A Trial for Segmentation and Annotation of Semantic Label for the Practical Medical Terms

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

山崎 誠（国立国語研究所研究系言語変化研究領域）†
相良 かおる（西南女学院大学保健福祉学部）
小野 正子（西南女学院大学保健福祉学部）
東条 佳奈（目白大学社会学部）
麻 子軒（大阪大学大学院文学研究科）

A Trial for Segmentation and Annotation of Semantic Label for the Practical Medical Terms

Makoto Yamazaki (National Institute for Japanese Language and Linguistics)
Kaoru Sagara (Seinan Jo Gakuin University)
Masako Ono (Seinan Jo Gakuin University)
Kana Tojo (Mejiro University)
Ma Tzu-Hsuan (Osaka University)

要旨

本発表では，電子医療記録に含まれる実践医療用語の語構成を明らかにするために，独自に設計した語構成要素への分割とそれに対する意味ラベルの付与を行い，意味ラベルによる語構成のパターンを調査した。調査対象は，ComeJisyoSjis-1（111,664 語）から，『分類語彙表 増補改訂版』に収録されている語を含む約 7,000 語から抽出した 1,000 語である。これらを短単位よりやや長めの語構成要素に分割し，意味ラベルを付与した。意味ラベルは，石井（2007）の複合名詞の語構造把握のための意味分類を参考にしたが，実践医療用語のために独自に設けたものも多い。分析結果から，以下のような点が明らかになった。(1) 語構成要素数が 2 個と 3 個のものが全体の 8 割以上を占める。(2) 意味ラベルは，「疾患」「身体部位」「状態」「症状」「医療行為」「時間」「生理」の 7 つで全体の約 8 割を占める。(3) 意味ラベルは，語頭により多く出現するもの（「身体部位」「時間」）や語末により多く出現するもの（「医療行為」「症状」「障害」）などがあり，分布に偏りが見られる。

キーワード：実践医療用語，ComeJisyo，語構成要素，意味ラベル，複合名詞

１．はじめに

近年，医療の現場に電子カルテの導入が進んでいるが，個人情報を含む医療記録が語彙研究の言語資源として公開されることはないため，電子カルテで用いられている専門用語については，まだ十分な考察が行われていない。

そこで筆者らは，電子カルテに含まれる専門用語（以下，「実践医療用語」という）を登録した辞書 ComeJisyo の見出し語を対象として語構成の解析に着手した。

ComeJisyo[1]は形態素解析器 MeCab の辞書として利用可能で，かつ，読み仮名や実際の医療記録における文書頻度などを付加した人間も可読な辞書であり，2008 年より無償公開

† yamazaki [AT] ninjal.ac.jp

[1] https://ja.osdn.net/projects/comedic/

され，2019 年 4 月に公開の ComeJisyoSjis-1 の登録語数は 111,664 語である（相良・小野 2018）。これらは臨床看護の経験者らが「一つのまとまった語」と判断したものを見出し語としており，専門用語をはじめ，略語，外来語を含む複合語および助詞が省略された臨時一語なども含まれている。

従って，ComeJisyo の見出し語を対象とした語構成の分析により，実際の医療記録で使われる合成語の構造の解明に加えて，実践医療用語の合成語を成す語構成要素の抽出が可能となる。

本発表では，ComeJisyo の見出し語より抽出した 1,000 語を対象に実施した語構成要素への分割と意味ラベルの付与について報告する。

2. データ

本発表で用いるデータは，成果物である語構成要素を「語構成要素表試案（仮称）」の一部として公開することを想定し，方言や医療施設特有の語を含む合成語を排除するために，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一致した ComeJisyo の見出し語（約 30,000 語）を合成語の対象データとしている。また，筆者らの殆どが医療の知識を持たないことから，先ず初めに，対象データの内，『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 7,000 語を抽出し，語構成の分析を行うこととした。本発表の対象データは，これらからランダムに抽出した 1,000 語である。

