調　査　報　告

テキストマイニングを用いた薬剤師国家試験問題の解析
Analysis of the national examination for pharmacists using text mining

白谷　智宣*,[a]，松延　千春[a]，清水　典史[a]，井上　寛[b]
Tomonori Shiratani*,[a], Chiharu Matsunobu[a], Norifumi Shimizu[a], Hiroshi Inoue[b]

[a]第一薬科大学　薬学部　薬学教育推進センター
[b]九州産業大学　経済学部
[a]Center for advancing Pharmaceutical Education, Faculty of Pharmaceutical Sciences, Daiichi University of Pharmacy
[b]Faculty of Economics, Kyushu Sangyo University

学生が薬剤師国家試験に対しての学習を計画するにあたって、それぞれの科目において既出問題を確認・精査することで傾向を把握し、出題頻度の高い項目や関連のある項目から優先的に取り組むことで、各科目の学習における効率の良い導入になることが期待できる。本研究では、テキストマイニングとネットワーク解析から国家試験問題の傾向を可視化し、テキストマイニングからは、頻出単語について 97～103 回と 104～106 回を比較した結果をもとに、国家試験に対する学習の導入時期において科目ごとに頻出単語を示すことで、それぞれの科目でまず取り組むべき指標を与え、さらにネットワーク解析からその単語が含まれるつながりを見ることで、どの項目でその単語が使われていたかが判別でき、重要項目や対応 SBO の抽出に適応できることを示すことができた。国家試験の既出問題を問題ごとに SBO に分類しただけでは、SBO としての出題頻度は見ることはできるが、その中のどの部分が頻出であったかまで見ることは難しい。ダイレクトにテキスト情報を見ることで、細かい出題部分について抽出することができ、その分野での修得するために必要な、具体的な指針を学生にまず提示することが可能になる。具体的な目標を与えることは、まだ国家試験に対する勉強法が確立されていない学生であったとしても、より主体的に学習に取り組めるようになることが期待できることから、薬学教育において十分活用できるものであることが示唆された。

はじめに

平成 18 年度から薬学教育課程として 6 年制課程が導入され、平成 24 年 3 月以降、6 年制薬剤師国家試験のあり方に関する基本方針である「新薬剤師国家試験について」を基準に薬剤師国家試験が実施されてきた[1)]。さらに、平成 28 年に医道審議会薬剤師分科会薬剤師国家試験制度改善検討部会により「薬剤師国家試験のあり方に関する基本方針」が策定され、基本方針の見直しの検討がなされ現在に至っている[2)]。この中で、薬剤師国家試験に関しては、「薬剤師国家試験を通じて、薬剤師資格を有する者として必要とされる倫理観・使命感や基本的な知識等のほか、薬学の全領域に及ぶ一般的な理論や、医療を中心とした実践の場において必要とされる知識・技能・態度等を確認する必要がある。また、薬学に関する基本的な知識等と実践に関する総合的能力が体系的に修得されているか否かを確認することも重要である。」と示されている[2)]。したがって、現在の薬学教育では、薬剤師国家試験に合格することが、大学での 6 年間の学習における知識・技能・態度領域に対する総括的な評価の 1 つとみなすことができる。また、本学では独自に薬学総合演習を開講し、この知識・技能・態度領域に対する総括的な評価を行っている[3)]。この平成 28 年に改訂された薬剤師国家試験のあり方に関する基本方針では、薬剤師国家試験の問題出題形式や合格基準が示されており、さらに、過去に出題された試験問題（既出問題）の取扱いについても言及されている。ここでは、既出問題に関して、「薬剤師に必要な資質を的確に確認することが可能な良質な問題として一定の評価が与えられた問題を活用することとし、その割合は、20%程度とする。」と明記されている[2)]。このことから、学生が薬剤師国家試験に対しての学習を計画するにあたって、それぞれの科目において既出問題を確認・精査することで傾向を把握し、出題頻度の高い項目や関連のある項目から優先的に取り組むことで、各科目の学習における効率の良い導入になることが期待できる。しかし、6 年生制度の国家試験になってからの問題だけでも 10 年分あり、全てを把握することは非常に困難である。そこで、膨大なテキストの中からテキストマイニングを行い有益な情報のみを抽出し、さらに、ネットワーク解析を行うことで出題頻度の高い語句や関連のある項目を効率良く抽出し可視化することで、国家試験の学習に有効な情報を提供できるかを検討した。

