実践医療用語_語構成要素語彙試案表 Ver.2.0 の構築

東条佳奈（大阪大学）
黒田航（杏林大学）
相良かおる（西南女学院大学）
高崎智子（西南女学院大学）
西嶋佑太郎（医師）
麻子軒（関西大学）
山崎誠（国立国語研究所）

Development of a Word Component Database for "Hands-On Medical Terms" (Version 2.0)

Kana Tojo (Osaka University)
Kou Kuroda (Kyorin University)
Kaoru Sagara (Seinan Jo Gakuin University)
Satoko Takasaki (Seinan Jo Gakuin University, Physician)
Yutaro Nishijima (Physician)
Tzu-Hsuan Ma (Kansai University)
Makoto Yamazaki (National Institute for Japanese Language and Linguistics)

要旨

医療記録データには、複数の単語が連結された合成語が多く存在する。そのため、自然言語処理を効率的に行うためには、合成語の語構成や、それらの構成要素の意味に着目し、合成語の構造を明らかにする必要がある。しかし、医療記録は非公開という資料的特質のため、言語学的な調査があまり行われてこなかった。また、医療関係者における意味のある言語単位も定まっておらず、整理の必要があった。こうした背景に基づいて作成した言語資源が『実践医療用語_語構成要素語彙試案表 Ver.2.0』である。本試案表は、『実践医療用語辞書 ComeJisyoSjis-1』より抽出した合成語より作成した『実践医療用語_語構成要素語彙試案表 Ver.1.0』を更新したもので、7,087 語の合成語について、それぞれを構成する語構成要素 6,633 種と、語構成要素に付与した意味ラベル 41 種を収録している。本発表では、Ver1.0 からの変更点と、本言語資源の特徴、意味ラベルに注目した語構成要素について概観を行った。

１．はじめに

本発表は、2022 年に言語資源協会より無料で配布を開始した電子データである『実践医療用語_語構成要素語彙試案表 Ver.2.0』[1]（以下,「本試案表」とする）について紹介するものである。

電子カルテシステムの普及により医療記録データが蓄積されていくにつれ、利用者が求

[1] https://www.gsk.or.jp/catalog/gsk2020-g

める情報の適切な検索や、二次利用のための技術の重要性が高まっている。自然言語処理では、処理を行うための最初のステップとして、テキストを意味のある最小の単位（一般には単語）に分割する前処理が必要となる。英語のテキストの場合、単語間は基本的にスペースで区切られているため、コンピュータでの単語に分割する処理も容易である。一方、分かち書きのされない日本語のテキストは、単語分割の処理が容易ではない。

特に、医療記録データには「腹腔鏡補助下低位前方切除 D2 郭清施行」の様に複数の単語が連結された合成語が多く、医療関係者にとって意味のある最小の単位がどういうものかという統一的な見解もないため、語分割のための情報が重要となる。加えて、医療記録は各々の医療現場で作成されるため、表記の揺れ、誤字、標準的でない用語などを含む専門的かつ実践的な医療用語が多用されており、効率的な自然言語処理のためには、それらの語構成や語種構成の実態の調査も必要である。

しかし、医療記録データは個人情報が豊富に含まれているために非公開とされており、こうした資料的特質から、実践医療用語の言語学的な調査は、十分に行われてはこなかった。

こうした背景のもと、発表者らは、医療記録文の自然言語処理支援のために相良が作成・公開した形態素解析器 MeCab 用のユーザー辞書『ComeJisyo』の登録語[2]より、医学的な知識が不足していても語の境界がわかりやすいように、「『分類語彙表―増補改訂版―』に収録されている語を含む合成語」を抽出し、語構成要素とそれぞれの意味ラベルを付与した『実践医療用語_語構成要素語彙試案表 Ver.1.0』（以下、「試案表 Ver.1.0」とする）を作成した（詳細な手順については相良他 2019、相良他 2020、相良 2021 を参照）。試案表 Ver.1.0 において見直しの必要があった意味ラベルの妥当性（相良 2021）を踏まえて、合成語や意味ラベルを見直し、新規の情報を追加するなどのアップデートを施したものが、Ver.2.0 にあたる本試案表である。

具体的には、試案表 Ver.1.0 より不要な語を除いた 7,087 語を対象合成語とし、これらの合成語を構成する語構成要素 6,633 要素、およびそれぞれの要素に付与した 41 種類の意味ラベルを収録したものとなっている。