以下に抽出手順の詳細を示す。

(1) 方言や医療施設特有の語を含む合成語の排除（汎用性の確保）：
公開予定の ComeJisyoSjis-1 の見出し語 111,664 語と研究用に収集した医療用語データ 52,974 語を照合し，一致する語 31,162 語（以下，「合成語データ」という）を抽出

(2) 『分類語彙表』に収録されている語（一般的な語）を含む合成語データの抽出：
① 「合成語データ」を MeCab0.996 と UniDic-cwj-2.2.0 により短単位に分割
② 「臨時一語」の認定要件（石井 2007）に該当しないもの 7,327 語を抽出
③ 更に，アルファベット，ひらがな，カタカナのみからなる語を削除した 3,728 語を抽出
④ 既に 2018 年 11 月公開の ComeJisyoUtf8-1（75,831 語）における出現頻度を求め，頻度の多いもの上位 25%となる 768 語を抽出
⑤ 『分類語彙表』の項目と照合し，一致した 231 語を抽出
⑥ 合成語データ（31,162 語）より⑤の 231 語を含む合成語データ 7,139 語を抽出

(3) 本調査データの抽出：
乱数を用いての 7,139 語より，本調査用の合成語データ 1,000 語を抽出

3. 語構成要素への分割

語構成要素としては，日本語コーパスの分析でよく用いられる短単位や長単位をそのまま利用するのではなく，医療従事者の直観に基づく独自の単位を設計することにした。この語構成要素は現在構築中であり，今回用いるものはあくまでも暫定的なものである。本来であれば，語構成要素に分割するための手順や規則を用意しなければならないが，それらはまだ整っていない。今後，分割作業を通して語構成要素および意味ラベルを整備していく予定である。実践医療用語の語構成要素の特徴としては，総体的に短単位より長いものとなっていることが指摘できる。その理由としては，接辞を独立した 1 構成要素としない，というルールがある。例えば，語構成要素「非中毒性」における漢字一字の接辞（「非，性」）などは，独立させないため，「非中毒性」全体が 1 構成要素となる。

分析した 1,000 語に用いられた語構成要素の総数（延べ）は 2,415 個，種類（異なり）は 1,146 個であった。そのうち，頻度 10 以上の 35 語を表 1 に示した。上位には，疾患，症状，状態を表すものが多く，身体部位は 3 つ（大腿骨，子宮，血管）のみであった。表 1 の

35 語の延べ語数は 647 語で，出現した全体語構成要素の約 26.8%を占める。

表 1　語構成要素（頻度 10 以上）

語構成要素	頻度
腫瘍	65
損傷	60
先天性	43
手術	28
障害	25
出血	21
挫傷	20
脱臼	20
大腿骨	19
多発性	19
中毒	18
皮膚炎	18

語構成要素	頻度
貧血	18
麻痺	17
感染症	16
狭窄症	16
異常	14
萎縮	14
熱傷	14
悪性	13
急性	13
結核	13
後遺症	13
骨折	13

語構成要素	頻度
慢性	13
一過性	11
子宮	11
中毒性	11
捻挫	11
運動	10
結核性	10
血管	10
術後	10
髄膜炎	10
慢性	10

4. 意味ラベルの付与

意味ラベル[2]は，石井（2007: 182）の複合名詞の語構造把握のための意味分類を参考にしたが，実践医療用語のための独自に設けたものも多い。例えば，「行為」のほかに「医療行為」という意味ラベルを設けた。「医療行為」は，「手術」「検診」などのもっぱら医療従事者が行う行為，「行為」は，医療従事者以外も行う一般的な行為，のように使い分けている。その一部は，東条他（2019：46-47）にも挙げているが，今回分析対象とした 1,000 語に現れた意味ラベルは図 1 に挙げた 70 種類であった[3]。

