

## 医療用語に含まれる序数詞について

東条佳奈<sup>1)</sup> 相良かおる<sup>2)</sup> 西嶋佑太郎<sup>3)</sup> 麻子軒<sup>4)</sup> 山崎誠<sup>5)</sup>

1) 大阪大学 2) 西南女学院大学 3) 医師 4) 関西学院大学非常勤講師 5) 国立国語研究所

**概要**：本発表は、医療記録データにおける数量表現の適切な抽出のために、医学分野の用語集・辞典と医師国家試験問題文より抽出した「第」を冠する序数詞について「数詞の表記のゆれ」と「名義尺度と順序尺度のどちらにあたるのか」の2点に注目して分析を行ったものである。調査の結果、立項される序数詞においては、同じ見出し語であっても数詞の表記法にゆれがあること、算用数字は順序尺度が多く名義尺度が少ないこと、ローマ数字は名義尺度が多いことなどを示した。ただし、序数詞の特性上、尺度の判別に迷うものも多かった。今後、表記のゆれを所与とし、医療記録データから数量表現を的確に抽出するためには、序数詞だけではなく、数詞と助数詞、および隣接する名詞との関係を明らかにする必要がある。

**キーワード**：医療用語、序数詞、表記の揺れ、名義尺度、順序尺度

### Analysis of ordinal numerals in medical terminology

Kana Tojo<sup>1)</sup> Kaoru Sagara<sup>2)</sup> Yutaro Nishijima<sup>3)</sup> Tzu-Hsuan Ma<sup>4)</sup> Makoto Yamazaki<sup>5)</sup>

1) Osaka University 2) Seinan Jo Gakuin University 3) physician

4) Kwansai Gakuin University 5) National Institute for Japanese Language and Linguistics

**Abstract**: In this study, for appropriate extraction of quantitative expressions from medical record data, we analyzed ordinal numerals including the Japanese character “第 (dai)” extracted from glossaries and dictionaries used in the medical field and the National Medical Examination for Doctors, focusing on two points: fluctuations in the notation of numerals and whether the numerals correspond to nominal or ordinal scales. The results of the survey showed that (1) there are variations in the notation of numerals in ordinal numerals, even for the same headword; (2) arithmetic numerals are most often on ordinal scales but sometimes on nominal scales; and (3) Roman numerals are most often on nominal scales. However, due to the characteristics of ordinal numerals, there were many cases where it was difficult to determine the type of scale. In the future, it will be necessary to clarify the relationships between the scale and not only ordinal numerals but also numerals, particles, and adjacent nouns to accurately extract quantitative expressions from medical record data, given the variation in notation.

**Keywords**: Medical terminology, Ordinal numeral, Orthographic variation, Nominal scale, Ordinal scale

### 1. はじめに

電子カルテシステムの普及により蓄積された医療記録データから、利用者が求める情報を適切に検索する技術が益々重要になっている。医療記録データには数値データが多く含まれており、目的に沿ってこれらの数量表現を適切に抽出する必要がある。医療記録は当然、その書き手や医療機関ごとにゆれが生じるものとなるが、それは数量表現の使用においても同様である。数量表現におけるゆれとしてまず、数の表記の問題が挙げられる。日本語では数の表記は漢数字・アラビア数字(以下、算用数字とする)・ローマ数字の3種に分けられるが、一般に、文書中に現れる整数、実数、分数などの量的データに漢数字やローマ数字が使われることはない。従って、誤植・誤変換を除き表記のゆれは質的データに限られると考えられる。

また、医療記録は医療従事者によって作られており、そこで使われる表現、すなわち専門用語などの知識の習得には、専門の教科書、辞書、国家

試験などと関わりがあると考えられる。しかし、医学用語の標準化については様々な事情から課題があり[1]、多くの専門学会が出版している専門用語集においても、学会相互での用語の標準化はなされていないのが現状である。そのため、医療記録を解析するためにまず、辞書や国家試験など規範・教育的な出版物である書籍を対象に、どのようなゆれが、どの程度あるかといった調査を行った上で、経過記録と照合する必要がある。

加えて、日本語では、ものの数量を表す際、数詞に続いて助数詞が示される。特に狭義の助数詞である類別詞は数える対象の特性に関わるため、医療分野に特有の助数詞も、数量表現抽出の際には把握しておく必要がある。しかし、助数詞自体は接辞であり、単独では辞書には立項されていない可能性が高い。従って、数量を含む語として一語化している語から収集する必要がある。

