留学生向け看護語彙リスト作成のためのコーパス構築における課題 -『系統看護学講座』シリーズ３巻のパイロットスタディから-

山元一晃[1] 浅川翔子[2] 稲田朋晃[3] 岩間裕司[4] 土屋ともえ[5]
[1]金城学院大学文学部 [2]東京慈恵会医科大学医学部看護学科 [3]十文字学園女子大学教育人文学部
[4]防衛医科大学校医学教育部 [5]国際医療福祉大学成田看護学部
yamagen.98@gmail.com s-asakawa@jikei.ac.jp t-inada@jumonji-u.ac.jp
iwama@ndmc.ac.jp t-tsuchiya@iuhw.ac.jp

概要

看護留学生への支援に活用できる語彙リストを作成するためのコーパス構築のパイロットスタディについて発表する.まず,全 70 巻の『系統看護学講座』シリーズのうち３巻を文字化し，語彙の多様性や，品詞・語種の観点から留学生が直面しうる困難について検討した.その結果,名詞が多いこと，漢語や外来語が多いことなどから，母語や背景によっては難しく感じることが示唆された.その後,本パイロットスタディを踏まえ，コーパス構築を前提とした整形作業において考慮すべきことについて検討した.教科書という特性上，コラムや練習問題などが入り組んでおり，学習の必要性や語彙の特徴の過大評価の可能性などを踏まえ，今後の整形の方針を述べた.

1 はじめに

日本の看護系大学には，2022 年度現在，少なくとも 174 名の留学生が在籍している[1]. この数は短期大学や専修学校における留学生の数は含んでおらず,多くの専修学校が留学生向けの選抜を行っていることを鑑みれば，相当数の留学生が看護を学んでいることが予想される.また,日本政府が外国人労働者の受け入れに舵を切っていること，看護師の有効求人倍率は依然として高く，看護師不足の状況が続いていることから，看護系留学生の受け入れは，今後，加速していく可能性が高い.一方で,看護師を目指す留学生の支援に活用することを目指した研究は少ない.

特に語彙については，5 名の留学生を対象としたインタビュー調査で，全員が専門用語に特化した教材が欲しいと述べていた [2]. また,読解の理解には,語彙のテキストカバー率の高さが重要であるという研究もあり[3]，看護教育において用いられる教材の理解には，一定量の語彙を習得していることが重要であると考えられる.

そこで，発表者らは，効率的な語彙学習を支援するための語彙リストを作成することを目指し，その客観的指標を提供する看護教科書コーパスを構築中である.本発表では，パイロットスタディとして，構築するコーパスの一部の分析を行い，コーパス構築の整形作業を進めるにあたって考慮すべき点を明らかにする.

2 先行研究

看護留学生の支援のための語彙研究としては，実習記録を対象としたものがあり，記録の種類や記録する項目により必要とされる語彙の特徴が異なっていることが明らかになっている[4].この分析は，留学生向けの教材へと活かされている[5].

また，看護師国家試験の語彙を分析した研究では，名詞語彙の多くが日本語能力試験の級外語彙であり，日本語母語話者にとって難しいことを明らかにしている[6].

これらの研究は，看護日本語教育に多くの知見を与えるが，看護学部での学びを通して必要とされる語彙については，まだ明らかとなっていない.

3 方法

本研究においては，『系統看護学講座』シリーズ（医学書院）を電子化しテキスト化を行う.『系統看護学講座』は体系化された教科書であり，全国の多くの看護学校・看護系大学で採用されているという[7]. このシリーズは，専門分野全 32 巻，専門基礎分野全 11 巻，基礎分野全 9 巻，別巻全 18 巻の計 70 巻に分かれている[8].本研究においては 2024 年 8 月 2 日時点のいずれも最新巻を購入し，スキャンを行い，JPEG 形式で保存した.その後，Anthropic 社

の Claude API [9]を用いてテキスト化を行うための Python プログラムを作成し，テキスト形式で保存した.

現在は，整形作業を行っている.具体的には，文字起こしと原本とに齟齬がないかを確認する作業や，原本にはない文章が生成されていることがあるため，それを削除する作業を行っている.また，並行してタグ付けの作業を行っている.

本発表においては，最低限の整形作業を行ったファイルを用いて，その一部の形態素解析を行った結果を示し，今後の整形作業において考慮すべきことを検討する.