位置　医療行為　医療行為・作用　動き　動き・状態　動き・変化・増減　学術　感覚　機器　器具　機能　空間　空間・身体部位　経済　形状　形状・状態　検査法　行為　サービス・支援　時間　施設　疾患　疾患・奇形　疾患・形状　疾患・症状　疾患・状態　手段　種類　使用　障害　症状　症状・疾患　状態　状態・症状　状態・程度　状態・変化・増減　食品　身体部位　身体部位・空間　身体部位・部分　振動　心理　数量　数量・身体部位　精神　生理　生理・動き　増減　体液　建物　単複　超過　治療　通過　動植物　人間　熱　能力　排泄物　破裂　光　物質　物品　部分　分泌物　文法的機能　変化　本末　薬品　様相

図 1　意味ラベル一覧（50 音順）

[2] 東条他（2019）では「意味カテゴリー」という名称を用いているが，本稿では「意味ラベル」を用いる。

[3] 「動き・状態」のように，複数の語が並列されている意味ラベルは，(1) 実際の語においてそのうちどちらかになるということを意味する場合（したがって，この例で言えば，「動き」か「状態」かを特定すべきであるが，それは出来ていない）。(2)「動き・変化・増減」のように，階層性のある意味ラベルを列挙した場合，など複数の基準が混在している。この点については，今後整備していく必要がある。

これらの意味ラベルは，どの語構成要素がどの意味ラベルに対応するか，具体的な分類知識としてまとめる必要がある。また，語構成要素に意味ラベルを当てはめる際，概念の重複が起こらないよう，意味ラベル間の調整を行わなければならない。最終的にはこれらの意味ラベルは構造化（階層化）された形で提示することになろう。

表 2 に頻度の高かった意味ラベルを示した。ここでは，「疾患」がもっとも多く，「身体部位」がそれに次いでいる。表 1 の個々の語構成要素としては，身体部位に該当するものは上位には来てなかった。意味ラベル「身体部位」が付く語構成要素の種類が多かったことが分かる。

表 2　意味ラベル（頻度 10 以上）

意味ラベル	頻度
疾患	606
身体部位	576
状態	466
症状	94
医療行為	85
時間	54
生理	49
動き	45
物質	44
空間	30

意味ラベル	頻度
人間	29
行為	25
障害	25
動き・状態	24
疾患・状態	24
種類	21
状態・変化・増減	17
動植物	16
薬品	15
形状	12

意味ラベル	頻度
疾患・形状	12
数量	12
身体部位・部分	11
精神	10

5. 語構成要素数

図 2 は，1 語あたりの語構成要素数を示したものである。語構成要素数 2 の実践医療用語がいちばん多く，全体の 5 割強を占めることが分かる。また，語構成要素数が 2 と 3 とで全体の約 8 割以上を占めている。ちなみに，語構成要素数がいちばん多かったのは，「坐骨/恥骨/軟骨/結合/若年性/骨軟骨症」の 6 であった（「/」は語構成要素の切れ目）。

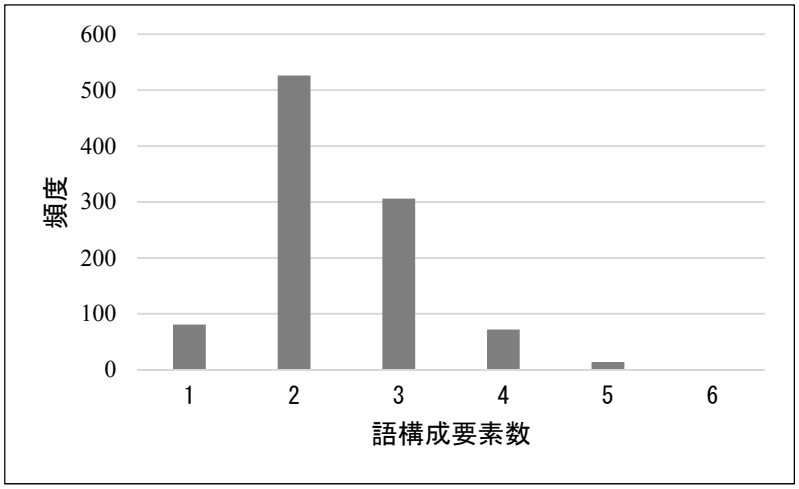

図 2　1 語あたりの語構成要素数

また，個々の医療用語を構成する文字数を語構成要素数で割った，1 語構成要素当たりの文字数は，最小値 1.5，平均値 2.75，中央値 2.5，最大値 11.0 となっている。1 語構成要素あたりの文字数が 4 を越えると，「エンテロウイルス髄膜炎」など，カタカナで表記される外来語が増える傾向にある。

