穴, 口, …
数量	二重, 三つ, 半, 距離, …
動き	連続, 自動, 回転, …
状態	平, 深, 薄, 速, ばら, …
時間	熱間, 冷間, 負荷時, …

(2) 分割した語構成要素について分割方法および、付与した意味ラベルについて分析者 5 名全員で見直し、統一の意味ラベル表を作成し、分析結果を修正した。

(3) 次に対象合成語より乱数を用いてランダムに抽出した合成語 1,000 語について分析者 5 名で分担して語構成要素に分割し、意味ラベルを付与した。[3)]

(4) 筆者ら 2 名(情報科学・看護学)により、これら 1,994 語(重複 106 語削除)に付与された意味ラベルを見直し、意味ラベル表を修正しながら、新たに 777 語について語構成分析を行った。

(5) 合成語 2,771 語(1,994 語＋777 語)から得られた語構成要素について、修正した意味ラベル表を基に意味ラベルを付与し、分析を行った。

3.2 標準マスターとの照合

①本合成語より抽出した語構成要素、②合成語 2,771 語(以下、「本合成語」という)、③対象合成語 7,139 語、④ComeJisyoSjis-1 の登録語 111,664 語、およびについて、以下の 4 種類の標準マスター[7)]の見出し語(114,883 語)と照合し、一致する割合を求めた。

① 病名マスターの索引 medis_ICD10V405 (103,697 語)
② 手術・処置マスター medis_prcdr20180911 (7,396 語)
③ 看護観察マスター medis_kansatsuVer3.3_20190815 (2,370 語)
④ 症状・所見マスター medis_phyxam-beta_20140306 (1,420 語)

4. 結果と考察

4.1 語構成要素

本合成語 2,771 語を分割して得られた語構成要素の総数(延べ)は 7,046 語、種類(異なり)は 2,070 語、本合成語 1 語あたりの語構成要素数の平均は、2.54 語であった。

表 2 に本合成語 1 語あたりの語構成要素数を示す。

今回、「非、不、性」等の漢字一字の接辞等を独立させずに分割したことで、短単位に分割すると「非｜糖尿｜病性」に分割される「非糖尿病性」は 1 語構成要素と長い単位となった。その結果、本合成語の 8 割以上の語構成要素数は、2 または 3 語となった。

表 2 合成語 1 語あたりの語構成要素数

語構成要素数	本合成語	
1	192	7%
2	1,251	45%
3	1,020	37%
4	258	9%
5	41	1%
6	9	0%
計	2,771	100%

本合成語の語構成要素数の最大値は 6 語であった。また、1つの合成語内に「～性」の付く語構成要素が複数あるものが 9 語あった。以下に、語構成要素数 6 で、かつ「～性」の語構成要素を複数含む合成語を示す。

「慢性｜びまん性｜メサンギウム｜増殖性｜糸球体｜腎炎」
「慢性｜びまん性｜半月体｜形成性｜糸球体｜腎炎」

本合成語の中に「兼」を含む臨時一語「先天性｜僧帽弁｜狭窄｜**兼**｜閉鎖｜不全症」が見つかった。「兼」は医療記録に出現する記号「・」や「＆」のような働きを持つ文法的な機能語である。なお本辞書には、「・」を含む「健康保持・増進」が登録されている。

表 3 語構成要素(頻度 30 以上)

語構成要素	頻度	語構成要素	頻度
先天性	437	骨折	48
多発性	170	狭窄症	46
麻痺	152	神経	45
多発	95	異常	44
損傷	80	癒着	44
貧血	77	髄膜炎	42
腫瘍	72	不全	42
皮膚炎	61	分娩	41
障害	60	脱臼	35
急性	59	手術	34
慢性	59	腎炎	34
移植	51	肥大	33
		出血	32

表 4 意味ラベルの出現位置（頻度 20 以上）

語 頭	頻度	割合	語 中	頻度	割合	語 末	頻度	割合
状態	1,258	49%	身体部位	646	38%	疾患	1,324	51%
身体部位	714	28%	状態	448	26%	症状	211	8%
時間	107	4%	数量	82	5%	状態	192	7%
物質	54	2%	疾患	70	4%	疾患・状態	185	7%
種類	48	2%	生理	35	2%	医療行為	120	5%
人間	45	2%	形状	30	2%	障害	61	2%
行為	28	1%	医療	30	2%	状態・変化・増減	59	2%
空間	28	1%	動き	28	2%	動き・状態	59	2%
薬品	26	1%	空間	27	2%	行為	52	2%
疾患	23	1%	物質	27	2%	身体部位	44	2%
			症状	26	2%	疾患・形状	34	1%
						動き	32	1%
						生理	31	1%
						精神	22	1%
						医療	20	1%
小計	2,331	90%		1,449	56%		2,446	95%
総計	2,579	100%		1,697	100%		2,579	100%