方法

テキストマイニングとは、膨大な量の文章(テキスト)を対象として単語の出現頻度を数え、頻度データとして置き換えることで、テキストの特徴を簡単に操作、集計、可視化できるようになり、新しい解釈や有益な情報を引き出す技術である。さらに、本研究ではネットワーク解析を用いて頻出語句・関連語句を比較的大きなネットワークを形成して可視化して過去の国家試験問題のテキスト解析を行った。国家試験の各回の既出問題に関しては、厚生労働省ホームページ内、「薬剤師国家試験のページ」[4)]で公開されており、そのデータを用いた。国家試験での出題科目は、「物理・化学・生物」、「衛生」、「薬理」、「薬剤」、「病態・薬物治療」、「法規・制度・倫理」、「実務」で

あるが、より詳細なデータを得るため、「物理・化学・生物」は、「物理」、「化学」、「生物」に分けて、9 科目としてそれぞれ解析した。それぞれの科目内の複合問題における実務分野の問題は「実務」としてまとめて解析をした。また、それぞれの科目の出題~~の~~傾向の特性を見るために、科目ごとに 97～103 回をまとめたものと直近の 104～106 回をまとめたものの 2 つの成分に分けて解析した。テキストデータの解析は、プログラミング言語「R」を用い、テキストマイニングによる頻度解析は、形態素解析に基づくテキストのデータ化を行うツールとして「Mecab」を用い[5,6]、各科目の 1 単語のみの頻度解析を、さらに、より視覚的にテキストデータを捉えることができるワードクラウドを作成した。ワードクラウドは、頻度解析結果に対して、単語の頻度の高さに応じて、プロットされる単語のフォントサイズが大きくなり、どの単語の頻度が高かったかを視覚的にわかりやすく表現できる特徴がある。さらに、テキスト同士のつながりやネットワークを見るために、ネットワーク解析手法の一つである「Link Community」を用い[6,7]、単語のつながりを分析した。この分析手法の特徴として、1 単語が分類されるクラスターは単一ではなく、1 単語に対して複数のクラスターに分類される点である。これにより、テキストデータの中で、重要なハブとなる単語を把握することができ、ハブとなる単語に関連するネットワークを見ることができるようになる。

結果および考察

今回は、データの解析手法の違いにより、頻度解析のグラフ、ワードクラウド、Link Community によるネットワーク解析図をそれぞれ 9 科目すべて作成し、検討した。

1. 頻度解析およびワードクラウド

「物理」での上位 20 単語の頻度解析のグラフ（Fig. 1）とワードクラウドに基づいた解析結果（Fig. 2）を示す。それぞれ、上は 97～103 回、下は 104～106 回のデータである。

頻度解析（Fig. 1）における縦軸が実際に使用されていた単語の出現回数を示している。このように、この手法を用いることで、国家試験問題のテキストから、実際に用いられた単語を集計し、1 単語としての頻度を直接見ることが可能となる。ここでは、上位 20 単語のテキストが図中に示されているが、ここでの単語についてはワードクラウドで示すことでより視覚的にとらえることが可能になることから、次に、得られた頻度解析の結果からワードクラウドの図を作成した（Fig. 2）。ワードクラウドは、単語に対する文字の大きさが、実際の頻度と比例しているので、大きいほどより多く用いられているものになる。すなわち、97～103 回は頻出順に、分子、イオン、温度、軌道、エネルギーであることが見てとれる。また、104～106 回は、標準、ナトリウム、k、滴、酵素であることがわかる。上位単語では単語の重複は少ないように見受けられるが、クロマトグラフィーやスペクトルはどちらにも見られ、さらにそれに関連する単語が、頻度の大小はあるが、どちらにも見られる。物理は上位 20 単語では 7 単語