6. 語の中の位置からみた語構成要素

語構成要素が語中のどの位置によく現れるのか，そのパターンを調べた。調べ方は語頭から見た場合と語末から見た場合の 2 種類である。いずれも表 2 で意味ラベルの出現頻度が 25 以上の 13 語について調べた。

6. 1 語頭からみた場合

意味ラベルの出現割合を語頭からの位置別に表したのが表 3 である。表 3 の見方であるが，例えば「疾患」という語構成要素は，語頭から 1 個目（すなわち語頭）の位置の出現割合は 3.60 であることを示す。この出現割合は，例えば，「疾患」が語頭から 1 の位置（すなわち語頭そのもの）に現れた数を，当該の位置を持つ語の数（語頭から 1 であれば，全ての語がその位置を持つので 1,000，語頭から 2 であれば，語構成要素数 1 の語が差し引かれるので，919 などとなる）で割って 100 を掛けた値である。出現した語の数自体は，語頭からの位置により「36，339，179，40，11，1」と少なくなっていくが，その位置に語構成要素を持つ語自体が少なくなっていくので，出現割合としては次第に高くなっていく。

表 3　語構成要素の出現割合（語頭から見た場合）

位置 / ラベル	1	2	3	4	5	6
疾患	3.60	36.89	45.55	45.98	73.33	100.00
身体部位	39.20	15.67	8.65	6.90	-	-
状態	29.60	13.60	8.40	10.34	20.00	-
症状	1.30	5.66	6.11	4.60	6.67	-
医療行為	1.00	2.72	8.91	17.24	-	-
時間	4.00	1.09	1.02	-	-	-
生理	1.70	2.83	1.53	-	-	-
動き	1.00	2.83	1.53	3.45	-	-
物質	3.10	0.87	1.27			
空間	1.50	1.09	1.27	-	-	-
人間	2.30	0.33	0.51	1.15	-	-
行為	1.30	0.98	0.76	-	-	-
障害	0.10	0.98	3.56	1.15	-	-

表 3 から，語構成要素の出現に偏りがあることが見て取れる。例えば，「疾患」「医療行為」のように語末に向かって出現割合が高くなるもの。その逆に，「身体部位」「時間」のように，語末に向かって出現割合が低くなるものがある。ほかにも「物質」「人間」は語頭に，障害は語末から 1 つ前に集中しているようである。「状態」は，語頭と語末が高く，中間が低いという分布を示している。

6. 2 語末からみた場合

表 4 は語末から見た場合の語構成要素の出現割合である。おおむね表 3 と一致する傾向

が見て取れる。表 4 の「位置」は，語末からの位置であり，1 が語末を表し，2 が語末から 1 つ前の位置を表す。表 3 では語末に向かって出現割合が高くなっていた「疾患」は表 4 では語頭に向かって出現割合が低くなっている。「身体部位」「時間」も表 3 と一致する傾向であり，語頭に向かって出現割合が高くなっているように見える。「空間」「人間」「時間」も語頭に向かって出現割合が高くなると言ってよいだろう。また，語末に出現が集中しているのは，「医療行為」「症状」[4]「障害」であり，とくに「障害」は語末のみに出現していることが分かった。

表 4　語構成要素の出現割合（語末から見た場合）

位置 / ラベル	1	2	3	4	5	6
疾患	54.30	5.98	2.04	-	-	-
身体部位	2.00	41.13	35.62	34.48	46.67	100.00
状態	6.40	23.94	38.42	28.74	40.00	-
症状	8.00	1.41	0.25	-	-	-
医療行為	7.30	0.87	0.76	1.15	-	-
時間	0.80	2.83	3.82	4.60	6.67	-
生理	1.70	3.05	0.51	2.30	-	-
動き	2.30	1.74	1.27	1.15	-	-
物質	0.30	2.83	3.31	2.30	-	-
空間	0.20	1.52	2.54	4.60	-	-
人間	0.40	1.52	1.27	6.90	-	-
行為	1.30	0.98	0.76	-	-	-
障害	2.50	-	-	-	-	-