そこで、本発表では、一語化して辞書に立項されている質的データを示す数量表現として、序数詞に着目し、数詞に冠して物の順序を数える接頭辞「第」を伴う序数詞(「第2期」「第Ⅲ因子」「第

二水俣病」など)について、複数の用語集、辞典類と数年分の国家試験問題を対象に実態の調査を行った。その上で、「数の表記法」と「序数詞を含む語の尺度」の二点に注目して分析を行い、その結果をもとに、医療経過記録の自然言語処理を行う上での留意点を示した。

以下、前提となる事柄として用語の説明と表記規則の現状の概観(2節)に触れ、3節にて調査方法と資料、4節に調査と分析の結果、5節に4節で示したそれぞれの結果についての考察を述べ、6節にてまとめと自然言語処理を行う上での影響について示し、今後の課題を述べる。

## 2. 前提となる事柄

### 2.1 用語の説明

#### 2.1.1 序数詞について

数を表す語群は広義に数詞と呼ばれる。広義の数詞は、日本語では多くの場合、「数の概念を表す語(本数詞、狭義の数詞)+数えられる対象の種類を表す語(助数詞)」の形式で用いられる。ここでいう助数詞は、数詞を構成する接辞である。数詞のうち、個数・数量・度数を表すものを基数詞、順序・位置づけを表すものを序数詞(順序数詞)という[2]。序数詞には、特有の助数詞として「第-」や「-目(め)」がある。基数詞であってもこれらの助数詞を付加すれば、序数詞となる。また、基数詞の中には、個数だけでなく順序数も同形で示しうる語があるため、数詞だけでは基数詞か序数詞か判断がつかないものもある[2]。

「第-」がつく序数詞は、計数意識が順次話し手の意識に上りやすい「-目」に比べ、計数意識を明確に感じさせないという相違点があるという[3]。例えば「第3冊」といった場合、「第1冊」「第2冊」といった表現がそれぞれの書物に予め割り当てられており、はじめから順序付けられているもの、という違いがあることが指摘されている。

本発表では、「第1子」のうち、「第」を接辞、前述の本数詞にあたる「1」を「(狭義の)数詞」または「数字」、「子」を「助数詞」、「第1子」全体を「序数詞」と扱う。なお、データの中には、「第1子妊娠」のように、序数詞(「第1子」)にさらに後部要素(「妊娠」)が付加しているものが見られるほか、「第四脳室」に対して「第四脳室蓋」がある場合の「脳室蓋」など、一つの助数詞と捉えるべきか、助数詞と隣接語に分けるべきか判断がつかない語もあったが、これらについても除外せず、隣接語と思われるものも含めて一括して「序数詞」として捉えた。

#### 2.1.2 尺度について

数字で表現されるデータは、質的データと量的データの二つに分けられ、更に質的データは①順序に意味をもたない名義尺度(「性別」「疾病名」

など)と、②順序に意味を持つ順序尺度(病気の経過「改善」「不変」「悪化」など)に分けられる。また量的データは③差のみが計算可能な間隔尺度と、④比と差の両方が計算可能な比尺度に分けられ、それぞれ統計処理の方法が異なる[4]。

### 2.2 数字の表記規則

『新しい国語表記ハンドブック第8版』(2018)によれば、横書きの日本語文書における数の書き表し方として、文部省の「文部省 公文書の書式と文例」を引用し、算用数字(アラビア数字)を用いることを原則とするとしている。例外的に「一般、一部、一時保留」など数の感じが少なくなった場合や、「ひとつ」「ふたつ」と読むなどの場合(和語系数詞の場合)、「億」や「万」などの位取りの単位、概数を示す場合には漢数字を用いて表すことが示されている[5]。

縦書き媒体となる新聞の規則においても、大きくは変わらない。『朝日新聞の用語の手引』(2002)では、日時・年齢・金額・数量などの数字の表記は原則として算用数字とし、慣用句・成句・専門用語・固有名詞に含まれる数字、ひとつ・ふたつ、万以上の単位語、あいまいな数の表記、読み誤る恐れがある場合は漢数字で表すことが定められている[6]。『毎日新聞用語集』(2002)においても、データの数字、計量の対象となる数字、ほかの数字に置き換えても文意が通じるものは算用数字で書く一方、数値をあいまいにした表現や、固有名詞・熟語・慣用句、数字が特別の意味合いをもつ場合、四大文明・五大陸・第二次世界大戦など、数字を含む語の中身が具体的に明らかなものや、伝統的な日本文化にかかわるものは漢数字で書く、と定められている[7]。単に数量を示しているのではなく、語の一部として意味を持つ数字は漢数字で表記するといえる。

