End-of-Word Survey of Compound Words Representing Disease Names

病名を表す合成語の語末調査

相良 かおる、高崎 智子（西南女学院大学）
東条 佳奈、麻 子軒（大阪大学）
山崎 誠（国立国語研究所）

End-of-word survey of compound words representing disease names

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

要旨

本研究では合成語の語末により病名の判別が可能か否かを確認するために合成語の語末調査を行った。具体的には、病名を表す合成語 5,465 語について語末の unigram、bigram、trigram を調べた。加えて、筆者等が着手している電子カルテに記載された合成語を対象とした語構成要素解析で定めた語単位で、合成語を分割した場合に語末となる語構成要素の頻度を調べた。

その結果、①右側主要部の規則による意味を用いた判別が可能なこと、②病名の末尾には「ヘモクロマトーシス」のようなカタカナ語があり、文字単位での判別より語単位の判別の方が適していることが分かった。また、意味ラベル「接尾語」が付与された語構成要素と、「病名」が付与された語構成要素の内「接尾語」を含まない要素を用いることで「病名」の機械的な判別が可能であることが示唆された。

１．はじめに

電子カルテに記載される情報には、医師が入力する初診時の記録、経過記録、退院時サマリ、そして看護師が入力する入院患者の観察記録や看護記録がある。日々の電子カルテの入力は時間に追われ断片的になりがちであり、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。

これら電子カルテデータの利活用を支援する為に、形態素解析器 Mecab で利用可能な実践医療用語辞書 ComeJisyo（登録語数約 11 万語）や、症状や病名に特化した万病辞書データ「万病辞書（登録語 362,866 語）」が無償で公開されている。

電子カルテデータを利活用する上で重要な情報に「病名」がある。万病辞書には、約 36 万語の症状および病名が登録されているが、それでも全ての病名を登録している訳ではなく、電子カルテに含まれる全ての用語を網羅した辞書の作成は困難である。

筆者等は電子カルテに含まれる合成語の意味的または統語的な規則を用いて合成語の抽出や意味分類ができないかと考え、現在、ComeJisyo の登録語から合成語 7,194 語を選定し、語構成の解析と意味分類に着手している。その過程で、語末が「症」「炎」「病」であれば、半機械的に意味ラベルを「病名」にしていることに気付き、右側主要部の規則により合成語が病名であるか否かを語末で判別できないかと考え、その可能性を確かめる為に今回、5,465 語の病名について語末調査を行った。本発表では、その結果と、合成語が「病名」か否かの判別方法についてのアイディアについて述べる。

２．用語の定義

語構成要素：

合成語を構成する要素で、本研究における語構成要素を、「医療の観点から意味的にまたは統語的に分割可能な語をすべて語」と定義する。

合 成 語 ： 上口唇皮下腫瘍
語構成要素： 上口唇、皮下腫瘍、皮下、腫瘍

語構成要素列：

合成語を構成する語構成要素の順序組。医療従事者が医療の観点から意味的にまたは統語的に妥当な一つの語単位を語構成要素とする語構成要素列（長）と、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語構成要素列（短）の 2 種類の語構成要素列が求まる。

合 成 語：上口唇皮下腫瘍
語構成要素列（長）：上口唇／皮下腫瘍
語構成要素列（短）：上口唇／皮下／腫瘍

意味ラベル：

語構成要素に付与する意味を表すラベルで、石井（2007）の意味的カテゴリーによる分類を参考に、筆者らが命名したものである。なお、「発育不全」のようにそれだけでは意味を限定できない多義の語構成要素の場合は「状態・症状・病名」のように複数の意味ラベルを列挙している。

現在、合成語 7,194 語より求めた語構成要素 5,350 要素に付与した意味ラベルは 101 種あり、妥当性について精査中である。

「病名」と意味的に近いラベルに「病因」「病態」「接尾語」がある。この場合の「接尾語」とは、自立語として使われることがなく、前に別の要素が付くことで「病名」となる要素を意味し、「失調症」「狭窄症」「欠乏症」など、82 要素に付与されている。

語構成要素列（長）：上口唇／皮下腫瘍
意 味 ラ ベ ル： 身体部位／病 名
語構成要素列（短）：上口唇／皮下／腫瘍
意 味 ラ ベ ル： 身体部位／身体部位／病 名

