要素の重複と不連続性を扱える抽出型の語構成要素解析
並列分散型形態素解析の提案

黒田 航[1]　相良 かおる[2]　東条 佳奈[3]　麻 子軒[4]　西嶋 佑太郎[5]　山崎 誠[6]
[1]杏林大学　[2]西南女学院大学　[3]大阪大学　[4]関西大学　[5]医師　[6]国立国語研究所

概要

重複や不連続な語構成要素を持つ医療用語を対象に，それらを効果的に認識するための抽出型の構成要素認識法を二種類提案する．一つ目は完全並列分散型の語構成解析法 PDMA で，もう一つは PDMA より性能が劣るがアノテーションが楽な MLMA である．『実践医療用語_語構成要素試案表 Ver. 2』のサンプル 15 事例の解析を元に，語構成要素認識に一般的に使用されている句構造解析 (PSA) の平均認識率が，PDMA が認識できる要素の 65%程度，MLMA の 70%程度である事を示した．

1　はじめに

医療 (関連) 用語には複雑な医学用語が数多く含まれる．こうした用語を含む文章の意味処理をしたいなら，これらの用語の高性能な解析が必須である．解析が高性能であるためには，精度が高いだけでなく，被覆率が十分に高くなければならない．これは，言語処理の究極の目的が内容理解であり，その中間目標が高性能検索であり，その実現手段が高性能解析であるという処理間の依存構造を考えると，当然の帰結である．この観点からすると，医療用語を含めた専門用語の語構成解析にそれなりの先行研究 [1, 2, 3, 4, 5] があるにもかかわらず，被覆率が十分に高くないが故に，十分な成果が得られているとは言いがたい．低被覆率の証拠は §4 で示す．

低被覆率の理由は単純であると同時に，根本的なものである．どの手法も用語の構成要素を**所与の語の分割によって得ようとしている**が，医療用語の形態素は部分的に重複している度合いが高く，重複を許さない分割では，必要な要素が取りこぼされる．例えば『実践医療用語_語構成要素試案表 Ver. 2』[1] [6] に収録されている 7,087 事例からほぼランダムに選んだ 99 事例中，重複が生じている事例は少なくとも 73 事例である[2] (付録 A を参照)．

語構成要素の認識の (単なる精度向上ではなく) 被覆率の向上を目標を含めれば，理想的な実装法は，要素の認識を非排他的に行う多層抽出式が望ましいとわかる．本論に先立って次の事を確立しておく:[3]

(1)複合語 c の語構成要素の認識は，c の**分割 (segmentation)** でなく，c 中の有意味な部分文字列の (可能な限り) 網羅的な**抽出 (extraction)** の方が望ましい．

この後の本論で，(1) の目標を実現する語構成要素抽出法の実装例として，**並列分散形態論解析 (Parallel Distributed Morphological Analysis: PDMA)** を紹介する (§2)．それに続けて，PDMA が与える結果をそれなりの性能で近似する簡便なアノテーション方法 (**複層化形態論解析 (MLMA)**) を紹介し (§3)，MLMA と通常の句構造解析 (Phrase Structure Analysis: PSA) の性能を，PDMA の結果をベースラインにして比較する (§4)．最後に §5 で考察と展望を述べる．

2　語構成要素認識の並列分散化

医療用語には構成関係が複雑な複合語が頻出する．例えば (2) のような疾患名がそうである:

(2)先天冠状動脈異常

これを形態素解析した結果は (3) である (ここでは MeCab +UniDic 2.1.2 の解析結果を示した):

(3)先天/冠状/動脈/異常

この解析の実質は有意味な要素 (≈ 形態素) への分

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

2)　第一著者の非公式の調査では形態素重複の発生率は一般に，学術用語で一般用語より高いが，医療用語での出現率の高さは異例である．

3)　最適でない分割の悪影響は，単語分かち書きが所与でない言語 (例えば日本語) で深刻な問題であるが，単語分かち書きをする言語が形態素の重複から免れていると考えるのは妥当でない．英語のような言語でも**カッコ入れの逆理 (bracketing paradox)** [7] はそれなりの頻度で起こっている．例は generative grammarian の意味の構成が (generative (grammarian)) ではなく，((generative grammar) -ian) である事．

割である．だが，なぜ分割なのか？それは構成素解析が**適正解析 (proper analysis)** [8] と同一視されているからである．

では，なぜ適正解析でなければならないのか？それは，産物が**連結 (concatenation)** で元に戻せるからである．ただ，自然言語で要素合成法が連結でなければならない根拠は弱い．**重ね合わせ (superposition)** であっていけない理由は，どこにもない [9, 10]．