医療分野においては、経過記録を含めて横書きすることが大半であり、縦書きの機会はほとんどない。そのため書字方向による表記差異は問題となりにくい。それ以外の数字の表記規則として分野全体でコンセンサスを得ている原則は見られなかった。一部の用語集の凡例に数字の表記について記載があるものがあり、『日本医学会医学用語辞典』では、「慣用もあるので、用語によって表記を使い分けた」とする。例として「一次免疫応答」「1日分泌量」「I型肺胞細胞」といった漢数字、算用数字、ローマ数字を使用する語が挙げられ、このうち一種を原則とはしていなかった。なお、同凡例には「血漿中の血液凝固因子(I~XIII)は国際符号としてローマ数字を用いる」こと、「数字の概念から外れ、1つの熟語的用語となったものは漢数字を用いる」こと(例:十二指腸、百日咳など)が記されている[8]。ほかに『文部科学省学

術用語集医学編』の凡例には「数を数えるという概念が強調されるものには、アラビア数字を用いる」、「英語の中に序数詞に用いられたローマ数字は、『第』を冠してアラビア数字又は漢数字で書き換える」とある[9].

### 3. 調査方法と資料

#### 3.1 言語資源

- (1) **用語集**：『日本医学会医学用語辞典』Web版(71,067語)[8]と『文部科学省 学術用語集 医学編』(約1万9千語)[9]の見出し語
- (2) **辞典**：『医学書院 医学大辞典 第2版』(約10万語)[10]と『南山堂 医学大辞典 第20版』(約4万語)の見出し語[11]
- (3) **国試問題文**：医師国家試験問題5回分(2001年, 2006年, 2011年, 2016年, 2021年)[12]

なお、医学用語を収録する用語集・辞典は数多くあるが、本研究では、その性格から(1)用語集、(2)辞典のように仮に名をつけ、二つに分けて調査した。(1)用語集とは、日本医学会や文部科学省が中心となって作成した「標準化」が目的に含まれるものを指し、(2)辞典とは、医学系出版社が編纂した「網羅性」をうたったものを指す。

#### 3.2 調査方法

(1)用語集、(2)辞典、(3)国試問題文より、接辞「第」を含む「第1回」「第III因子」「第九条」などの序数詞を抽出した。数詞の種類(算用数字、ローマ数字、漢数字)および尺度ごとに集計し、数詞と尺度の関係についてクロス集計表を作成した。また各グループについて表記のゆれがあるものを列挙した。

### 4. 調査結果

#### 4.1 数詞の表記ごとの集計

表1に、各言語資源から抽出した序数詞を数詞ごとに集計したものを、延べ語数、異なり語数に分けて示した。(1)用語集では174語、(2)辞典では253語が立項されており、(3)国試問題文には70語が見つかった。これらを数詞の表記種別(算用数字、漢数字、ローマ数字)ごとにまとめて異なり語数を調べると、用語集は88語、辞典は110語、国家試験問題は40語であった。更に表記種を考慮せずに数詞としてまとめた場合の異なりでは、用語集71種、辞典81種、国試問題41種であった。(1)用語集と(2)辞典を比べると、延べ語数、異なり語数ともに(2)辞典のほうが多く、総収録語数の大小と一致していた。

使用した言語資源はいずれも横書きであったが、いずれの資源においても算用数字、漢数字、

ローマ数字の3種が見られた。その内訳をみると、(1)用語集では算用数字の割合が半数以上を占め、(2)辞典では漢数字が半数近くを占めているという違いが見られた。(3)国試問題文は、割合から見れば(1)用語集に近いものであった。

表 1. 数詞ごとの集計

	(1) 用語集	(2) 辞典	(3) 国試問題文
算用数字	115   66%	66   26%	59   84%
漢数字	32   18%	124   49%	8   11%
ローマ数字	27   16%	63   25%	3   4%
延べ語数 計	174   100%	253   100%	70   100%
算用数字	61   69%	31   28%	31   78%
漢数字	23   26%	65   59%	7   18%
ローマ数字	4   5%	14   13%	2   5%
異なり語数 計	88   100%	110   100%	40   100%