ファイル種別は Excel ファイル（拡張子 xlsx）で、「はじめに」（用語の定義及び各シートの解説）、「対象語」（対象語一覧）、語構成要素（語構成要素一覧）、意味ラベル（意味ラベル一覧）の 4 つのシートから構成されており、それぞれのシートには、以下の情報が収録されている。

（1）「対象語」シート（7,087 語）：

通し番号（ID）・対象合成語（見出し）・対象合成語の読み・語構造・文字長情報

「語構成要素」シート（6,633 種）：

通し番号（ID）・語構成要素・語構成要素の読み・意味ラベル・対象語における語構成要素の出現頻度・当該語構成要素が語頭に出現した頻度・当該語構成要素が語末に出現した頻度・語構成要素が対象語と一致する場合の対象語 ID

「意味ラベル」シート（41 種）：

通し番号（ID）・意味ラベル・当該意味ラベルに分類された語構成要素の個数・意味ラベルの属性（具体か抽象か）

---

[2] 分かち書き用実践医療用語辞書 ComeJisyo
https://ja.osdn.net/projects/comedic/
試案表作成の際には ComeJisyoSjis-1 の登録語（111,664 語）を使用した。

以下、本試案表における用語と、試案表 Ver.1.0 からの変更点について触れながら、本言語資源の特徴を述べる。

２．本試案表における用語について

本試案表が対象とする「実践医療用語」は、医療施設で使われる医療記録に含まれる「学術上の専門用語」と「それ以外の専門用語」を指す（相良 2021）。

また、本試案表で扱う「語構成要素」は、相良（2021:559）に従い、「合成語を構成する要素で、合成語を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部分文字列」と定義し、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とらえる。例えば（2）のように、「脳幹多発性硬化症」という合成語は、「脳幹」「多発性」「硬化症」という短い語構成要素のほか、「多発性硬化症」という語構成要素同士が結合した形も一つの語構成要素として捉えて集計している。

（2）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

また、本試案表においては、それぞれの語構成要素に、医療の観点での意味を表すラベルとなる「意味ラベル」を付与している。意味ラベルは語構成要素と必ずしも 1:1 の関係ではなく、語構成要素が多義の場合は複数の意味ラベルを「,」で区切り、列挙して示している。例えば語構成要素「流動食」の意味ラベルは《医療行為，食品，治療食》を付与している[3]。「流動食」は、食品という点では一般の食品が含まれており、同時に《治療食》という《医療行為》の一つとも考えられる。そこで《医療行為》と《食品》と《治療食》の意味ラベルを併記して示している。なお、ここでの「意味」は、用語の意味を表すもの（例：《衛生物品》）と，医療者の使用面での心的な捉え方を表すもの（例：《病因》）が含まれている。

３．『実践医療用語_語構成要素語彙試案表 Ver.1.0』からの主な変更点

1 節で述べたように、本試案表は、2021 年に公開した試案表 Ver.1.0 を見直し、変更を加えたものである。主な変更点は、①不要な対象合成語の削除、②合成語の語構造情報の付与、③意味ラベルの見直し・統合である。そのほか、対象合成語・語構成要素の見直しに伴い変更した情報についても更新を行っている。

①については、異体字による重複・誤字・現在では使われない語等を削除した。②の語構造情報は、(3) のように対象合成語の分割位置を医師が確認し、[ ]で区切りを入れたものである。これにより、(4) のような長大な合成語についても意味理解がしやすくなったといえる。また、合成語がどのような結合パターンになっているかについての検討なども可能になった。

（3）対象合成語：悪性脳腫瘍
　　語構造：悪性[脳[腫瘍]]
　　語構成要素：悪性，脳，腫瘍，脳腫瘍

[3] 本発表では便宜上、意味ラベルは《》に入れて示す。

（4）対象合成語：脳淡蒼球内オイルプロカイン注入療法
語構造：[脳[淡蒼球内]][[オイル[プロカイン]][注入[療法]]]

次に③についてだが、本試案表では、80 種類とかなり細分化されていた試案表 Ver.1.0 の意味ラベルを見直して統合を行い、41 種類にまとめた。（5）に試案表 Ver.1.0 より削除したラベル 46 種類を、（6）には本試案表のラベル 41 種類を示す。なお、今回新たに追加したラベル 7 種類には下線を付している。