実際，「(自然) 言語の複合的な単位は要素の連結で得られる」という (現実反映の保証のない) モデル化が処理上の問題を幾つも発生させている．適正解析を想定した解析には辞書使用の単位や境界認定アルゴリズムを変えても回避できない次の難点がある．

(4) a.複数の構成要素の間に要素の共有 (= 重複) がある場合，それを認識できない．
b.不連続な構成要素を認識できない．

一般用語の解析では (4) は (見かけは大して) 問題にならない．だが，医療用語を始めとする専門用語の語構成解析では深刻な問題が生じる．実例として (4) の制限を取り外した解析を図 1 に示した．このような結果を得る解析法を並列分散形態論解析 (PDMA) と呼ぶ事にする．

unit.id	成語性	要素数	語構成要素	先	天	冠	状	動	脈	異	常	serialized	文字数	不連続
1	1	1	.....異常	0	0	0	0	0	0	1	1	00000011	2	0
2	0.5	1	.....脈.....	0	0	0	0	0	1	0	0	00000100	1	0
3	1	2	.....脈異常	0	0	0	0	0	1	1	1	00000111	3	0
4	1	2	.....動脈.....	0	0	0	0	1	1	0	0	00001100	2	0
5	1	3	.....動脈異常	0	0	0	0	1	1	1	1	00001111	4	0
6	1	1	.....冠状.....	0	0	1	1	0	0	0	0	00110000	2	0
7	1	3	.....冠状動脈.....	0	0	1	1	1	1	0	0	00111100	4	0
8	1	4	.....冠状動脈異常	0	0	1	1	1	1	1	1	00111111	6	0
9	1	1	先天.....	1	1	0	0	0	0	0	0	11000000	2	0
10	1	2	先天.....異常	1	1	0	0	0	0	1	1	11000011	4	1
11	0.5	3	先天.....脈異常	1	1	0	0	0	1	1	1	11000111	5	1
12	1	4	先天.....動脈異常	1	1	0	0	1	1	1	1	11001111	6	1
13	1	5	先天冠状動脈異常	1	1	1	1	1	1	1	1	11111111	8	0

図 1　(2) の PDMA [serialized 列のビット列で降順ソート]

図 1 中の表で，手作業で入力しなければならないのは，i) 成語性，ii) 要素数，iii) 構成性を表わす 0/1 の値のみで，他要素は指定値から自動生成される[4]．

図中の表は (2) の構成要素が次だと規定している．

(5) a.要素数 1: 異常, 脈, 冠状, 先天
b.要素数 2: 先天異常, 動脈, 脈異常,
c.要素数 3: 動脈異常, 先天動脈異常
d.要素数 4: 冠状動脈異常, 先天動脈異常
e.要素数 5: 先天冠状動脈異常 [= 全体]

図中の表を得るための PDMA のアルゴリズム:

(6) a.対象語 w を文字に分割する
b.w を構成する要素ごとに，その部分に 1 を付与し、非部分に 0 を付与する．
c.これらを，より大きな単位を得るように再帰的に結合を繰り返す．が，この際，i) 重複を避けない，ii) 不連続な構成を認める．

PDMA の結果は FCA を使った語構成解析法 [11] と互換である．それを示す結果を図 2 に示した．この Hasse 図は，[. . . 異常], [. . . 脈. . .], . . . , [先天冠状動脈異常] を対象とし，⟨. ∗ 先.∗⟩, ⟨. ∗ 天.∗⟩, . . . , ⟨. ∗ 異.∗⟩, ⟨. ∗ 常.∗⟩ を属性とする形式文脈から生成されたものである．重複している構成要素，不連続な構成要素を含めた構成要素の部分/全体関係が網羅的で体系的に表わされている[5]．

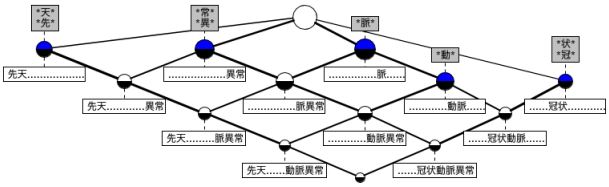

図 2　(1) の該当部分の FCA [より単純な様子は束の上に，より複雑な要素は束の下に配置される]