重複しており、重複率は 35%であった。ワードクラウドでは上位 50 単語に関して図示しているが、97～103 回と 104～106 回で頻度に違いはあるが、重複している単語は 46%占めている。重複の度合いを見ることで、出題頻度の高い SBO が 10 年間において変化しているかの指標になると考える。

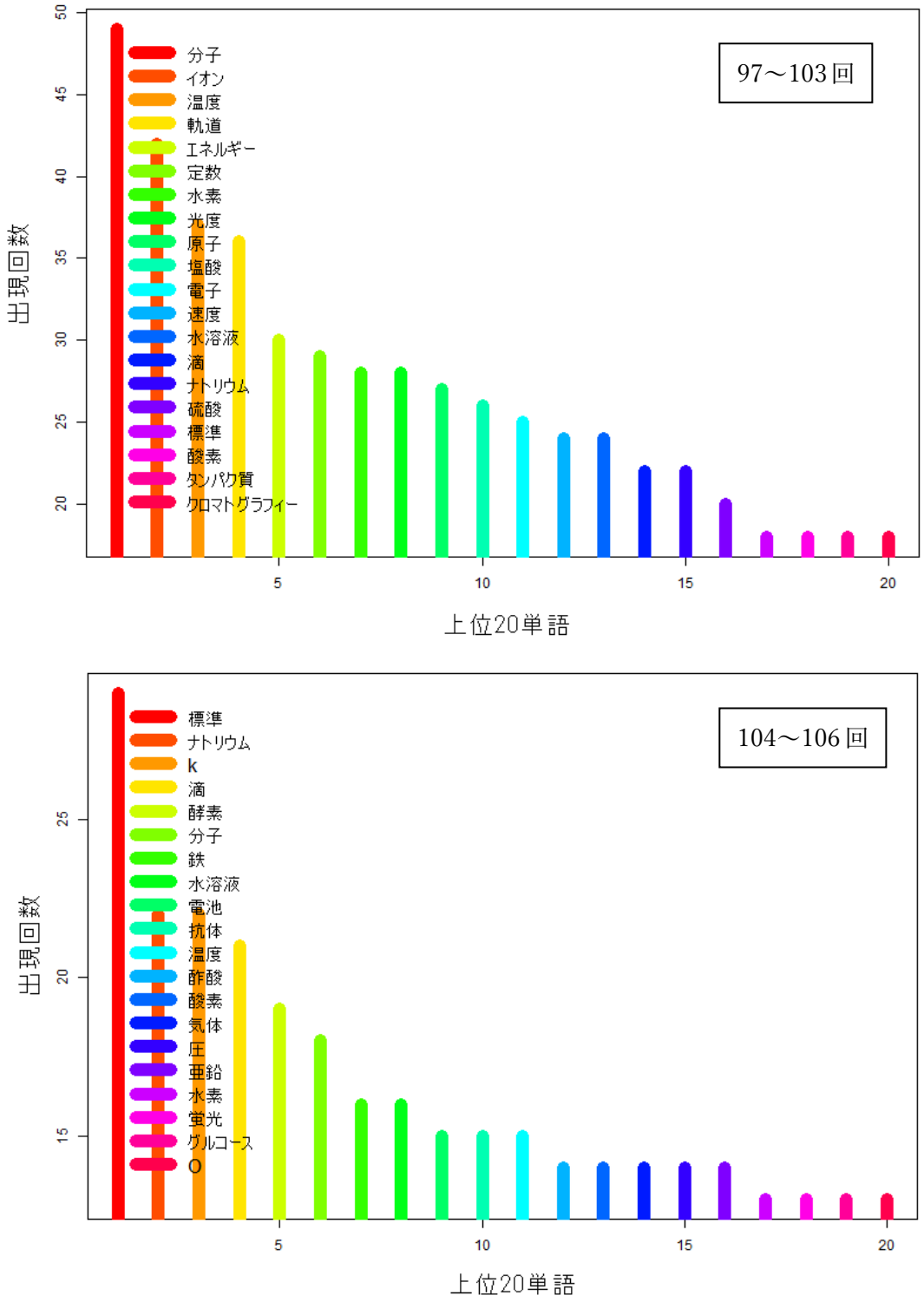

Fig. 1 「物理」の頻度解析結果（上：97～103 回、下：104～106 回、上位 20 単語）

したがって、物理分野は、直近の過去 3 年間の過去問題を網羅しただけでは、全体の出題傾向をつかむうえでは不十分であることがわかり、さらにさかのぼって幅広く学習し理解する必要があることがわかる。しかし、重複率がそこまで高くないので、重要部分の指標は示せるが、これだけでは不十分で全体的に網羅していく必要があることがわかる。

97～103 回

104～106 回

Fig. 2「物理」のワードクラウド（上：97～103 回、下：104～106 回）

そこで、他の科目についても傾向の特徴を知るために、同様に頻度解析とワードクラウドによる解析を行い、97～103 回と 104～106 回での上位 20 単語および 50 単語における重複率を求めた（Table 1）。

	物理	化学	生物	衛生	薬理	薬剤	病態*	法規**	実務
上位 20	35	55	55	80	80	50	60	55	55
上位 50	46	40	40	80	60	48	56	40	42

*病態：病態・薬物治療、**法規：法規・制度・倫理

Table 1　9 科目の上位 20 および 50 単語での重複率（%）

ここから、衛生および薬理は上位 20 単語の重複率が非常に高いことがわかる。さらに衛生は上位 50 単語においても重複率が非常に高い。物理と比較するために、重複率の高かった衛生のワードクラウドを示す（Fig. 3）。

Fig. 3「衛生」のワードクラウド（上：97～103 回、下：104～106 回）

このことから衛生および薬理は直近 3 年分をまず精査し学習することで重要部分