「兼」、「・」、「＆」等を独立した一つの語構成要素とするか否か、例えば「先天性｜僧帽弁｜狭窄｜兼｜閉鎖｜不全症」を 1 語の合成語（臨時一語）とするか、「先天性｜僧帽弁｜狭窄」と「閉鎖｜不全症」に分け、2 合成語とするかは、今後の検討課題である。

表 3 は、頻度 30 以上の 25 語の一覧である。

本合成語 2,771 語には、分析者が興味を持った「先天性（396 語）」「多発（244 語）」を含む合成語が含まれ、「先天性」と「多発性」の頻度が高くなっている。

なお、山崎（2019）の乱数を用いて選定した合成語 1,000 語における頻度では「先天性」は 4 位、「多発性」は 10 位となっている。これら 25 語の内、岩波国語辞典第 5 版 [9]の見出し語と一致するものは 20 語、見出し語に無い語は、「先天性」「多発性」「皮膚炎」「狭窄症」「髄膜炎」の 5 語であるが、これらから造語成分「性」、「炎」、「症」を除いた「先天」、「多発」、「皮膚」、「狭窄」、「髄膜」は全て岩波国語辞典第 5 版の見出し語にある。従って、頻度 30 以上の語構成要素の専門度は比較的低いと考えられる。

4.2 意味ラベル

表 4 は、語構成要素数が1の合成語 192 語を除いた 2,579 語の語構成要素に付与された意味ラベルの出現位置を「語頭」、「語中」、「語末」毎に集計し、頻度 20 以上をまとめたものである。「語頭」に位置する語構成要素の半数 49%は「状態」であった。一方「語末」に位置する語の半数 51%は「疾患」であり、「語中」の 38%は「身体部位」であった。

なお、「疾患」を表す意味ラベルには、「疾患（肺炎）」、「疾患・状態（麻痺）」、「疾患・形状（腫瘤）」など、意味ラベルを 1 つに定められず、複数の意味ラベルを付与したものがある。

表 5 は頻度 50 以上の意味ラベル 20 種の一覧である。「状態（27%）」、「疾患（21%）」、「身体部位（20%）」で全体の約 7 割を占めていた。参考意味分類（表 1）になく、新たに意味ラベルを設けたものが 13 ラベル（表 5 の※印）あった。

ランダムに抽出した合成語 1,000 語の語構成要素に付与された意味ラベルは 70 種類 [3]、今回、本合成語 2,771 語の語構成要素に付与された意味ラベルは 102 種類であった。この内、参考意味分類と一致する意味ラベルは共に 13 種類、意味ラベル 102 中、89 種類は新たに設けたラベルであり、これらには、前述の「疾患」を表す意味ラベルの他、「動き・状態（穿孔）」、「状態・変化・増減（肥大）」等の複数の意味ラベルを付与したものが 34 種類あった。

表 5 意味ラベル（頻度 50 以上）

	意味ラベル		頻度	
1	状態		1,923	27%
2	疾患	※	1,462	21%
3	身体部位	※	1,421	20%
4	症状	※	262	4%
5	疾患・状態	※	200	3%
6	医療行為	※	154	2%
7	時間		147	2%
8	数量		99	1%
9	行為	※	94	1%
10	物質	※	92	1%
11	生理	※	87	1%
12	動き		74	1%
13	動き・状態	※	70	1%
14	種類	※	69	1%
15	形状		68	1%
16	人間		64	1%
17	障害	※	62	1%
18	空間		59	1%
19	状態・変化・増減	※	59	1%
20	医療	※	51	1%
	小計		6,517	93%
	総計		7,046	100%

4.3 意味ラベルのパターン

本合成語 2,771 語の意味ラベルのパターンは 664 種類あった。表 6 は頻度が 20 以上の意味ラベルのパターンをまとめたものである。語末が「疾患」のパターンが 8 種類と最も多くなっており、これら 8 種類を併せると 34%であった。「状態」を複数含む語構成要素、例えば、「状態｜身体部位｜状態」の多く

は、「先天性｜呼吸器｜異常」のように、「～性」が付く語構成要素が語頭に、「性」の付かない語構成要素が語末に来るパターンが多く、これは、「～性」の意味ラベルは全て「状態」としたことに起因している。「身体部位｜身体部位｜疾患」には「手指｜血管｜損傷」等があった。

表 6 頻度 30 以上の意味ラベルパターン

	意味ラベルパターン	頻度	
1	状態｜疾患	339	12%
2	身体部位｜疾患	185	7%
3	状態｜身体部位｜疾患	155	6%
4	状態｜状態｜疾患	81	3%
5	状態｜身体部位｜状態	79	3%
6	状態｜症状	63	2%
7	身体部位｜疾患・状態	49	2%
8	疾患	45	2%
9	身体部位｜状態・変化・増減	38	1%
10	身体部位｜身体部位｜疾患	36	1%
11	身体部位｜数量｜疾患	36	1%
12	身体部位｜状態｜疾患	34	1%
13	状態｜疾患・状態	29	1%
14	身体部位｜動き・状態	29	1%
15	身体部位｜医療行為	27	1%
16	状態	25	1%
17	状態｜状態｜症状	23	1%
18	時間｜疾患	20	1%
	小計	1,293	47%
	総計	2,771	100%