3　厳密な並列分散解析の近似解

§2 で解説した PDMA は，語構成要素の取りこぼし = 偽負例 (false negatives) の発生を理論的に 0 にできる手法である．それ故に，語構成要素の認識過剰 = 偽正例 (false positives) の発生を心配しなければならない．ただ，偽正例の抑制は教師あり学習で実現できるので，これは PDMA の弱みというより強みだろう．それに対し，偽負例の発生 = 真正例の不足は機械学習でも補う事が難しい．

ただ，PDMA はアノテーションの方法としては非現実的に複雑でもある．そのため，被覆率が 100% に達していなくても，それなりの被覆率で PDMA の結果を近似できる簡便なアノテーション手法があると都合が良い．その手法を次に紹介する．

3.1　構成関係の多重アノテーション

句構造解析 (PSA) では要素の始まりと終わりに標識を付ける．この用途には通常，次のように “[” と “]” の対が使われる．

(7)a. [[動脈] 硬化]　　b. [動脈 [硬化]]

このような記法は頻繁に用いられるものだが，二つ難点がある．第一の難点は，§1 で述べたように，形

4)　要素数の自動算出も理論的には可能だが，Excel の作業シート上で実装するのは手間である．

5)　これは PDMA を解釈した解析束 (parse lattice) で，PSA の解釈である解析木 (parse tree) に対して上位互換である．

態素の重複や不連続性を扱えない事である．第二の難点は，語構成の主従関係と医療用語としての意味的な主要部と修飾部 (あるいは述語/項構造) を正しく反映している保障がない事である．上の (7) の a, b のいずれが妥当な解析なのかは二点目に依存し，一方を選び，それを一貫して使用しなければならない．ただ，この方針を複数の作業者の間で統一するのは，実地の観点からすると難しい．

相良ら [12, 13, 14] の目的は，言語学的な意味での医療用語の語構成を知る事に加えて，専門用語として述語/項構造関係を明示する事である．その理解の下では，上の例で言えば，[硬化] が病名の主要部で，[動脈] が病変の起きている場所 (事態性名詞 [硬化] の一種の項) であるという意味関係も明示したい．このために，医療用語の意味的主要部の標識づけは，言語学的な意味での主要部と概念的に区別し，別の基準で認定されるべきである．

この要件を満足するために，意味的主要部 m_0 から m_n までの遠心拡大構造を “<” と “>” の対を使って i) $<m_n<\ldots<m_0>>>$ や ii) $<<<m_0>\ldots>m_n>$ で標識する[6]．この表記では，“動脈硬化” の解析結果は <動脈 <硬化>> である．同様に，“冠状動脈硬化” の解析結果は (8) である．

(8)<冠状 <動脈 <硬化>>>

この解析は，この病変が <硬化 (症)> の一種で，また <動脈硬化 (症)> の一種である事を記述している．

その一方，解析 (8) は [冠状動脈] という構成要素を取りこぼしている．対応づけありの境界記号が一種類しかない場合，解析精度と被覆率が両立しないジレンマを避けようがない．それを避ける方法は，i) §2 で示したように，境界指定を根本から定義し直すか，ii) 複数次元の構成関係を並列化し，それぞれに別の境界記号を使う事である．二つ目の案の実装例を，**複層化形態論解析 (Multi-Layered Morphological Analysis: MLMA)** と呼ぶ事にする．

境界記号の優先順位を <...>，[...]，{..}，(...) と定めた上であれば，多重構造アノテーションのアルゴリズムは次の通り:

(9) a.Step 1: 用語の (通常右端にある) 主要部 A を <...> で括り出し，それを元に修飾構造を遠心的に再帰的に認定する．
b.Step 2: この境界認定で認識されない構成要素がある場合，A の左側に別の遠心構造の中心 B を見つけ，[...] を使って遠心構造を指定する．

必要に応じて，{..}，(...) を使って Step 1, Step 2 と同様に解析する．現実的には，3 重以上の多重性が必要な場合，§2 で記述した図 1 に例示した PDMA の利用を考える方が無難である

3.2 実例

この解析法を事例 (10) に適用すると，(11) を得る．

(10)冠状動脈硬化症

(11) a.Step 1: < 冠状 < 動脈 < 硬化 < 症 >>>>
b.Step 2: <[冠状 <[動脈 <[硬化]]] < 症 >>>>
c.Step 3: <[{冠状 <[動脈 } <[硬化]]] < 症 >>>>
d.Step 4: (...) で認定すべき要素なし
NB: Step 3 は [{(冠状) <[動脈 } <[硬化]]] < 症 >>>> と等価