の傾向をつかむことが可能になり、効率よく国家試験に対する学習の導入が可能になる。病態・薬物治療もこの 2 科目ほどではないが比較的重複率が高いので、同様な学習計画が適用できると考える。しかし、前述の物理と同様、化学、生物、法規・制度・倫理、実務は上位 50 単語の重複率が 40%程度とそこまで高くない。これらの科目は、重要部分の提示はできるので、まず最初に取り組むべき指標として、これらの単語を有する SBO を中心に学習するが、過去問題に固執することなく幅広い学習に取り組む必要がある。

2. ネットワーク解析

物理、衛生以外はすべて上位 20 単語での重複率のほうが高く、50 単語になると低くなる。これらは、国家試験における出題のコアとなる部分は過去 10 年においてあまり変化していないことを示している。しかし、1 単語での頻度解析の欠点は、頻度の多かった単語の前後にどのような単語が書かれていたか、単語同士のつながりを見ることができない。当然、上位 20 単語に連動して文章の前後に位置する単語が頻度として上位に来ると思われるが、上位 20 単語でカウントされた単語が同じであったとしても同じ SBO、内容から出題されているとは限らないため、50 単語まで広げると重複率は下がることになる。そこで、単語同士のつながり、すなわちネットワークを見ることで、どの部分の内容からの出題であったかを精査する必要がある。今回は、Link Community によるネットワーク解析を行い、それぞれ共起ネットワーク図を作成した。「物理」でのネットワーク解析図を示す（Fig. 4）。上は 97～103 回、下は 104～106 回のデータである。

どちらにも共通している単語にクロマトグラフィーがある。それぞれネットワークを見ると、97～103 回は分配や液体などと組んでいて、104～106 回では薄層や板などと組んでいる（Fig. 4）。ネットワーク解析をすることで、同じ単語であったとしても、文章としてのつながりを精査することで、細かく出題内容の違いを見ることができることがわかった。上位 50 単語での重複率が 40%である科目は、共通のネットワークが組まれている可能性が低く、同一単語であったとしても、それぞれのネットワークを比較して、内容について精査する必要がある。また、104～106 回の頻度解析で k が上位 3 位の出現回数であったが、ネットワーク解析図では見ることができなかった。これは、文章でなく数式内などで出現していたため、ネットワーク解析にかからなかった可能性がある。この部分は今後検討していく必要がある。