#### 4.2 表記のゆれ

表2に、抽出した語のうち、各グループ内で表記のゆれがあったものを尺度と共に示した。「漢」は漢数字、「ロ」はローマ数字、「算」は算用数字、☆は数詞が入る箇所を表す。

表 2. 表記のゆれと尺度

	序数詞	ゆれ	尺度
用語集	第☆因子	ロ・算	名義
	第☆因子欠乏症	ロ・算	名義
	第☆斜位	漢・算	名義
	第☆性病	漢・算	名義
	第☆極体	漢・算	順序
	第☆胃	漢・算	順序
辞典	第☆脳神経	漢・算	名義
	第☆胎向	漢・算	名義
	第☆経路	漢・算	名義
	第☆度房室ブロック	ロ・算	順序
	第☆回旋	漢・算	順序
	第☆頸椎	漢・算	順序
	第☆級動揺	漢・算	順序
	第☆減数分裂	漢・算	順序
	第☆度無月経	漢・算	順序
	第☆次硝子体過形成遺残	漢・算	順序
	国試問題文	第☆選択	漢・算
第☆肋間胸骨左縁		漢・算	順序

(1)用語集のうち、『日本医学会医学用語辞典』では「第1極体」と「第二極体」、「第1斜位」と「第二斜位」のように同じ助数詞でも表記にゆれが生じているものがあつた。さらには「第4性病：inguinal lymphogranuloma」「第四性病：fourth venereal disease」のように同一の疾患を表すものであつても、二つの表記が立項され、それぞれ紐づけされる英語用語も異なってしまう例もあつた。血液凝固因子については、『文部科学省学術用語集医学編』の凡例で「英語の中に序数詞に用いられたローマ数字は、『第』を冠してアラビア

数字又は漢数字で書き換える」とあるため、日本医学会との間でゆれが生じていた。これらを除くと両方で表記がゆれるものはなかった。

(2)辞典では、医学書院で「第1度房室ブロック」「第2度房室ブロック」のようにゆれが見られるものもあったが、同一辞典内のゆれは少なく、多くは辞典間でのゆれであった。「脳神経」は南山堂でローマ数字、医学書院で算用数字を用い、「回旋」は南山堂で算用数字、医学書院で漢数字、「頸椎」は南山堂で漢数字、医学書院で算用数字というように、それぞれの辞典でどちらかの表記に偏るわけではなく傾向は見いだせなかった。

(3)国試問題文では、同じ年度の問題文の中に「第1選択」「第一選択」(第105回)、「第4肋間」「第二肋間胸骨左縁」(第115回)のようにゆれが見られ、それ以外のゆれは見られなかった。

### 4.3 尺度ごとの集計

表3に、序数詞を順序尺度か名義尺度かに二分し、集計したものを示した。

表3. 尺度ごとの集計

	(1) 用語集	(2) 辞典	(3) 国試問題文
順序	89   51%	107   42%	61   87%
名義	85   49%	146   58%	9   13%
延べ語数 計	174   100%	253   100%	70   100%
順序	50   57%	60   55%	32   80%
名義	38   43%	50   45%	8   20%
異なり語数 計	88   100%	110   100%	40   100%

各語が順序尺度か名義尺度のいずれに該当するかは、その後に含まれる数字が「順序を意味するか否か」について医師1名が判断した。「第1頸椎」「第2頸椎」などは頸椎(首にある脊椎の骨)を頭側から順に呼んでおり、順序尺度である。「第1頭位」や「第1胎向」は胎児の背中が母体の左右どちらかにあるかを示す表現であり[11]、名義尺度である。しかし、ある数字が順序を意味するか否かは截然と区別できるものではなく、判断に悩む例がいくつか見られた。以下に例を挙げる。

第I因子から第XIII因子までである血液凝固因子は、発見された順番に命名され、1950年代にローマ数字で表記することが国際的に決まったものである[13]。元来、順序を意味しているものの、第VI因子は欠番となっており、また第I因子はフィブリノゲン、第II因子はプロトロンビンのように別名で呼ばれることが多い[11]。そのため医療の場面では、順序を意識して使用することがまざらないと思われるため、名義尺度とした。