(11c) のような MLMA の結果は可読性が低く，そのままでは利用可能性が低い．そのため，このようなアノテーションから構成要素を網羅的に自動抽出する Perl スクリプト[7]を用意した．それで (11c) の構成要素を抽出した結果は次である:

```
## input 1: <[{冠状<[動脈]><[硬化]]]<症>>>>
# A components found with matching 3 pairs of [ and ]
# B components found with matching 4 pairs of < and >
# C components found with matching 1 pairs of { and }
# D components not found: ( and ) missing or mismatching
# summary:
item 1 component  1: 冠状
item 1 component  2: 冠状動脈
item 1 component  3: 冠状動脈硬化
item 1 component  4: 冠状動脈硬化症
item 1 component  5: 動脈
item 1 component  6: 動脈硬化
item 1 component  7: 動脈硬化症
item 1 component  8: 症
item 1 component  9: 硬化
item 1 component 10: 硬化症
-------------------------
```

図 3 (11c) を元にした語構成要素の自動抽出結果

4 PDMA と MLMA と PSA の比較

4.1 比較 1

性能比較の概要を掴んでもらうために，事例 (10) の PMDA の結果と MLMA の結果と PSA の結果を比較する．事例 (10) の MLMA と PSA として，そ

6) 第一著者が調査した限りでは，日本語の意味解析では i) が支配的な構造であり，ii) は “非–”，“脱–”，“前–” のような否定性の接頭辞の作用を記述する時にのみ現われる．i) は右枝分かれ構造であるので，補足があった方が良いだろう．<...> で右枝分かれが支配的なのは，主要部後置型言語では統語構造と語形成論で左枝分かれが支配的なのと矛盾しているように見えるが，ここでは意味的要素の作用域 (一種の依存関係) を記述しており，**厳密には語構成を記述している訳ではない**のが，その乖離の理由である．論を一般化すると，統語解析の結果 (例えば句構造) と意味解析の結果は (多くの期待に反して) 一致するとは限らない．

7) https://github.com/kow-k/MLMA-extractor で公開.

れぞれ <[{(冠状)<[{ 動<[脈 }]< 硬化]]]< 症>>>> と [[冠状 [動脈]][硬化 [症]]] を考え，PDMA が認識するどの要素をそれらが認識するかを評価する．PDMA と MLMA の偽正例を人手除去したものを，PDMA.filtered と MLMA.filtered とする．

component	PDMA	PDMA.filtered	MLMA	MLMA.filtered	PSA	note
.....症	1	1	1	1	1	
.....硬化...	1	1	1	1	1	
.....硬化症	1	1	1	1	1	
.....脈硬化...	1	0	1	0	0	hand filtered
.....脈硬化症	1	0	1	0	0	hand filtered
.....動脈.....	1	1	1	1	1	
.....動脈硬化...	1	1	1	1	0	
.....動脈硬化症	1	1	1	1	0	
冠状.....	1	1	1	1	1	
冠状動脈.....	1	1	1	1	1	
冠状動脈硬化...	1	1	1	1	0	
冠状動脈硬化症	1	1	1	1	1	
異なり数	12	10	12	10	7	
coverage			1.00	1.00	0.70	対PDMA.filtere
					0.70	対MLMA.filtere

図 4　(10) の PDMA.filtered をベースラインとした MLMA と PSA の被覆率の比較

図 4 にある通り，MLMA.filtered は PDMA.filtered の要素を 100%認識するが，PSA は (不連続な要素を含まない事例の解析でも) PDMA.filtered, MLMA.filtered の要素をそれぞれ 70%しか認識しない．

4.2　比較 2

比較 1 と同じ手法で，他の事例で性能比較した結果が図 5 の表である (被覆率は PDMA.filtered の個数に対する割合)．ID は『実践医療用語_語構成要素試案表 Ver. 2』 [6] の行番号，Term はその ID を持つ事例である．ID が空欄の場合，理由を Note に示した．