今回、物理では全く同じネットワークを見ることはできなかったが、単語の重複率の高かった衛生では、ほぼ同じネットワークをもつものがあった（Fig. 5）。

物理と同様、上は 97～103 回、下は 104～106 回のデータである。衛生のネットワーク解析図（Fig. 5）から、例えば、どちらにも、中という単語を中心に、尿、大気、食品、成分といった単語とネットワークが組まれていることが見られる。衛生は単語の重複率が 80%と高かったため、そのネットワークも同じものが多いことが示された。

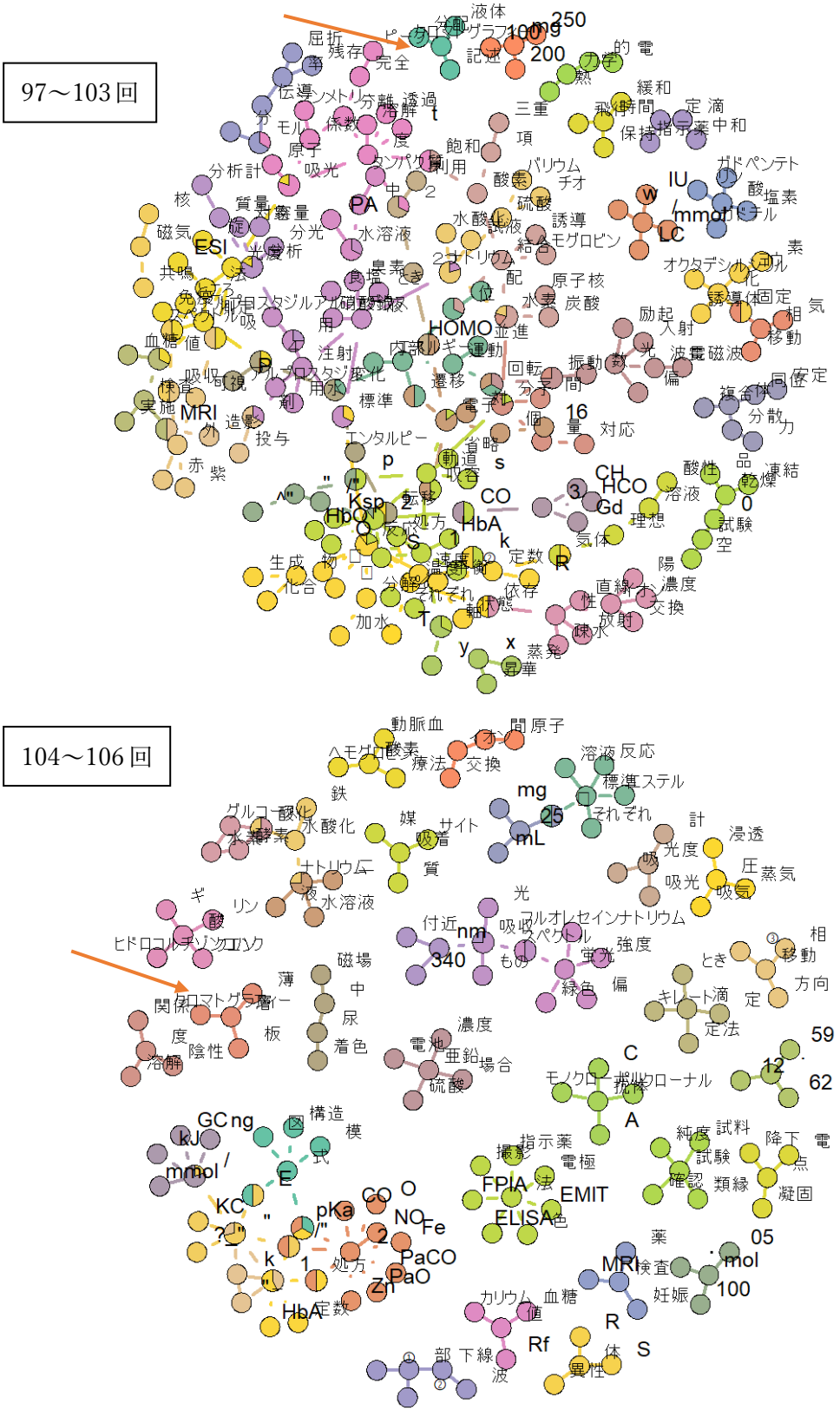

Fig. 4「物理」のネットワーク解析図（上：97～103 回、下：104～106 回）
どちらにも共通している単語にクロマトグラフィーがある。それぞれネットワークを見ると、97～103 回は分配や液体などと組んでいて、104～106 回では薄層や板などと組んでいる（矢印）。