左右12対ある脳神経は、その起始部が脳の前方にあるものから後方へ順に、第I/I脳神経などと呼ばれる。しかし各脳神経は機能面から嗅神経、視神経、動眼神経のように別名があり、病名では

「顔面神経麻痺」(顔面神経は第7/VII脳神経にあたる)のように数詞を用いる表現は使われない。起始部の位置による順番は意識されていないことから、現在の使用法として順序に意味はないと考え、名義尺度に分類した。

このように、命名当初は順序を意味したものであっても医療用語として使用される場面で順序の意味では使われないものがあり、これらを名義尺度とした。全体として順序尺度のものが多いものの、名義尺度とおおよそ拮抗するほどであった。また、(1)用語集と(2)辞典とで尺度の割合に大きな差は見られなかった。

### 4.4 数詞と尺度の関係

表4は、(1)用語集と(2)辞典より抽出した序数詞の異なり語171語の数詞と尺度のクロス集計表である。左の列に観測値を右の列に期待値を示している。期待値より、漢数字や算用数字に比べてローマ数字に有意な差がありそうである。また期待値に5未満の値がないことからコクランの規則によりカイ二乗検定が適用できるが、ここではフィッシャーの正確確率検定によるP値を求めた。その結果、ローマ数字、漢数字、算用数字の3群間で比率が有意に異なることが分かった。

表4. 数詞と尺度の関係

観測&期待値	算用数字	漢数字	ローマ数字	計
順序	55   46.53	45   46.53	2   8.95	102
名義	23   31.47	33   31.47	13   6.05	69
計	78   78	78   78	15   15	171

注)Fisherの正確確率検定(両側)でのP値=0.0001389

ローマ数字に有意な差があることは期待値から想像に難くない。しかし、算用数字の場合は、期待値と比べて順序尺度は若干少なく、名義尺度は多くなっているものの有意な差か否かは分からない。そこで、残差分析を行った。その結果、各群のP値より、ローマ数字だけでなく、算用数字においても有意差が認められた(表5)。

表5. 観測値と残差分析

	算用数字	漢数字	ローマ数字
順序	55 ▲	45	2 ▽
名義	23 ▽	33	13 ▲
P値	0.008	0.6329	0.0001

注)▲有意に多い, ▽有意に少ない, p<.05

## 5. 各項目についての考察

### 5.1 数詞の表記ごとの集計について

(1)用語集は上述の通り、表記は慣用によるしつつ、おおまかな方向性を凡例に示している。一方(2)辞典では、数字表記についての記載は見られず、表記一般について「異なる表記が併存してい

る場合は、慣用を重視しつつも、歴史的な経緯や国語学的な視点を加味し、本辞典独自に整合・統一を行った」（『医学書院医学大辞典』）とあり、こちらも慣用を重視している点で(1)用語集と方向性の大きな違いはない。そのうえで数詞の割合に差が生じたのは、(2)辞典の編集時に漢数字に統一されるケースが多かったことが推測される。

(3)国試問題文については、『医師国家試験出題基準(平成30年度版)』を参照しても、検査項目の表記について例示がなされているほかは、使用する語の基準についての記載はなく、表記の整理統一についてどの程度の配慮がなされているか判断としない。算用数字、漢数字、ローマ数字のいずれも見られ、かつ割合も(1)用語集と大きく隔たりがないという点で、(3)国試問題文も慣用におおよそ則っていることが推察される。

## 5.2 表記のゆれについて

医療用語のうち医学用語は、(1)用語集の二冊によって標準化が図られている。『文部科学省学術用語集医学編』は日本医学会が作成に参加していたこともあり、両者の表記の傾向は似通ったところが多かった。しかし表2に示したようなゆれも存在し、特に『日本医学会医学用語辞典』内部のゆれが存在していることが明らかになった。これは紙媒体の英和第3版からWeb版に移行した際に、これまでの不統一が顕在化したものと思われるが、さらなる校正作業が必要と思われる。

(3)国試問題文も同様に、内部での不統一のみであるため、さらなる校正作業が必要と思われた。

(1)用語集と(2)辞典の間でのゆれの調査は行っていないが、少なくとも同じく慣用に従っている(2)辞典の間に表記のゆれが表2のように見られた。これは、同じ語であっても出版社によって整理統一する方向が異なるほど、序数詞の表記統一の困難さを示しているといえる。