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：上口唇皮下腫瘍
短単位列：上／口唇／皮下／腫瘍

右側主要部の規則：

形態的に複雑な語（合成語）の主要部はその語の右側の要素にあるという規則である。

３．方法

３.１ 言語資源

語構成解析と意味分類用に ComeJisyoSjis-1 の登録語 111,664 語より選定した合成語 7,139 語の内、医療情報システム開発センター（MEDIS-DC）の病名マスターに登録されている 5,465 語を本解析データとする。なお、合成語 7,139 語の選定は、先ず始めに ComeJisyoSjis-1 の登録語を 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照合し、一致した 31,162 語を求め、次いで専門外の分析者による意味分類の作業を考慮し、『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）を含む合成語を抽出している。

３.２　調査方法

Step.1　対象合成語の末尾 unigram、bigram、trigram を求める。

Step.2　語構成要素列（短）の末尾の語構成要素の頻度を求める。

Step.3　語末の語構成要素の意味ラベルの頻度を求める。

Step.4　病名と判別できる語末を調べる。

４．結果

表 1 は、対象合成語 5,465 語の末尾の unigram、bigram、trigram と語末の語構成要素の種類をまとめたものである。

表 1　語末の n-gram と語構成要素、そして意味ラベルの種類

語末	種類
1 文字	226
2 文字	591
3 文字	1,830
語構成要素	1,138
意味ラベル	52

表 2　語末 unigram、bigram、trigram の結果

順位	1 文字	語数	相対度数	累積相対度数	2 文字	語数	相対度数	累積相対度数	3 文字	語数	相対度数	累積相対度数
1	症	754	13.8%	13.8%	＊腫瘍	407	7.4%	7.4%	＊性腫瘍	117	2.1%	2.1%
2	傷	664	12.2%	25.9%	＊損傷	401	7.3%	14.8%	＊狭窄症	100	1.8%	4.0%
3	炎	494	9.0%	35.0%	麻痺	189	3.5%	18.2%	＊感染症	85	1.6%	5.5%
4	＊瘍	452	8.3%	43.3%	出血	167	3.1%	21.3%	＊性貧血	81	1.5%	7.0%
5	血	268	4.9%	48.2%	＊挫傷	158	2.9%	24.2%	＊脈損傷	78	1.4%	8.4%
6	痺	189	3.5%	51.6%	障害	135	2.5%	26.7%	＊皮膚炎	64	1.2%	9.6%
7	腫	136	2.5%	54.1%	＊中毒	127	2.3%	29.0%	後遺症	62	1.1%	10.7%
8	害	135	2.5%	56.6%	骨折	123	2.3%	31.2%	＊部挫傷	61	1.1%	11.9%
9	毒	134	2.5%	59.0%	＊膜炎	100	1.8%	33.1%	＊経損傷	60	1.1%	13.0%
10	折	123	2.3%	61.3%	＊窄症	100	1.8%	34.9%	＊髄膜炎	57	1.0%	14.0%
11	臼	95	1.7%	63.0%	脱臼	95	1.7%	36.6%	＊症候群	56	1.0%	15.0%
12	縮	94	1.7%	64.7%	貧血	91	1.7%	38.3%	＊部腫瘍	53	1.0%	16.0%
13	核	92	1.7%	66.4%	＊結核	88	1.6%	39.9%	＊動脈瘤	53	1.0%	17.0%
14	瘤	83	1.5%	67.9%	萎縮	87	1.6%	41.5%	＊ン中毒	40	0.7%	17.7%
15	常	71	1.3%	69.2%	＊染症	85	1.6%	43.1%	＊欠乏症	38	0.7%	18.4%
16	裂	58	1.1%	70.3%	＊熱傷	74	1.4%	44.4%	＊経麻痺	36	0.7%	19.0%
17	窄	58	1.1%	71.4%	＊脈瘤	71	1.3%	45.7%	＊管損傷	36	0.7%	19.7%
18	群	56	1.0%	72.4%	異常	71	1.3%	47.0%	＊臼骨折	34	0.6%	20.3%
19	＊挫	56	1.0%	73.4%	＊膚炎	64	1.2%	48.2%	放骨折	33	0.6%	20.9%
20	創	54	1.0%	74.4%	遺症	62	1.1%	49.3%	＊パチー	33	0.6%	21.5%

表 2 は、末尾 1 文字、2 文字、3 文字について高頻度のもの上位 20 位迄の頻度と相対度数、累積相対度数をまとめたものである。末尾 1 文字では、異なり 226 種の内、「症」

「傷」「炎」「瘍」「血」の上位 5 位で全体の 48.2%を占め、上位 20 位迄では全体の 74.4%を占めている。末尾 2 文字では「腫瘍」と「損傷」の相対度数がそれぞれ 7.4%と 7.3%であり、全 591 種の内、上位 20 位で全体の 49.3%を占めている。そして末尾 3 文字では 1,830 種の内、上位 20 位で全体の 20%を占めている。