パイロットスタディとしての分析であるため，看護を学ぶ前提となると考えられる基礎分野 9 巻のうち『物理学』『化学』『生物学』の 3 巻を対象として分析を行った.

最低限の整形作業として，作成したプログラムにより自動的に付与した箇所の削除，ページが分かれていることにより一文が途中で切れてしまっている箇所の修正，生成 AI により挿入された図表の説明やページが白紙であることの説明の削除を行った.

その後，MeCab 0.996[10], 形態素解析用辞書として現代書き言葉 UniDic ver.2023.02 [11]を用い，短単位に形態素解析を行った.

形態素解析済みのテキストから品詞大分類が「記号」「補助記号」「空白」に分類されたものを除外し，全体の延べ語数，異なり語数を集計した.その上で，語彙多様性および品詞ごとの PMW (100 万語あたり頻度)を集計した.

語彙多様性の指標としては，単純な Type/Token Ratio よりも頻度の影響を受けにくい修正 Type/Token Ratio (CTTR) を用いた[12].なお CTTR は，異なり語数を延べ語数の 2 倍の平方根で除した値である.

4 集計結果と考察

３冊の延べ語数，異なり語数，語彙多様性の指標である CTTR は表 1 のとおりとなった.

表 1　３巻の集計結果

タイトル	異なり語数	延べ語数	CTTR
物理学	3006	64397	8.38
化学	3987	103869	8.75
生物学	6487	158941	11.5

『物理学』『化学』に比べ，『生物学』の CTTR が高く，より多様な語彙が使われていることが分かる.このことは，生物学に関する語彙の学習の負担が大きいことを示唆している. 『生物学』には『物理学』の２倍近くの異なり語が含まれていることからも，そのことが分かる.

次に，品詞ごとに，語の頻度を延べで集計した結果を表 2〜表 4 に示す.参考として『現代日本語書き言葉均衡コーパス』(BCCWJ) [13]における品詞ごとの頻度[14]を表 5 に示す.

表 2　品詞ごとの集計（物理学）

品詞	延べ語数	PMW
名詞	28519	442862.3
助詞	17182	266813.7
動詞	7854	121962.2
助動詞	3132	48635.81
記号	2161	33557.46
接尾辞	1924	29877.17
形容詞	1033	16041.12
連体詞	650	10093.64
形状詞	579	8991.1
代名詞	468	7267.42
副詞	366	5683.49
接頭辞	317	4922.59
接続詞	196	3043.62
感動詞	16	248.46

表 3　品詞ごとの集計（化学）

品詞	延べ語数	PMW
名詞	57669	555209
助詞	21120	203333
動詞	8597	82767.72
記号	5268	50717.73
接尾辞	3774	36334.23
助動詞	3430	33022.36
接頭辞	962	9261.67
形容詞	756	7278.4
形状詞	628	6046.08
連体詞	608	5853.53
副詞	429	4130.2
接続詞	316	3042.29
代名詞	284	2734.21
感動詞	28	269.57

表 4　品詞ごとの集計（生物学）

品詞	延べ語数	PMW
名詞	80574	506942.8
助詞	37391	235250.8
動詞	16009	100722.9
助動詞	7170	45111.08
接尾辞	6977	43896.79
記号	3515	22115.12
接頭辞	1557	9796.09
形状詞	1292	8128.8
連体詞	1251	7870.85
形容詞	921	5794.6
接続詞	778	4894.9
副詞	755	4750.19
代名詞	713	4485.94
感動詞	38	239.08

表 5　品詞ごとの集計（BCCWJ）

品詞	延べ語数	PMW
名詞	36651583	350355.9
助詞	31428580	300428.8
動詞	14148216	135244.1
助動詞	10279970	98267.21
接尾辞	3346976	31994.06
副詞	1830329	17496.29
形容詞	1588226	15182
代名詞	1516372	14495.14
形状詞	1314004	12560.69
連体詞	997276	9533.056
接頭辞	868076	8298.021
接続詞	481094	4598.823
感動詞	161716	1545.859

『物理学』『化学』『生物学』のいずれにおいても，BCCWJ に比して，名詞が多いことが分かる.『化学』『生物学』では，それぞれ，PMW で 1.58 倍，1.45 倍であり，名詞の多さが際立っている.一方で『物理学』は 1.26 倍であり，索引や目次などの名詞が中心の部分も含まれていることを考えると，顕著に多いとまではいえない.