ID	Term.id	Term	PDMA.count	filtered.count	MLMA.count	filtered.count	PSA.count	PDMA.rate	filtered.rate	MLMA.rate	filtered.rate	PSA.rate1	PSA.rate2	Note
4706	5	大腸癌転症	10	10	10	10	7	1.00	1.00	1.00	1.00	0.70	0.70	
989	5	肝癌骨転移	10	10	8	8	7	1.00	1.00	0.80	0.80	0.70	0.88	
4469	5	細織癌変症	9	7	10	7	5	1.29	1.00	1.43	1.00	0.71	0.71	
5252	5	前代謝障害	6	6	6	6	5	1.00	1.00	1.00	1.00	0.83	0.83	
5549	5	乳がん再発	8	8	8	8	7	1.00	1.00	1.00	1.00	0.88	0.88	s/癌/がん/ の編集あり
1101	5	肝切除術後	10	10	10	10	5	1.00	1.00	1.00	1.00	0.50	0.50	
1004	5	眼球内出血	8	8	8	7	7	1.00	1.00	1.00	0.88	0.88	1.00	
4721	5	大脳萎縮症	12	12	10	10	7	1.00	1.00	0.83	0.83	0.58	0.70	
2862	5	集中治療室	6	6	6	6	5	1.00	1.00	1.00	1.00	0.83	0.83	
	6	大動脈瘤破裂	13	11	14	11	6	1.18	1.00	1.27	1.00	0.55	0.55	1530, 1546, 6347 の共通要素
	6	睡眠薬依存症	10	10	10	10	7	1.00	1.00	1.00	1.00	0.70	0.70	3566 の語末に[]]を追加
6108	7	脾臓静脈吻合術	14	14	13	13	8	1.00	1.00	0.93	0.93	0.57	0.62	
	7	冠状動脈硬化症	12	10	12	10	7	1.20	1.00	1.20	1.00	0.70	0.70	自作例
4719	7	大動脈瘤切除術	21	17	17	17	7	1.24	1.00	1.00	1.00	0.41	0.41	
3949	9	先天性冠状動脈異常	26	26	15	15	12	1.00	1.00	0.58	0.58	0.46	0.80	
2468	11	三尖弁狭窄兼閉鎖不全症	32	32	17	17	10	1.00	1.00	0.53	0.53	0.31	0.59	
		average	12.94	12.31	10.88	10.31	7.00	1.06	1.00	0.97	0.91	0.64	0.71	
		median	10.00	10.00	10.00	10.00	7.00	1.00	1.00	1.00	1.00	0.70	0.70	
		max	32.00	32.00	17.00	17.00	12.00	1.29	1.00	1.43	1.00	0.88	1.00	
		min	6.00	6.00	6.00	6.00	5.00	1.00	1.00	0.53	0.53	0.31	0.41	
		stdev	7.29	7.18	3.50	3.55	1.86	0.10	0.00	0.22	0.15	0.17	0.16	

図 5　PDMA と MLMA と PSA の 16 事例の被覆率の比較

PSA の平均認識率は，対 PDMA.filtered の比較 (PSA.rate1) で 65% 程度，対 MLMA.filtered の比較 (PSA.rate2) で 70% 程度だった．認識率が検索の成功率だと解釈すれば，PSA は明らかに低性能である．

5　議論

5.1　関連研究

PDMA は並列分散型の統語解析 Parallel Distributed Parsing (PDP) [15] を語構成解析に適用したものである．MLMA の方法論は Autolexical Syntax (AS) [16] と類似している．ただし類似性は表面的なものに過ぎない．AS の想定は統語論と形態論の部門単位の並列化であり，形態論内，あるいは統語論内の解析の並列化ではない．

5.2　過剰認識の問題と機械学習との関連

図 1 に例示した解析は，作成に手間がかかる．その理由で，性能が劣っていても MLMA のような簡易版や PSA の方が望ましいと思う人がいるかも知れないが，そうではない．PMDA でない解析では偽負例を原理的に回避できず，語構成解析が用語抽出だと考えた場合，欠点となる．

PDMA は逆に，解析で偽正例が生じる可能性を心配しなければならない程，解析精度は高い．偽正例を回避する手順の本質は成語性の問題であり，専門的知識がない者にこれが容易に実現できないのは仕方がない．ただ，成語性の判定は，候補の列挙よりは遥かに楽な作業である．

真正例を取りこぼさず，偽正例の個数を専門知識を持った者による事後検証によって 0 に近づける事ができたとすると，図 1 中の表にあるような解析結果は，複合語 c が与えられた時の c の構成要素認識の正例と負例の混合となる．この特性を利用すれば，機械学習 (例えば Deep Learning) を使って，専門用語の構成要素認識を自動化できる見込みは十分にある．PDMA は確かに遂行にそれなりの手間はかかるが，それに見合った性能を備えている．