文字を単位とした 2 文字と 3 文字の上位 20 位の中で、言語学的に意味を持つものは 2 文字では 14 種、3 文字では 8 種である。

なお、表 2 の各文字の前に付与された「＊」は、語構成要素 5,350 要素から該当する文字を末尾に持つ要素を調べた結果、その意味ラベルに「病名」「接尾語」「病因」「病態」以外のものがないことを示している。例えば「症」を含む語構成要素に「軽症」があり、意味ラベルは「状態・程度」が付与され、末尾に「症」がくる合成語を「病名」と機械的に判断できないことを示している。一方、trigram では、上位 20 種の内、末尾から合成語を「病名」と判断しても大きく間違うことのないものが 18 種類あることを示している。

なお、意味ラベル「病因」が付与された語構成要素として「中毒」が、「病態」が付与されたものに「梗塞」がある。

表 3　語末の語構成要素の頻度

順位	語構成要素(短)	語数	相対度数	累積相対度数	意味ラベル
1	＊腫瘍	406	7.4%	7.4%	病名
2	＊損傷	376	6.9%	14.3%	病名
3	出血	164	3.0%	17.3%	症状
4	＊挫傷	132	2.4%	19.7%	病名
5	麻痺	122	2.2%	22.0%	状態
6	＊中毒	118	2.2%	24.1%	病因
7	障害	113	2.1%	26.2%	障害
8	＊狭窄症	100	1.8%	28.0%	接尾語
9	貧血	91	1.7%	29.7%	症状・病名
10	＊感染症	85	1.6%	31.2%	病名
11	脱臼	78	1.4%	32.7%	状態・病名
12	＊熱傷	74	1.4%	34.0%	病名
13	＊皮膚炎	63	1.2%	35.2%	病名
14	＊結核	63	1.2%	36.3%	病名
15	後遺症	61	1.1%	37.4%	症状
16	狭窄	58	1.1%	38.5%	状態
17	＊捻挫	56	1.0%	39.5%	病名
18	＊髄膜炎	55	1.0%	40.5%	病名
19	破裂	49	0.9%	41.4%	動き・状態
20	＊肉腫	41	0.8%	42.2%	病名

表 3 は、語構成要素列（短）の末尾の語構成要素の高頻度のもの上位 20 位迄の相対度数、累積相対度数、そして意味ラベルをまとめたものである。第 1 位は「腫瘍」で 7.4%、次いで「損傷」が 6.9%となり、語構成要素 1,138 種の内、上位 20 位迄で全体の 42.2%を占めている。語構成要素に付与された意味ラベルを見ると、上位 20 位の内、10 要素は「病名」であり、「病名」を含む要素は「症状・病名」「状態・病名」の 2 要素である。また、「病名」「接尾語」「病因」「病態」以外の意味ラベルが付与されたものは 9 要素である。

表 4 は、語末の語構成要素に付与された意味ラベルの頻度の上位 20 位迄の相対度数と累積度数をまとめたものである。第 1 位は「病名」で全体の 49.3%を占めている。上位 10 位迄で全体の 91.4%を占め、「病名」「接尾語」「病因」「病態」以外の意味ラベルに「状態」「症状」「障害」「動き」があり、これらの相対度数和は 27.2%である。意味ラベル「状態」が付与されたものには「麻痺」が、「症状」には「出血」が、「障害」には「障害」が、「動き」には「破裂」がある。

表 4　語末にくる語構成要素の意味ラベルの頻度

	意味ラベル	語数	相対度数	累積相対度数
1	病名	2,692	49.3%	49.3%
2	状態	737	13.5%	62.7%
3	症状	505	9.2%	72.0%
4	接尾語	306	5.6%	77.6%
5	状態・病名	190	3.5%	81.1%
6	障害	127	2.3%	83.4%
7	動き・状態	122	2.2%	85.6%
8	病因	118	2.2%	87.8%
9	症状・病名	114	2.1%	89.9%
10	病態	82	1.5%	91.4%
11	行為	62	1.1%	92.5%
12	生理	62	1.1%	93.6%
13	身体部位	57	1.0%	94.7%
14	状態・病態	56	1.0%	95.7%
15	動き	41	0.8%	96.5%
16	医療行為・状態	37	0.7%	97.1%
17	医療行為	23	0.4%	97.5%
18	人間	17	0.3%	97.9%
19	精神	14	0.3%	98.1%
20	体外物質	10	0.2%	98.3%

