Random Forests による英語理学療法論文からの特徴語抽出

——Corpus of Contemporary American English Full Text 版を参照コーパスとして——

八野幸子

Abstract

The purpose of this study was to extract key words in English physical therapy articles using Random Forests. For the data analysis, the author compiled a corpus of English physical therapy articles (PT). The Corpus of Contemporary American English (COCA) Full-Text version, especially its sub-corpus "Academic Medicine" (CM), was used as a reference. Random Forests (RF), an ensemble classifier originally developed by Breiman (2001), was used to extract key words. Tabata (2012-a) utilized RF to spotlight lexical items that Charles Dickens consistently used. In the study, Tabata pointed out that measures used for key word analysis in previous studies, such as Log likelihood and Chi square tests, extract words that frequently appear in a single text as the key words in a whole corpus and proposed Random Forests as an alternative measure. The author hypothesized that using RF as a measure would extract the key words more consistently since previous studies on physical therapy English have not use RF for key word analysis and corpuses from other medical fields have not been used as references.

In the results, words such as *rehabilitation, motor,* and *mobility* which are important in the field of physical therapy were extracted and the validity of the key words was demonstrated by an experienced physical therapist. These results confirmed that Random Forests can extract the key words which are consistently used in a corpus.

1. はじめに

1.1 研究背景

近年の日本の大学において、実学系学部は増設傾向にあり、大学は教育研究機関であることから、これらの実学系学部においても文献購読等の観点から、英語教育は重要であると考えられる。理学療法分野でも同様に、大学における学部増設傾向がみられ、このような傾向は理学療法分野の英語研究の必要性を感じさせる。宮本ほか（2007）では「理学療法教育においては、学内教育や臨床実習の中で、英語文献の抄読が学生への課題とされる場面が散見される。」とされており、また宮本ほか（2011）では、「理学療法分野における国際化の波は近年なおも高まってきている。（中略）このような現状にあって、理学療法士養成課程に在籍する学生の中にも、国際化への関心を寄せる者も増加してきており、学内の英語教育へのニーズと重要性も高まってきている。」としている。さらに「理学療法士養成課程の英語教育は、依然構成根拠に乏しい状況にあるが、ESP の下位分野のひとつに位置づけ、詳細な分析結果に基づき教育を展開していく必要がある」としており、理学療法分野の英語が ESP の１つの分野として成り立っていくことに期待が寄せられている。しかしながら、Mitsuda（2009）では現役理学療法士を対象としたニーズ調査の結果、英語力に関しての質問項目で「英検３級程度」が最も多かったことが報告されており、理学療法学科を含む複数の実学系学科における、英語学習に関するニーズ調査を行った藤原（2011）でも調査対象者の英語力は「英検４級から英検３級程度」としている。このような点から英検３級程度の英語力を持つ学習者が、理学療法分野においては一定数いることが想定され、英検３級程度の英語から ESP への橋渡しを目的とした語彙研究は当該分野では意義があると考えた。

このような背景のもと、筆者は英検３級程度の英語から ESP への橋渡しを目的とした理学療法分野の語彙、語法に関して研究を行っているが、本研

究では、理学療法以外の医療系分野の論文コーパスを参照コーパスとした、理学療法分野の特徴語抽出に関して、どのような語が抽出されるか、抽出された語は理学療法分野において信頼性があるか、また作成された語彙リストは妥当なものであるかについて調査した結果を報告する。理学療法分野以外の医療系論文コーパスを参照するのは、医療系語彙の中にも、理学療法分野に特徴的でないものがあると考えられ、学習者の負担を軽減する観点から、とりわけ理学療法英語に特徴的な語彙を明らかにすることは意義があると考えたためである。また、本研究で Random Forests を用いるのは、Random Forests が近年、文学作品の著者判別などに用いられ、似通った性質を持つ資料の判別に有効性を発揮し、かつ判別に寄与した語彙の抽出が同時に行えることが報告されているためである。筆者が分析しようとする理学療法論文と、他の医療系分野の論文も同じ医療系分野に分類され、似通った性質を持っていることから、この方法を用いることとした。その他、この手法の詳細は後述する。このような背景から、本研究では Random Forests を用いる。抽出を目指す語は英検 3 級程度の英語から ESP への橋渡しを目的としたもので、一般的な文脈にも出現するが、理学療法分野の特徴も併せ持ち、英検 3 級程度から大学初級程度のレベルのまでの語とする。