5.3　専門用語の語構成は特別な例外か？

専門用語の語構成論には並列分散解析が必要かつ有用という主張が本稿の骨子であるが，次の疑問が発展的に生じる．Q1. 非専門用語の語構成論に並列分散解析は不要なのか？ Q2. 並列分散解析が必要なのは，語構成論に限定されるのか？

Q1 への答えが否なのは，専門用語と非専門用語の境界が曖昧である事から明らかである．非専門用語で並列分散解析の必要性が低いのは，語の内部構造の複雑度が専門用語に較べて低いからである．

Q2 への答えは Q1 への答えに関係している．内部構造の複雑度が並列分散解析の必要性を増すのであれば，文の規模でも要素が並列分散解析が必要な複雑度を持っていると考えるのが自然である．それが不要に見えるのは，主流言語理論に目を晦まされ，現実の複雑性を見損なっているからではないか？

謝辞

本研究は JSPS 科研費 JP21H03777 の助成を受けたものである．

FCA は Concept Explorer 1.3 (http://conexp.sourceforge.net) で実行した．

参考文献

[1] 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, 第 6 巻, pp. 115–124. 1994.

[2] 山田恵美子, 松本裕治. 専門用語の内部構造解析. 言語処理学会第 15 回年次大会発表論文集, pp. 340–343, 2009.

[3] 山田恵美子, 松本裕治. 文字係り受けに基づく専門用語の内部構造表現と解析. 研究報告音声言語情報処理 (SLP), Vol. 2009, No. 20, pp. 1–6, May 2009.

[4] 内山清子, 岡照晃, 東条佳奈, 小野正子, 山崎誠, 相良かおる. 実践医療用語の語構成要素抽出の試み. 言語資源活用ワークショップ発表論文集, 第 3 巻, pp. 463–467. 国立国語研究所, 2018.

[5] 麻子軒, 黒田航, 相良かおる, 東条佳奈, 西嶋佑太郎, 山崎誠. 実践医療用語における語構成要素の結合順序に関する量的調査. 計量国語学会第 66 回大会発表論文集, 2021.

[6] 東条佳奈, 黒田航, 相良かおる, 高崎智子, 西嶋佑太郎, 麻子軒, 山崎誠. 実践医療用語_語構成要素語彙試案表 ver.2.0 の構築. 言語資源ワークショップ, 2022.

[7] Andrew Spencer. Bracketing paradoxes and the English lexicon. **Language**, Vol. 64, pp. 663–682, 1988.

[8] Charles C. Fries. Meaning and linguistic analysis. **Language**, Vol. 30, No. 1, pp. 57–68, 1954.

[9] Kow Kuroda. "Pattern Lattice" as a model for linguistic knowledge and performance. In **Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Vol. 1**, pp. 278–287, 2009.

[10] 黒田航, 長谷部陽一郎. Pattern Lattice を使った (ヒトの) 言語知識と処理のモデル化. 言語処理学会第 15 回大会発表論文集, pp. 670–673, 2009.

[11] 黒田航, 相良かおる. 医療用語の is-a オントロジー構築の FCA を使った効率化. 言語処理学会第 28 回年次大会発表論文集, pp. 705–709, 2022.

[12] 相良かおる. 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, 2021.

[13] 相良かおる, 山崎誠, 麻子軒, 東条佳奈, 小野正子, 内山清子. 実践医療用語の語構成要素: 意味を基準とした分割. じんもんこん 2019 論文集, 2019.

[14] 相良かおる, 小野正子, 高崎智子, 東条佳奈, 麻子軒, 山崎誠. 実践医療用語の語構成と意味: 語構成要素語彙表試案表の作成にむけて. じんもんこん 2020 論文集, 2020.

[15] Kow Kuroda. Arguments for *Parallel Distributed Parsing*: Toward the integration of lexical and sublexical (semantic) parsings. In **Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation**, pp. 455–462. Institute of Digital Enhancement of Cognitive Processing, Waseda University, 2010.

[16] Jerrold M. Sadock. **Autolexical Syntax: A Theory of Parallel Grammatical Representations**. University of Chicago Press, 1991.