５．考察とまとめ

本調査の過程で、「軽症」「重症」に意味ラベル「状態・程度」を付与していることから、末尾が「症」「炎」「病」の場合、機械的に「病名」の意味ラベルを付与しているのではなく、瞬時に意味を解釈していることに気付いた。また病名を表す 5,465 語の内、末尾が「症」で終わる合成語は 754 語と多いものの、次いで多いものは「傷」で終わる合成語 664 語であり、「炎」で終わる合成語は 76 語（第 23 位）と予想より少なかった。そして「炎」を末尾に持つ語構成要素に意味ラベル「治療」が付与された「消炎」を見つけ、末尾に「炎」がくる合成語は病名を表すという仮説は正しくないこと、専門外の人の思い込みの危うさに気付いた。

しかしながら、病名を表す合成語の語末にくる語構成要素の意味ラベルの 72.8%に「病名」「接尾語」「病院」「病態」が含まれることから、意味的な判別に右側主要部の規則が利用可能であることが示唆された。

本研究の目的である語末の文字を用いた病名の機械的判別であるが、unigram では、病名を精度よく判別できる可能性の高い文字は上位 20 種の内、「瘍」と「挫」の 2 文字であり、これらの相対度数和が 9.3%であることから、病名を表す合成語の約 1 割の判別に寄与すると考えられる。また bigram では、上位 20 種の内 11 種の相対度数和は 30.6%となり、約 3 割の判別に寄与し、trigram では、上位 20 種の内 18 種の相対度数和は 19.8%となり、約 2 割の判別に寄与すると考えられる（表 2）。「瘍」の相対度数 8.3%に対し、「腫瘍」の相対度数は 7.4%であり、残り 0.9%は第 34 位の「膿瘍」と第 59 位の「潰瘍」であった。

病名の末尾にしか出現しない 2 文字のリストがある場合、異なりの種類（表 1）と n-gram の結果から、末尾の文字による判別では 2 文字による方法が良いと考えられる。

一方、語構成解析により求めた語構成要素では、上位 20 要素の内 12 要素の相対度数和は 28.7%となり、約 3 割の判別に寄与することが分かる（表 3）。

末尾にくる語構成要素の種類は 1,138 種（表 1）で、trigram の 1,830 より種類は少なく、

文字長の平均は、3.6 文字である。文字長が最も長いものは 12 文字の「フリードライヒ運動失調症」であり、意味ラベルは「病名」が付与されている。今回の調査で「病名」の意味ラベルが付与された語構成要素には、「ヘモクロマトーシス」「パラミオクローヌス」のようなカタカナ語が含まれ、これらを末尾に持つ合成語において、末尾の文字単位では高精度の判別は難しい。末尾がカタカナ語である合成語の判別においては文字単位ではなく、語単位である語構成要素を用いた判別が良い。

人間の健康の維持、回復、促進などを目的とする医療の世界では、結果の過程を明らかにし、言語化することが重要である。従って、言語学的に意味を持たないものを含む文字単位での判別では、文字リストをどのようにして作成したかの説明が必要となる。

一方、語構成要素と付与された意味ラベルを用いた判別では、語構成要素に意味ラベルが付与されているため、人間可読である。右側主要部の規則に則れば意味ラベルが「病名」のみの語構成要素を用いることで大まかな判別が可能となる。しかしながら、語構成要素は「医療の観点から意味的にまたは統語的に分割可能な語」であることから「フリードライヒ運動失調症」などの合成語が含まれる。短単位による分割は「フリードライヒ／運動／失調／症」、一般的な意味で分割すると「フリードライヒ／運動／失調症」となる。「フリードライヒ（人名）」は、医療の観点から意味的にまとまった語ではないことから、分割されず、その結果、より結合の強い合成語「運動失調症」の分割が出来ず「フリードライヒ運動失調症」が一つの語構成要素となる。しかしながら、病名の判別では、病名の接尾語「失調症」だけで判別が可能である。

以上のことから、①意味ラベル「接尾語」が付与された語構成要素と、②「病名」が付与された語構成要素の内「接尾語」を含まない要素を用いることで、「病名」の機械的な判別ができると考えられる。

そこで語構成要素と意味ラベルを精査した上で判別用の要素を選定し、病名の判別実験を行う予定である。

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。

文　献

国立国語研究所（2004）：分類語彙表　増補改訂版，大日本図書
石井正彦（2007）：現代日本語の複合語形成論，ひつじ書房
鈴木良太編（2006）：言語科学の百科事典，丸善

関連 URL

実践医療用語辞書 ComeJisyo　https://ja.osdn.net/projects/comedic/stats/frs
万病辞書　http://sociocom.jp/~data/2018-manbyo/index.html
MEDIS 標準マスター・インデックス　https://www.medis.or.jp/4_hyojyun/medis-master