## 5.3 尺度ごとの集計について

上述のように、もとは順序に意味を持たせて命名されたと思われる語でも、使用場面から考えて、名義尺度と判断した語がいくつもあった。

こうした語の特徴の一つとして別名の存在がある。血液凝固因子や脳神経の場合、序数詞以外の別名が存在し、序数詞と同等かそれ以上に使用される。別名がある場合、別名の表す機能や形態の情報が数詞の順序の情報よりも重要視されうると考えられる。

もう一つの特徴として、欠番の存在がある。血液凝固因子は第VI因子が欠けており、ほかに「脳室」は「第1脳室」「第2脳室」がなく「第3脳室」以降が存在する。欠番となった用語は過去に存在したものの、別概念に統合されたか、別名で

呼称されるようになったものである。医学の知見が蓄積された結果と思われるが、これによって序数詞が不連続なものとなってしまい、順序としての意味は見出しにくくなる。

ただ、用語集や辞典への収録の有無のみでは、欠番の判断や名義尺度か否かの判断は困難である。例えば順序尺度と判断した「頸椎」は第1から第7までであるが、(2)辞典ではどちらも「第1頸椎」「第2頸椎」「第7頸椎」のみを立項している。これらは他と比べて形態が異なる、あるいは体表から触知できるという特徴を持ち、それぞれ環椎、軸椎、隆椎という別名を持つ。用語集や辞典の特性上、順序尺度で表される語のうち、特徴的なもののみ立項していることも考えられる。

こうした特徴を持つ語を名義尺度と判断したこと、そして順序尺度の特徴的なもののみが用語集・辞典に収録される可能性があることで、相対的に全体に占める順序尺度の割合が低く出ている可能性も考えられる。

なお、奥津[14]は序数詞について、指し示す対象が一つしかなく、基数詞が備える副詞的性質を持たないことから、普通名詞や固有名詞と同様に、事物そのものを表わしているものと捉えるべきだと指摘している。名義尺度か順序尺度かの判断が困難なことのひとつに、序数詞のもつこうした性質が関わっている可能性がある。とはいえ、病状の進行度合などの時間的な順序や、前後などの空間的な順序の差が重視される場合もある。順序尺度か名義尺度かは、明確に二分できるものではなく、両方の性質を備えるものもあると捉えるのが妥当であろう。

## 5.4 数詞と尺度の関係について

(1)用語集の凡例には「第」から始まる序数詞に限定したものではないが、熟語的な用語を漢数字に、数を数えるという概念が強調されるものには算用数字を用いるというおおまかな表記の方針が示されていた。この凡例に尺度についての言及はないものの、前者は名義尺度、後者は順序尺度を表しているのとらえることもできる。そこで数詞と尺度の関係を調べた結果、表5では、算用数字は順序尺度が多く、ローマ数字に名義尺度が多く有意差が認められた。これにより凡例で示されるおおまかな方針が実際にもある程度は反映されていることが確認できたが、漢数字の場合は、表記から尺度の判別は困難である。

## 6. まとめと今後の課題

本発表では、標準化がなされていない医療記録から数量表現を適切に抽出するために、医学分野の用語集、辞典、国試問題文より、「第」を冠した序数詞を抽出し、これらについて「数詞の表記の

ゆれ」と「名義尺度と順序尺度のどちらにあたるか」に注目して実態の把握と分析を行った。

その結果、横書きという書字方向の環境下であっても、全ての言語資源において、算用数字、漢数字、ローマ数字の3種の使用が見られ、用語集と国試問題文では算用数字の割合が半数以上を、辞典では漢数字が半数近くを占める結果となった。また、各言語資源に、同じ見出し語であっても数詞の表記法にゆれがあるものが少数見られ、表記統一が困難であることが示唆された。

尺度の判別では、全体的に順序尺度と判別したものの方が多いが、名義尺度も拮抗しており、順序を示す序数詞であっても名義尺度の数も少なくないことがわかった。また、元々は順序づけにより命名された語でも、順序の情報が希薄となったために、尺度の判別が困難なものもあった。