A　重複のサンプルデータ

表 1　見本中の重複の有無 (抜粋)

ID	MLMA	<, >	[,]	{, }	(,)	重複
56	<[[アミノ酸<欠乏][性]]<貧血>>>	3	3	0	0	1
474	<壊疽 [性]<[歯<肉]<炎>>>>	4	2	0	0	1
642	<外傷 [性]<脊髄<出血>>>	3	1	0	0	0
676	<[外側<半月 [板]]<障害>>>	3	2	0	0	1
690	<[外転<神経]<麻痺>>>	3	1	0	0	1
879	<[{肩<[関節}[後方]]]<脱臼>>>	3	3	1	0	1
913	<[カテーテル<感染]<症>>>	3	1	0	0	1
989	<[肝<癌]<骨<転移>>>>	4	1	0	0	1
991	<[[肝 [機能]]<[検査]]<異常>>>	3	4	0	0	1
1004	<[眼球<[内]<出血>>>	3	1	0	0	1
1101	<[{肝<[切除}<術]]<後>>>	4	2	1	0	1
1170	<[完全<<埋伏]>歯>>	3	1	0	0	1
1172	<[完全<[無] 歯]<症>>>	3	2	0	0	1
1248	<顔面<多発<刺<創>>>>	4	0	0	0	0
1307	<[{気<[管}<内]]<出血>>>>	4	2	1	0	1
1308	<[{気<[管]<内]]<洗浄>>>>	4	2	1	0	1
1309	<[気管 [内]<注] 入>>	2	2	0	0	1
1332	<[基礎<代謝]<率>>>	3	1	0	0	1
1405	<急 [性]<腎盂<腎<炎>>>>	4	1	0	0	0
1514	<[胸 [椎]<破裂]<骨折>>>	3	2	0	0	1
1601	<菌 [状]<[息<肉]<腫>>>>	4	2	0	0	1
1921	<[幻覚<薬]<依存>>>	3	1	0	0	1
2074	<[[口][唇]<麻痺>>	2	3	0	0	1
2151	<[{後<[頭}<神経]]<ブロック>>>>	4	2	1	0	1
2469	<[{三尖<[弁}<狭窄]]<症>>>>	4	2	1	0	1
2562	<[趾<関節]<損<傷>>>>	4	1	0	0	1
2760	<[疾病<恐怖]<症>>>	3	1	0	0	1
2775	<自発<運動>>	2	0	0	0	0
2788	<[社会<恐怖]<症>>>	3	1	0	0	1
2797	<若年 [性]<子宮<機能<出血>>>>	4	1	0	0	0
2848	<[銃<自殺]<未遂>>>	3	1	0	0	1
2862	<[集中<治療]<室>>>	3	1	0	0	1
2903	<[{手<[指}<知覚]]<異常>>>>	4	2	1	0	1
2912	<[{手術<[創]<[部]]]<膿瘍>>>	3	3	1	0	1
3130	<[小<脳]<萎縮>>>	3	1	0	0	1
3157	<[{{上<[腕}<筋}]]<挫<傷>>>>>	5	2	2	0	1
3201	<食事<療<法>>>	3	0	0	0	0
3239	<処置 [後]<腎<不全>>>	3	1	0	0	0
3367	<新生 [児]<壊死 [性]<腸<炎>>>>	4	2	0	0	0
3391	<新生 [児]<点 [状]<出血>>>	3	2	0	0	0
3438	<[身体<発育]<遅滞>>>	3	1	0	0	1
3479	<[膵<移植]<不全>>>	3	1	0	0	1
3513	<[[膝 [体]<[尾 [部]]]<<腫}>瘍>>>	4	4	1	0	1
3523	<[水分<欠乏]<症>>>	3	1	0	0	1
3566	<[睡眠<薬]<依存>>>	3	1	0	0	1
3567	<[睡眠<薬]<自殺>>>	3	1	0	0	1
3599	<[精索<狭窄]<症>>>	3	1	0	0	1
3619	<生殖<器>>	2	0	0	0	0
3731	<[{脊<[髄}<膜]]<結核>>>>	4	2	1	0	1
3763	<[[舌] 咽<神経]<損<傷>>>>	4	2	0	0	1
3820	<[線維<脂肪]<肉<腫>>>>	4	1	0	0	1
3921	<先天 [性]<横隔 [膜]<ヘルニア>>>	3	2	0	0	0.5
4077	<先天 [性]<水晶 [体]<<偏>位>>>	4	2	0	0	0
4119	<先天 [性]<大葉 [性]<肺<気<腫>>>>>	5	2	0	0	0
4349	<[{{前 [腕}<軟 [部]]]<<腫}>瘍>>>	4	3	2	0	1
4357	<[臓器<移植]<法>>>	3	1	0	0	1
4553	<{帯状 [疱疹}[性]]<髄膜<脳<炎>>>>	4	2	1	0	1
4560	<[大<[静<脈]]<先天<異常>>>>	5	2	0	0	1
4695	<[[大<腸]<癌]<検診>>>>	4	2	0	0	1
4895	<[[胆<管]<狭窄]<症>>>>	4	2	0	0	1
4970	<[窒素<酸化]<物>>>	3	1	0	0	1
4991	<中隔 [性]<肝<硬<変>>>>	4	1	0	0	0
5153	<調節 [性]<斜<視>>>	3	1	0	0	0
5240	<適応<行動>>	2	0	0	0	0
5255	<手<熱<傷>>>	3	0	0	0	0
5537	<二次 [性]<糖尿<病>>>	3	1	0	0	0.5
5582	<乳房<肥大>>	2	0	0	0	0.5
5747	<{{{{脳}室}[内]}<<腫}>瘍>>	3	1	4	0	1
5805	<[背<筋]<挫<傷>>>>	4	1	0	0	1
6041	<非感染 [性]<[空<腸]<炎>>>>	4	2	0	0	1
6121	<[非{代償}[性]]<肝<硬変>>>	3	2	1	0	0
6301	<複合<母斑>>	2	0	0	0	0
6343	<腹 [部]<損<傷>>>	3	1	0	0	0
6416	<[{{分娩<[後}遺]<症>]]<尿道<狭窄>>>>	4	3	2	0	1
6612	<[哺乳<力]<低下>>>	3	1	0	0	0.5
6819	<[[網<膜]<深層]<出血>>>>	4	2	0	0	1
6855	<[薬剤 [性]<糖 [尿]]<病>>>	3	3	0	0	1
6925	<腰 [椎]<開放 [性]<<脱臼><骨折>>>>	5	2	0	0	1
6979	<[[卵<管]<結紮]<術>>>>	4	2	0	0	1
7031	<リウマチ [性]<舞踏<病>>>	3	1	0	0	0