また，語種ごとに，語の頻度を延べで集計した結果を表 6〜表 8 に示す.なお，3 節で述べたように品詞が「記号」「補助記号」のものは除外しているが，品詞が「名詞」等で，語種が「記号」のものは除外していない.参考として BCCWJ における語種ごとの頻度を表 5 に示す.

表 6〜表 9 については，いずれも頻度の高いものから順に並べている.

表 6　語種ごとの集計（物理学）

語種	延べ語数	PMW
和語	36803	571501.8
漢語	21597	335372.8
記号	2610	40529.84
外来語	1837	28526.17
[語種なし]	855	13277.02
混種語	533	8276.78
固有名	162	2515.65

表 7　語種ごとの集計（化学）

語種	延べ語数	PMW
漢語	44515	428568.7
和語	41365	398242
記号	7839	75470.06
外来語	5737	55233.03
[語種なし]	3487	33571.13
固有名	511	4919.66
混種語	415	3995.42

表 8　語種ごとの集計（生物学）

語種	延べ語数	PMW
和語	75057	472231.8
漢語	68682	432122.6
外来語	5680	35736.53
記号	5021	31590.34
[語種なし]	2874	18082.18
固有名	1024	6442.64
混種語	603	3793.86

表 9　語種ごとの集計（BCCWJ）

語種	延べ語数	PMW
和語	71518076	683648.05
漢語	26106078	249550.47
外来語	2945152	28152.99
固有名	2661023	25436.97
混種語	1125516	10758.91
記号	256511	2452.01
その他	62	0.60

『物理学』『化学』『生物学』のいずれにおいても，BCCWJ に比して和語が少なく，漢語・外来語が多く用いられていることが分かる.特に『化学』『生物学』において，その頻度が高い.非漢字圏出身の留学生にとっては漢字が，また，外来語を苦手

とする留学生にとっては外来語が障壁となる可能性がある.『化学』については，『外来語』が PMW で約 1.96 倍用いられており，外来語も習得の鍵を握る.

語種が付与されていない[語種なし]は形態素解析用辞書に含まれていない「未知語」であると考えられる.この「未知語」が『物理学』では 1.33%，『化学』では 3.36%，『生物学』では 1.81%含まれており，無視できない割合になっている.

特に化学や物理においては，未知語として解析された外来語が多いことが予想されるため，未知語の扱いについても今後検討する必要がある.

5 整形において考慮すべきこと

本節では，パイロットスタディーとしての最低限の整形作業を行う際に，また，形態素解析を行う際に明らかとなった，整形において考慮すべきことを述べる.

表紙・目次・奥付・索引等の扱い

表紙・目次・奥付・索引についてもスキャンを行っている.実際に留学生が困難を抱えると考えられるのは本文であるため，これらについては，除外してもよいと考えられる.整形作業において，タグ付けを行い，機械的に削除できるようにし，分析時には含めない.

図表の扱い

教科書であるため，読者の便宜のため図表が多く含まれている. この中にも，多くの語彙が含まれている. ただ，図表のタイトルや図表内の語を含めることで，図表による説明が必要な箇所で用いられる語が，過剰に特徴的な語として抽出されてしまう可能性を考慮し，分析には含めない.

枠囲み箇所の扱い

本文には「コラム」のような記事や，「例題」「ゼミナール」のような練習問題が枠囲みされている. このような箇所は，教科書を読み進める上で重要であること，「例題」や「ゼミナール」は，全体にわたってあることから，過大評価はされにくいと考え，分析に含める.

脚注の扱い

本文には脚注が含まれている. これは，読みながら参照するものであると考えられるため，分析に含める.

巻末資料・解答の扱い

たとえば『物理学』には，算数・数学の基礎知識が巻末資料として掲載されている. このような巻末資料は必要に応じて読むものであり，分析には含めないこととする. 解答には，解答の他に簡単な説明が含まれるが，こちらも必要に応じて読むものであるため，分析には含めない.

生成 AI 特有の問題

OCR については，既存の OCR アプリケーションを使うよりも，その精度が極めて高い. 一方で，たとえば，図表の説明など原本にない情報が追加されてしまうという問題がある. これについても本文と照らし合わせ人手で整形作業を行う必要がある.

6 まとめと今後の展望

看護留学生の支援での活用を視野に，看護教科書のコーパスを構築するにあたってのパイロットスタディーについて述べた.