（5）「試案表 Ver.1.0」より削除したラベル 46 種：
意向，動き，音，化学現象，課題，感覚，関係，基準，軌跡，基礎，規則，機能，教育，距離，傾向，行為，作用，時間，色彩，社会，手技，主体，順序，数量，成育，性質，制度，増減，属性，体外物質，体内物質，知的産物，調節，程度，認識，熱，能力，波動，光，文法用語，変化，方向，保健衛生，保留，様相，例示

（6）「試案表 Ver.2.0」のラベル 41 種：
維持行為，位置，医薬品，医療行為，衛生物品，化学物質，患者属性，患部，機器，経過，形状，検査，サービス，施設，指標，種類，症状，状態，食品，身体機能，身体部位，精神，生体物質，生理，組織，治療食，動植物，排泄物，場所，ヒト，費用，病因，病原体，病態，病名，物品，部分，法規，方法，予防行為，#未定

試案表 Ver.1.0 から削除した意味ラベルの中で、最も語構成要素が多かったのは《体内物質》、次いで《体外物質》であった。これらはより分類しやすいラベルとして、《生体物質》《化学物質》のラベルに変更した。また、（5）には、《感覚》（例：「感覚」「知覚」「聴覚」など）《動き》（例：「感覚運動」「眼球運動」「関節運動」など）《機能》（例：「肝機能」「呼吸機能」など）のように、どのように異なるのか区別が判然としない類義の意味ラベルも含まれていた。これらのラベルについては上位ラベルとして《身体機能》を想定し、《身体機能，精神》《身体機能，生理》のように複数のラベルを付与することで整理を行った。結果として、《感覚》《動き》《機能》などのラベルは、《身体機能，生理》という 2 種のラベルを付与することで、「自我意識」「欲求」などの精神的な身体機能と区別できるようになった。

加えて、試案表 Ver.1.0 に付与されていた意味ラベルのうち、該当する語構成要素がごくわずかしかなかったものは、合成語の中でどのような意味合いかを再度検討し、より上位の意味ラベルに変更した（例：「認知」《認識》→《身体機能，生理》、「白色」《色彩》→《状態》など）。削除した意味ラベルの中には、過分割していた語構成要素の見直しを行った結果、不要になったものもある。例を挙げると、「任意」《意向》→「任意入院」《医療行為》のような変更である。一方で、本来は分割すべき語構成要素を短くすることも行っている（例：「ぶどう膜炎」《病名》より、「ぶどう膜」《身体部位》を語構成要素として追加）。このように本試案表では、より医療の観点で有用になるように語構成要素の分割位置に関する調整も行った。

そのほか、命名が困難なものや、医療用語特有のものではない一般的な語構成要素[4]に「#未定」のラベルを付与したことも大きな変更点である。

---

[4] 『岩波国語辞典第五版タグ付きコーパス 2004』に立項されているかどうかを判断の基準とした。

４．意味ラベルからみる語構成要素の概観

本試案表では、新たに意味ラベルごとに、分類された語構成要素の個数の情報を追加した。語構成要素数の降順に 20 位まで示したものが表 1 である。

最も多いものが《病名》（1918 種）で、《状態》（1669 種）、《身体部位》（1639 種）、《病態》（1202 種）と続く。この 4 種の意味ラベルで全体の約 70%という割合となる。本試案表に収録されている対象語のうち、大部分が病名である[5]ため、《病名》ラベルが最も多くなる。同様に、病名にはその症状が発生している患部が同時に含まれることが多いため、《身体部位》ラベルが付与される語構成要素も高頻度となった。

《状態》ラベルが付与された語構成要素のほとんどは、《病態》ラベルと併記しているものになった。《病態》は、病的な状態や、単独では病名を構成しない語構成要素に付与しており、《状態》の下位にあたる意味ラベルである。さらに、《病態》の中には、「細菌性食中毒」における「細菌性」のように、《病因》であり、かつ《病原体》であるというものもある。この場合、(7) に示すように、意味ラベル同士の関係は等位ではなく、上位下位関係にあると解釈される。

表 1　語構成要素数上位 20 の意味ラベル

意味ラベル	語構成要素数
病名	1,918
状態	1,669
身体部位	1,639
病態	1,202
医療行為	319
症状	249
化学物質	241
部分	239
位置	235
経過	176
病因	145
＃未定	141
病原体	123
身体機能	96
生体物質	89
種類	83
生理	77
医薬品	72
形状	60
ヒト	54