表 1 は調査に用いたサンプル 99 のうち，ページに収まる 81 例を示した．

ID は『実践医療用語_語構成要素試案表 Ver. 2』の行番号 (= ID).

重複の値が 1 のものは確実に重複ありの事例，0.5 のものは <...> の設定次第で重複ありの事例，0 のものは重複なしの事例．値が 1 事例の割合は 0.74% (= 73/99)．値が 0 より大きい事例の割合は 0.78% (= 77/99).

B　FCA の概要

B.1　FCA は何をする道具か？

FCA は属性によって定義された対象を自動分類するためのアルゴリズムの一つである．それは概念 (concept) を次のように形式化する事で実現される．

(12)　a.(形式的) 概念 c とは，外延 o と内包 a の対である (記号で書けば，$c := (o, a)$).

a′.ただし，外延 o と内包 a はそれぞれ，対象の全体集合 O と属性の全体集合 A の部分集合とする．

b.c_i と c_j は，〈一方が他方を含む〉という関係について順序構造をなし，c_i, c_j を要素にもつ概念の全体集合 C に束構造 (lattice) がある．

要するに，FCA は，概念を (12a) のように (数学的に) 定義した上で，概念間の関係を (12b) が定義する束構造として明示化する手順である．

もっと一般的な理解では，対象の集合 $O = \{o_1, \ldots, o_n\}$ があり，それらを記述する属性の集合 $A = \{a_1, \ldots, a_m\}$ が (暫定的に) 定められた時，O と A の直積 $O \times A$ に真理値を割り当てる．これが表現する状態を離散的に自動分類するためのアルゴリズムの一つが FCA である．

B.2　FCA の実行例

図 6 は形式文脈として表 2 を与え，結果として得られる Hasse 図である．

表 2　太陽系の惑星の FCA を使った分類

	large	medium	small	near	far	satellites	no satellites
Mercury	1	0	0	1	0	0	1
Venus	1	0	0	1	0	0	1
Earth	1	0	0	1	0	1	0
Mars	1	0	0	1	0	1	0
Jupiter	0	0	1	0	1	1	0
Saturn	0	0	1	0	1	1	0
Uranus	0	1	0	0	1	1	0
Neptun	0	1	0	0	1	1	0
Pluto	1	0	0	0	1	1	0

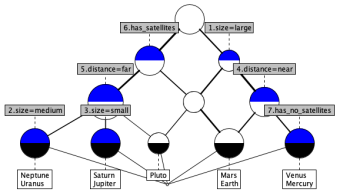

図 6　(2) を形式文脈として構築した Hasse 図