全 70 巻のうちの 3 巻であったが，品詞や語種などに特徴が見られ，看護留学生が困難を抱えると考えられるポイントが分野によって異なることが分かった.

またパイロットスタディにより，整形において考慮すべきことも明らかとなった.

今後は，全 70 巻の整形・タグ付け作業を進めていく.その上で，先行研究を参考に以下の作業を行っていく.

- 全 70 巻の形態素解析.多様な分析を可能にするため，UniDic による短単位での形態素解析のほか，医療施設で使われる用語を集めた辞書である ComeJisyo[15]を使用した形態素解析も行う.
- 特徴語の抽出.対数尤度比を指標として BCCWJ の語彙表との比較を行い，看護教科書それぞれに特徴的な語彙を明らかにする.
- 新たに学ぶべき語の抽出.抽出された語のリストを『日本語教育語彙表』[16]と対照させ，日本語教育で扱われないと考えられる語を抽出する.

上記によって抽出された語のリストから語彙リストを作成し，テスト等によって，語彙の観点から看護学生のレディネスを明らかにし，スムーズに専門教育へと移行できるような支援を考えていきたい.

謝辞

本研究は JSPS 科研費 JP24K03990 の助成を受けたものです.

参考文献

1. **日本看護系大学協議会・日本私立看護系大学協会**.「看護系大学に関する実態調査」報告書（2022 年度状況）. 出版地不明：https://www.jspcun.or.jp/wp/wp-content/uploads/2024/06/47910c7a8a0b11a05ea380f9e9406e85.pdf, 2025-1 閲覧.
2. 看護師を目指す留学生と看護教員が日本語教師と日本語の教材に期待すること―留学生・看護教員へのインタビュー調査から―. **山元一晃, 浅川翔子, 加藤林太郎**. 名古屋：金城学院大学論集 人文科学編, 2023 年, 第 19 巻 2 号. 222-229.
3. **Nation, I. S. P.**. Learning Vocabulary in Another Language (2nd Ed.). New York：Cambridge University Press, 2013.
4. 看護実習記録に用いられる語彙の特徴の分析. **山元一晃・浅川翔子**. 東京：社会言語科学, 2021 年, 第 23 巻 2 号. 67-80. https://doi.org/10.19024/jajls.23.2_67
5. 看護師を目指す留学生のためのライティング教材の開発とその活用. **山元一晃, 浅川翔子, 加藤林太郎**. 名古屋：金城学院大学論集 人文科学編, 2022 年, 第 18 巻 1 号. 129-139.
6. 看護師国家試験対策と「やさしい日本語」. **岩田一成**. 東京：日本語教育, 2014, 第 158 号. 36-48. https://doi.org/10.20721/nihongokyoiku.158.0_36
7. -. 看護教員の方へ | 医学書院. (オンライン) (引用日: 2025 年 1 月 9 日.) https://www.igaku-shoin.co.jp/booklist_educator?series=21.
8. -. 系統看護学講座 2025_パンフレット. (オンライン) (引用日：2025 年 1 月 9 日.) https://www.igaku-shoin.co.jp/prd/catalog/2025/textbook_keikan/#page=1.
9. -. Build with Claude Anthropic. (オンライン) (引用日：2025 年 1 月 9 日.) https://www.anthropic.com/api.
10. -. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. (オンライン) (引用日：2025 年 1 月 9 日.) https://taku910.github.io/mecab/.
11. -. 「UniDic」国語研短単位自動解析用辞書. (オンライン) (引用日：2025 年 1 月 9 日.) https://clrd.ninjal.ac.jp/unidic/.
12. On Sampling from a Lognormal Model of Word Frequency Distribution. **Carrol, J. B.**. Computational Analysis of Present-Day American English. Providence：Brown University Press, 1967. 406-424.
13. **前川喜久雄（監修）, 山崎誠（編）.** 書き言葉コーパス―設計と構築―. 東京：朝倉書店, 2014.
14. -. 「中納言」版公開データ 現代日本語書き言葉均衡コーパス（BCCWJ）. (オンライン) (引用日：2025 年 1 月 9 日.) https://clrd.ninjal.ac.jp/bccwj/bcc-chu.html.
15. -. ComeJisyo Project. (オンライン) (引用日：2025 年 1 月 9 日.) https://comejisyo.com/.
16. -. 日本語教育語彙表. . (オンライン) (引用日：2025 年 1 月 9 日.) https://jhlee.sakura.ne.jp/JEV/.