医師国家試験コーパスの構築と語彙分析—対数尤度比に基づく特徴語の算出—

山元一晃（やまもとかずあき）・品川なぎさ（しながわ）・稲田朋晃（いなだともあき）（国際医療福祉大学）

留学生が日本語で医学を学習するにあたって、もっとも障害になると予想されるのが、膨大な専門語彙の習得である。語彙習得のためのよりよいカリキュラム作りを目指すための第一歩として、本研究では、医師国家試験のコーパスを作成し、さらにそのコーパスから医師国家試験に特徴的な名詞語彙を抽出し、一般的な語彙との差を明らかにした。

１．先行研究

医学部医学科の留学生への日本語教育に関する報告はいくつかあるが、その中に医師国家試験を扱ったものはない。同じ医療系である看護師国家試験、介護福祉士国家試験については、試験問題の分析が報告されている（岩田 2014、大場 2017)。これらの研究は、いずれの国家試験でも１級・級外の専門語彙が多数を占めるという結果を得ており、専門語彙をいかにして学習・習得していくかという学習方法について今後検討していく必要があることを示している。

医師国家試験についても同様の結果が予想される。本発表では、まず、医師国家試験コーパスの構築について述べ、次に、そのコーパスを使って語彙面、特に名詞に絞って行った分析について述べる。

２．コーパスの概要

医師国家試験のうち 106 回(2012 年実施)～111 回(2017 年実施)の計 6 回分をデジタル化し、形態素解析（MeCab 0.996 および UniDic 2.1.2 を使用)を行った。医師国家試験は、各年問題 A～問題Ｉまであり、それぞれが問題冊子と別冊に分かれている。そのうち問題冊子のみを対象とした。形態素解析を行ったのち、空白・空行を機械的に除去し、名詞として分類されていた記号（「」など）を目視により確認、修正を行った。それ以外は修正していない。

３．分析の観点と方法

本発表では、名詞のみの分析を示す。全異なり語 8,948 語中、名詞が 7,629 語(85.6%)と最も多く、かつ、全体の８割以上を占めていたためである。そのうち、漢字・ひらがな・カタカナのいずれかを含む名詞 6,571 語(73.4%)を分析対象とし､以下の観点から分析した。

（１)「現代日本語書き言葉均衡コーパス(BCCWJ)」と対照した対数尤度比（近藤 2011 を参照）に基づく指標（「特徴度」とする）により医師国家試験に特徴的な語を客観的に抽出する。対数尤度比は、当該テキストに比して参照テキストに特徴的である語であっても高い値となるため、そのような語については負の値が出るよう-1 を乗じる。

（２）上記（１）で抽出した特徴的な語を、日本語学習者が学ぶ機会があるのか、また、どのレベルの学習者であれば既知だといえるのかを「日本語教育語彙表」(Sunakawa, et. al 2012)と照らし合わせながら明らかにする。

４．分析結果

臨界値 3.84 (p<.05) を上回る特徴度を示す異なり語数は、2,894 であった。その例として特徴度が高い順に 20 語を示す。

所見　来院　別冊　球　主訴　呼吸　血圧　整*　異常　血液　血球**　脈拍
胸部　腹部　検査　分　基準　投与　体温　小板***

*「脈拍 76/分、整。」(第 111 回Ａ問題)などの例がある。

**「赤血球」または「白血球」などの語の構成要素である。

***　「血小板」の構成要素である。

そのうち「日本語教育語彙表」の全６レベルに該当する語数は以下の表１の通りである。国家試験に特徴的な名詞の大半が初級前半～上級後半までに含まれず、一方で、特徴度が低い語は初級前半～上級後半までの間に 95%以上の語が含まれることが分かる。

【表１　日本語教育語彙表の各レベルに該当する異なり語数】

	特徴度が有意に高い(p<.05)名詞		特徴度が有意に低い(p<.05)名詞		全名詞（異なり）	
	語数	割合	語数	割合	語数	割合
初級前半／初級後半	59	2.0%	254	20.3%	408	6.2%
中級前半／中級後半	401	13.9%	796	63.8%	1971	30.0%
上級前半／上級後半	327	11.3%	155	12.4%	945	14.4%
該当なし	2107	72.8%	43	3.4%	3247	49.4%
合計	2894	100.0%	1248	100.0%	6571	100.0%

BCCWJ には含まれていない 201 語は表１に含まれていない。そのうち、頻度が 2 以上のものは 86 語、そのうち頻度が 5 以上のものに限っても「開大」「カニューラ」など 20 語ある。これらの語についても、医師国家試験に特徴的な語であるととらえられる。

５．まとめと今後の課題

本発表では、医師国家試験のコーパスから特徴的な語を抽出し、日本語教育語彙表との対照をした。その結果、医師国家試験に有意に特徴的な語の大半が、一般的な上級レベルの日本語教育ではカバーされない可能性が示唆された。今後は、ほかの品詞についても分析を進め、テキストカバー率も考慮したリストを作成し、教材化を図る。

参考文献：

岩田一成(2014)「看護師国家試験対策と「やさしい日本語」」『日本語教育』第 158 号 36-48.

大場美和子(2017)「介護福祉士国家試験の筆記試験における文法・語彙項目の分析―日本語能力試験の観点から―」『小出記念日本語教育研究会論文集』25 号 5-20.

近藤明日子(2011)「中学校・高校教科書の教科特徴語リストの作成」『特定領域研究「日本語コーパス」言語政策班報告書 (JC-P-10-01)』145-152.

Sunakawa,Y.,Lee,J.&Takahara,M.(2012)The construction of a database to support the compilation of Japanese learners' dictionaries. Acta Linguistica Asiatica, 2(2).