4.4 標準マスターの割合

表 7 は、標準マスターの見出し語 114,883 語と照合した結果である。本辞書の登録語から抽出した『分類語彙表』に収録されている語を含む対象合成語および本合成語共に約 9 割が病名マスターの見出し語と一致し、4 種の標準マスターと一致する合成語は 94%であった。また、本合成語を分割して得られた語構成要素では 73%が一致した。

表 7 標準マスターとの照合

		語構成要素	本合成語	対象合成語	ComeJisyo Sjis-1
①	病名マスター	4,595	2,467	6,208	23,785
	索引	65%	89%	87%	21%
②	手術・処理	245	116	426	3,912
	マスター	3%	4%	6%	4%
③	看護観察	260	14	39	927
	マスター	4%	1%	1%	1%
④	状・所見	10	1	7	480
	マスター	0%	0%	0%	0%
⑤	不一致	1,935	173	459	82,560
		27%	6%	6%	74%
	計	7,046	2,771	7,139	111,664
		100%	100%	100%	100%

5.まとめ

医療記録に含まれる合成語を構成する語構成要素、すなわち、「意味」を基準とした最少の単位の語への分割規則が定まれば、語構成要素に分かち書きするための機械可読の辞書の作成が可能となり、合成語の自動生成が可能となる。

また、語構成要素の意味ラベルの付与規則が定まれば、その結合パターンから合成語の意味の推測が可能となり、医療記録の要約や言い換え等に応用できる。さらに、医療情報学の分野だけでなく、日本語学分野における実践医療用語の語彙研究の対象とも成り得る。

そこで筆者らはこれらの解明に着手し、本合成語 2,771 語より 2,070 語の語構成要素を抽出した。また分析の過程で、以下のことが分かっている。

(1) 「神経(身体部位)｜麻痺(疾患・状態)」と「脳性(状態)｜麻痺(疾患・状態)」はそれぞれ 2 要素となるが、本合成語に含まれない同義語「神経性(状態)｜麻痺(疾患・状態)」と「脳(身体部位)｜麻痺(疾患・状態)」では、異なる意味ラベルが付与される。

(2) また 1 要素とした「球麻痺｜(疾患)」と「片麻痺(疾患)」の同義語「延髄(身体部位)｜麻痺(疾患・状態)」と「片側(空間)｜麻痺(疾患・状態)」は、2 要素となり当然、付与される意味ラベルも異なる。

(3) 「肥大(状態・変化・増減)」のように複数の意味ラベルが付与される場合がある。

(4) 「麻痺(疾患・状態)」と「片麻痺(疾患)」等、意味ラベルの不統一問題、そして「疾患」ではなく「症状」ではないかという妥当性の問題

方言、略語、業界用語を含む合成語を、実用性の高い適切な「意味」を基準とした最少の単位に分割し、意味ラベルを付与することは困難である。また、分割した要素に付与する意味ラベルの命名も容易ではない。

本合成語の分析結果について十分に検討・議論し、対象合成語 7,139 語より「語構成要素」を求め、実用性の高い人間可読の語構成要素の語彙表を作成し公開する予定である。

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，p.1491-1494，2019.
2) 東条佳奈，相良かおる，小野正子，山崎誠：「実践医療用語における構成要素の意味分類試案―「先天性」を例に―」，『現代日本語研究』，11，pp.40-58，大阪大学大学院文学研究科日本語学講座現代日本語学研究室.
3) 山崎誠，相良かおる，小野正子，東条佳奈，麻 子軒：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，国立国語研究所，言語資源活用ワークショップ 2019，p.161-168.
4) 野村雅昭，石井正彦：学術用語語基連接表,国立国語研究所，1988.
5) 文部科学省 研究振興局 学術研究助成 課：学術用語の標準化について，2016.
[http://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu4/siryo/_icsFiles/afieldfile/2016/08/31/1376140_005.pdf (参照 2019-08-23)]
6) 石井正彦：『現代日本語の複合語形成論』，ひつじ書房，2007.
7) 医療情報システム開発センター(MEDIS-DC)：MEDIS 標準マスター・インデックス(2019 年 7 月 20 日ダウンロード)
[https://www.medis.or.jp/4_hyojyun/medis-master/index.html]
8) 国立国語研究所：分類語彙表 増補改訂版，大日本図書，2004.
9) 言語資源協会：岩波国語辞典第五版タグ付きコーパス 2004