[5] 《病名》もしくは《病態》ラベルが付与されたものが対象語全体のうち 73.6%（5,218 語）であった。

（7）対象合成語：細菌性食中毒

語構成要素	意味ラベル
「細菌性」	《状態》＞《病態》＞《病因》＞《病原体》
「食中毒」	《病名》

つまり、本試案表に付与されている意味ラベルのうち、「,」で列挙しているものの中には、包含関係で表されるものもあるということである。《状態》ラベルは意味ラベルの中では上位概念にあたるラベルとなっており、他にも、《患者属性》（例：「老人性貧血」の「老人性」）・《形状》（例：「線状網膜炎」の「線状」）・《経過》（例：「進行性筋萎縮」の「進行性」）などが《状態》の下位概念にあてはまる。

試案表 Ver.1.0 では、「～性」となる語構成要素をはじめ、あらゆるものに《状態》ラベルが付与されており、「意味」による分類が出来ているとは言い難いものになっていたが、本試案表で意味ラベルを複数付与・併記したことで、《状態》ラベルの細分化を行うことができたといえる。例えば、同じ「～性」でも、「サルコイドーシス性」は《病態》で、「コクシジオイデス性」は《病因》かつ《病原体》である、というように、「性」の前要素に関する知識がなくとも意味判断ができるようになった。

しかし、本試案表ではこのような上下関係または包含関係にあたるものも、そうではなく、多義のために併記しているもの（例：《位置》であり《経過》を示す「前」など）もいずれも「,」で区切って列挙をしており、見かけ上区別がつかず混在しているため、意味ラベルの関係性の整理が必要と思われる。

また、表 1 で示したように、《#未定》ラベルを付与している語構成要素が少なからずある。これらは前述の通り、命名が困難なものや、医療用語特有のものではない語構成要素があてはまるが、ラベルを作るほど合成語や語構成要素の数がない、というものも含まれる。本試案表の作成の目的の一つである、自然言語処理の効率化等のためには、未定ラベルの整理も課題であるといえよう。

５．おわりに

本発表では、『実践医療用語_語構成要素語彙試案表 Ver.2.0』の概要について、主に Ver1.0 との変更点と意味ラベルを中心に説明した。ComeJisyo より抽出した実践医療用語の合成語について、語構成要素と、それぞれに対応する意味ラベル、合成語の語構造情報を付与した本試案表を利活用することで、医療用語の語彙的な特徴の分析のほか、将来的に新たな医療記録データの解析に応用していくことが可能であると思われる。

意味ラベル間の関係の整理や、未定ラベルの検討についても引き続き行いながら、本試案表の精緻化を今後も行っていく予定である。また、これまでは、合成語の抽出の際に『分類語彙表―増補改訂版―』に収録されている「一般語を含む語」であることを条件にしていたため、語構成要素がすべて専門用語からなる合成語についても対象にしていく必要があるだろう。いずれも今後の課題としたい。

謝　辞

本研究は JSPS 科研費 JP18H03499 ならびに JP21H03777 の助成を受けている。

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.
相良かおる・山崎誠・麻子軒・東条佳奈・小野正子・内山清子（2019）「実践医療用語の語構成要素意味を基準とした分割」『人文科学とコンピュータシンポジウム論文集』pp.57-64.
相良かおる・小野正子・高崎智子・東条佳奈・麻子軒・山崎誠（2020）「実践医療用語の語構成と意味―語構成要素語彙試案表の作成にむけて―」『人文科学とコンピュータシンポジウム論文集』pp.289-296.
相良かおる（2021）「実践医療用語における語構成要素の意味ラベルについて」『言語処理学会第 27 回年次大会発表論文集』pp.559-562.
https://www.anlp.jp/proceedings/annual_meeting/2021/pdf_dir/P3-10.pdf

関連 URL

『岩波国語辞典第五版タグ付きコーパス 2004』https://www.gsk.or.jp/catalog/gsk2010-a/
『分かち書き用実践医療用語辞書 ComeJisyo』https://ja.osdn.net/projects/comedic/
『実践医療用語_語構成要素語彙試案表 Ver.2.0』https://www.gsk.or.jp/catalog/gsk2020-g