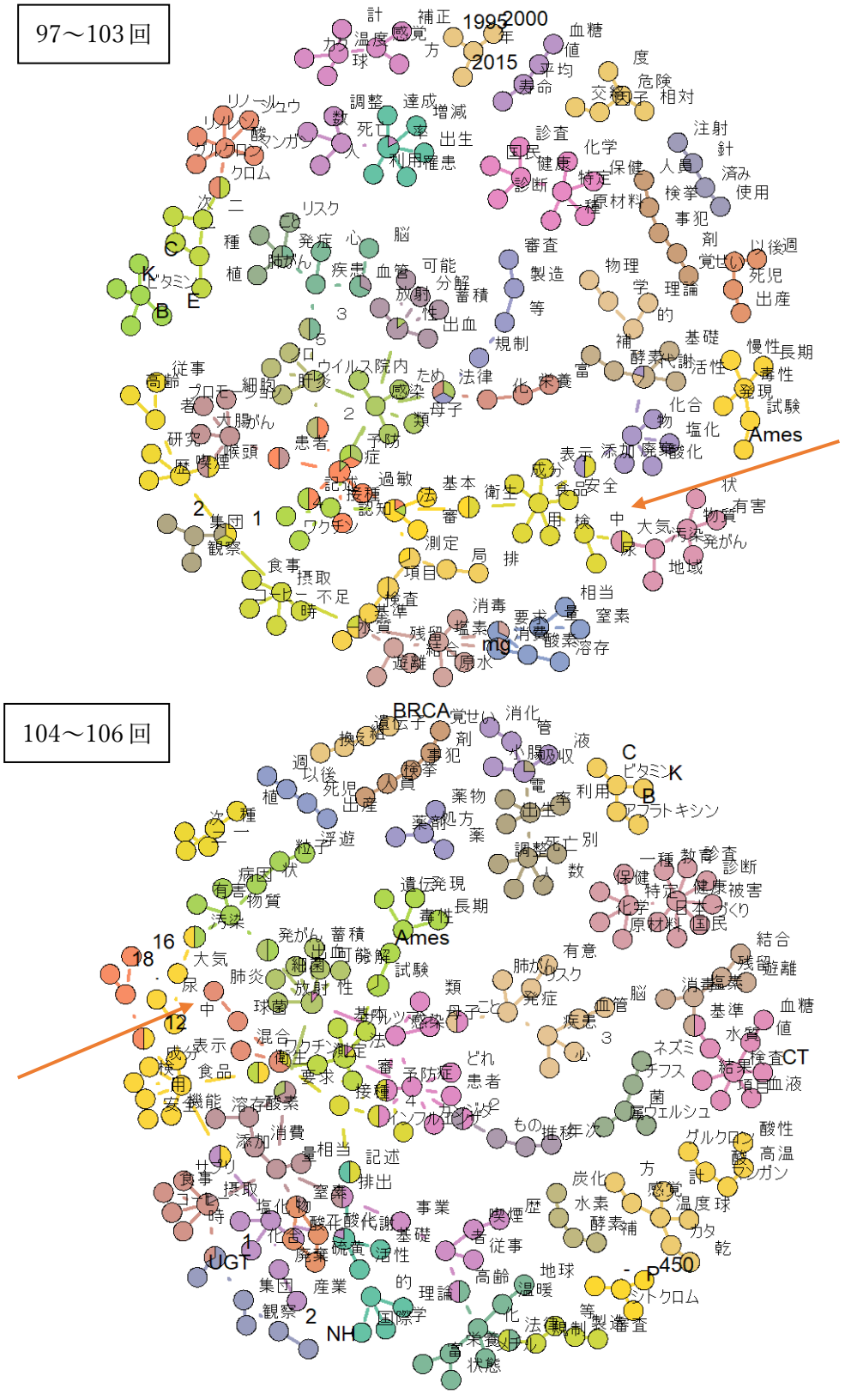

Fig. 5「衛生」のネットワーク解析図（上：97～103 回、下：104～106 回）
どちらにも、中という単語を中心に、尿、大気、食品、成分といった単語とネットワークが組まれていることが見られる（矢印）。

したがって、ネットワーク解析からも衛生は、国家試験の過去の既出問題をまずしっかりと学習することで、出題傾向を把握し、重要部分の抽出が可能になることが示唆された。薬理に関しても重複率は高かったが、上位 50 単語になると重複率は下がったことから、衛生ほど同じネットワークが出現していないことがわかる。

まとめ

このように、学生が国家試験に対する学習に取り掛かるにあたって、国家試験の既出問題についての頻度解析やワードクラウドを提示することで、ここで示された単語は、科目ごとに重要なキーワードとして認識させ、国家試験に対する学習において、常に意識してこの単語が内容に含まれる SBO について学習することで、科目ごとの知識の最初の骨格形成に大きく寄与できるようになる。さらに、ネットワーク解析から得られた情報から、細かく出題された内容や、具体的な SBO との紐づけが可能になり、重要な部分を明確にすることが可能になる。国家試験の既出問題を問題ごとに SBO に分類したものはよく見る。しかし、SBO を示すだけでは、SBO としての出題頻度は見ることはできるが、さらに細かくその中のどの部分が頻出であったかまで見ることはできない。このようにダイレクトにテキスト情報を見ることで、細かい出題部分について抽出することができ、その分野での修得するために必要な、具体的な学習目標を学生に提示することが可能になる。抽象的な SBO だけの情報よりも、このように、まず必要となる具体的な細かい指針を提示することで、まだ国家試験に対する勉強法が確立されていない学生であったとしても、より主体的に学習に取り組めるようになることが期待できる。