7. 語構成要素数別に見た場合

本節では，語構成要素数別に語構成要素の出現の傾向を見る。

7. 1　語構成要素数 1 の場合

表 5 に示したように「疾患」「身体部位」「状態」が上位に来ている。

表 5　語構成要素数 1 の場合（上位 5 語）

意味ラベル	出現数
疾患	23
身体部位	12
状態	10
症状	6
医療行為	4

7. 2　語構成要素数 2 の場合

表 6 から，語構成要素数 2 の場合，第 1 要素と第 2 要素の間の上位には共通する意味ラベルがなく，これらの組み合わせにより，医療用語が構成されていることが示唆される。ちなみに，意味ラベルの 2-gram の集計では，いちばん多かったのは「身体部位」+「疾患」

[4] 「症状」は表 3 では明確な傾向が認められなかった。

で 270 回出現，その次が「状態」＋「疾患」で 157 回の出現であった。

表 6　語構成要素数 2 の場合（上位 7 語）

意味ラベル	第 1 要素での頻度
身体部位	228
状態	140
時間	20
物質	18
生理	13
動植物	13
人間	13

意味ラベル	第 2 要素での頻度
疾患	304
症状	46
状態	25
医療行為	22
疾患・状態	16
動き	15
動き・状態	14

7．3　語構成要素数 3 の場合

語構成要素 3 では，位置的な分布に特徴があるものを抜き出して表にした。表 7 から「身体部位」「状態」は第 1 要素（語頭）と第 2 要素（語中）に多く出現すること，「疾患」「医療行為」「症状」「障害」は第 3 要素（語末）に多く出現することが分かる。

表 7　語構成要素数 3 の場合

意味ラベル	第 1 要素での頻度	第 2 要素での頻度	第 3 要素での頻度
身体部位	118	114	0
状態	117	70	20
疾患	5	32	166
医療行為	3	3	32
症状	1	6	23
障害	0	0	14
生理	1	13	4

7．4　語構成要素数 4 の場合

語構成要素 4 でも，位置的な分布に特徴があるものを抜き出して表にした。表 8 から「身体部位」は第 1（語頭）～第 3 要素にかけて広く出現すること，「状態」は第 1 要素（語頭）と第 2 要素に多く出現すること，「疾患」「医療行為」は第 4 要素（語末）に多く出現すること，「空間」は語中の要素のみに出現することが分かった。

表 8　語構成要素数 4 の場合

意味ラベル	第 1 要素での頻度	第 2 要素での頻度	第 3 要素での頻度	第 4 要素での頻度
身体部位	27	19	30	1
状態	23	28	7	7
疾患	0	3	13	38
医療行為	1	0	3	15
空間	0	3	4	0

8. まとめと今後の予定

本発表では，実践医療用語を医療従事者の立場から，語構成要素に分割し，意味ラベルを付与する試みを行い，その結果に基づき，実践医療用語の語構成のパターンの一端を明らかにした。今回用いた語構成要素は，分割基準がまだ明確でなく，意味ラベルの付与規則も十分に検討されているとは言いがたい。今後これらを文書化し，オープンな形で充実させていく予定である。

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）の助成を受けたものである。

文　献

石井正彦（2007）『現代日本語の複合語形成論』，ひつじ書房.
国立国語研究所（2004）『分類語彙表 増補改訂版』，大日本図書.
相良かおる，小野正子（2018）「実践医療用語辞書 ComeJisyoSjis-1 の作成」，p.1491-1494，言語処理学会第 25 回年次大会発表論文集.
東条佳奈，相良かおる，小野正子，山崎誠（2019）「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.