数詞の表記と尺度の関係においては、算用数字は順序尺度が、ローマ数字は名義尺度が有意に多かったものの、漢数字は、表記から尺度の判別は困難であった。

以上のことから、医療記録データの自然言語処理において、(1)序数詞「第」を全て順序尺度として扱うことができないこと、(2)「第+ローマ数字」を名義尺度に、「第+算用数字」を順序尺度と見做すことは可能であっても、漢数字については尺度の判別が困難であることが示唆された。

大量の医療記録データには誤変換や入力ミスを含め多様な表記のゆれがあり、これらを全て統制することは困難である。また、入力のし易さは一般に、算用数字>漢数字>ローマ数字と考えられる。よって、医療記録データの自然言語処理において、統計処理などの量的データの処理では、算用数字のみを半角算用数字に統制し、意図して入力されたと考えられる漢数字・ローマ数字を含む数量表現はそのまま文字データとして処理の方が精度の低下は少ないと考えられる。

換言すれば、「横書きの日本語文書では算用数字を原則とする」という表記規則は、自然言語処理の精度の低下を招く可能性があると考えられる。また、縦書きの時代から蓄積された医療用語を、慣用に合わせて立項した結果生じる表記のゆれは、語源を調べる上で貴重な情報となり得る。

医療記録には、「第5,6肋骨骨折」「第1~3趾」などの辞典に立項されていない表記や、「第」が省略されている場合もある。算用数字を含む序数詞を質的データと判断するためには、数詞と助数詞、および助数詞と隣接語の関係を明らかにする必要がある。また、全ての数量表現を語分割用の辞書に網羅的に登録することは困難であるため、語彙規則と文法規則による自然言語処理が要望される。

本調査で得られた結果を踏まえ、医療記録データにおける助数詞と隣接語の関係を明らかにし、医療分野での数量表現の語彙規則の提案に繋がりたいと考えている。

## 謝辞

本研究は JSPS 科研費 JP21H03777 の助成を受けたものです。

## 参考文献

- [1] 開原成允：医学用語の現状と課題，日本語学，Vol.29，No.15，pp.14-24(2010).
- [2] 日本語教育学会(編)：新版日本語教育辞典，pp.264-268，大修館書店(2005).
- [3] 田野村忠温：現代日本語の数詞と助数詞一形態の整理と実態調査一，奈良大学紀要，No.18，pp.194-216(1990).
- [4] 福富和夫，永井正規，中村好一，柳川洋：ヘルスサイエンスのための基本統計学第3版，pp.9-14，南山堂(2006).
- [5] 三省堂編集所(編)：新しい国語表記ハンドブック第8版，p.253，三省堂(2018).
- [6] 朝日新聞社用語幹事(編)：朝日新聞の用語の手引最新版，pp.122-131，朝日新聞社(2002).
- [7] 毎日新聞社(編)：最新版毎日新聞用語集，p.p.424-435，毎日新聞社(2002).
- [8] 日本医学会医学用語辞典 WEB 版(オンライン)．入手先 <<https://jams.med.or.jp/dic/mdic.html>> (参照 2021-10-22).
- [9] 文部科学省，日本医学会：学術用語集 医学編，独立行政法人日本学術振興会(2003).
- [10] 伊藤正男，井村裕夫，高久史磨(編)：医学大辞典第2版，pp.1713-1759，医学書院(2009).
- [11] 南山堂医学大辞典第20版，電子辞典，南山堂(2015).
- [12] 厚生労働省(オンライン)．95回(2001)入手先 <<https://www.mhlw.go.jp/topics/2005/10/tp1028-3/index.html>> 100回(2006)入手先 <<https://www.mhlw.go.jp/topics/2006/04/tp0419-1.html>> 105回(2011)入手先 <<https://www.mhlw.go.jp/topics/2011/04/tp0414-2.html>> 110回(2016)入手先 <[https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou\\_iryuu/iryuu/topics/tp170425-01.html](https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryuu/iryuu/topics/tp170425-01.html)> 115回(2021)入手先 <[https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou\\_iryuu/iryuu/topics/tp210416-01.html](https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryuu/iryuu/topics/tp210416-01.html)> (参照 2021-10-22).
- [13] Giangrande, P. L. F.: Six Characters in Search of An Author: The History of the Nomenclature of Coagulation Factors, *BJH*, Vol.121, No.5, pp.703-712 (2003).
- [14] 奥津敬一郎：数量的表現の文法，日本語教育，No.14，pp.42-60(1969).