今回、テキスト情報でしかない国家試験の既出問題について、テキストマイニングおよびネットワーク解析を行うことで国家試験問題の情報を可視化することができた。本手法は、薬学教育の中で活用していくうえで、十分価値のあるものと考える。今後、改良を加えながら、継続してこの手法を活用し、国家試験に対してより正確で詳細な傾向を見出したい。

謝辞

本研究は、令和 3 年度 第一薬科大学研究奨励金の助成により研究が遂行されたものです。この場を借りて深く御礼申し上げます。

引用文献

1) [internet]厚生労働省、「新薬剤師国家試験について」,
https://www.mhlw.go.jp/shingi/2010/01/dl/s0120-3a.pdf
2) [internet]厚生労働省、医道審議会薬剤師分科会薬剤師国家試験制度改善検討部会、「薬剤師国家試験のあり方に関する基本方針」,
https://www.mhlw.go.jp/file/06-Seisakujouhou-11120000-Iyakushokuhinkyoku/0000112014.pdf.

3) 2021 年 第一薬科大学薬学部シラバス.
4) [internet]厚生労働省、「薬剤師国家試験のページ」, https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iyakuhin/yakuzaishi-kokkashiken/index.html.
5) 石田 基広, R によるテキストマイニング入門 (第 2 版) 森北出版株式会社, (2017).
6) 井上 寛, テキストマイニングとネットワーク分析を用いたオープンキャンパスアンケート自由記述の分析, 第一薬科大学研究年報, 35, 45-53 (2018). (平成 30 年度学内学術奨励金 成果報告).
7) Ahn Y-Y, et al. Link communities reveal multiscale complexity in networks, Nature, 466, 761-764 (2010).