1.2 Random Forests

Random Forests は Breiman（2001）開発のアンサンブル学習による分析法である。波部（2012：2）によると「Random Forests はその名が示しているとおり、複数の木（tree）を用いて森（forest）を構成して、判別などを行う機械学習アルゴリズムである。ここでいう木は決定木（decision tree）のことであり、個々の決定木は高い識別精能を持つわけではないが、それらを複数用いてそれぞれの結果を補うことによって高い予測性能を得ることが一つの特徴である。これは機械学習の分野ではアンサンブル学習（ensemble learning）と呼ばれており、個々の決定木がアンサンブル学習に

おける弱識別器（weak classifier）に相当している。（中略）Breiman による Random Forests は自らが提案したバギング（Bagging）を利用して、弱識別器となる決定木を構築するものである。」[1]とされている。

単純化して言い換えると、統計的な判別においては、1回の判断を「枝」とみなし、その組み合わせを1本の「木」と見立てそれを決定木と呼んでいる。大きなデータから無作為（Random）に抽出された複数のデータに基づき、それぞれから、決定木を構成した場合、個々の決定木の識別力は弱いが、それらを組み合わせること（アンサンブル、ここでは森）により、強い識別力を得る手法が Random Forests である。

また、言語研究において有用な Random Forests の特徴として、田畑（2012b：16）では「Random Forests の利点は、比較するコーパス（あるいはテクスト群）間において、一貫して生起に差異のある語彙項目が抽出されるため、特定のテクストに生起が偏る語彙項目が判別マーカーリストに混入する問題を未然に抑止できることであり、従来の特徴語抽出の問題点を補うものである。」とされており、分析対象資料において、特徴的かつ汎用性の高い語の抽出が可能であると考えられることから、この手法を特徴語抽出へ応用することとした。

なお Random Forests には抽出される語にランダム性がある。そのため、本研究では抽出語の安定性の検証を行ったうえで、この手法を用いる。

2. 先行研究

2.1　コーパスを用いた理学療法分野の英語に関する研究

理学療法分野の英語に関する研究は少ないが、本研究同様にコーパス分析を行ったものでは、次の4つの研究が挙げられる。宮本ほか（2007）では理学療法論文コーパスを分析し、後述の宮本ほか（2011）の基礎となる300語の教育用語彙選定が行われている。Mitsuda（2009）では一般英語から専門英語への橋渡しをする読解教材の開発が行われ、教材に用いる語彙選定の手

段として、理学療法論文コーパスの分析が行われた。宮本ほか（2011）では、宮本ほか（2007）の理学療法論文コーパスの再編纂と理学療法教科書コーパス編纂が行われ、これらの分析により ESP 語彙表の作成が行われた。宮本ほか（2012）では、宮本ほか（2011）と同等のコーパスで、共起パターンの特性について分析し、“severely disable”（先行研究表記通りに記載）のような「-ly 型副詞＋他動詞」などの 2 語の共起表現の抽出が行われている。このような中、理学療法以外の医療分野の論文データを参照コーパスとした特徴語の抽出、及び Random Forests による特徴語抽出は行われていない。

2.2　Random Forests を用いた言語研究

近年、Random Forests による、言語分析がいくつか行われている。金・村上（2007）は Random Forests とその他の分類法を用いて、異なる複数の著者が書いた、小説、作文、日記を分類法の正解率と、学習に用いた標本サイズとの関係に着目して分析を行った。この研究では Random Forests がその他の方法より、正解率が高く、学習に用いる標本サイズの減少による影響が小さいこと（標本サイズが小さい場合でも高い正解率が得られること）を明らかにしている。小林・田中・冨浦（2011）では、談話表現の頻度をもとに Random Forests により英語科学論文の分類を行い、さらに母語話者の論文と非母語話者の論文それぞれに特徴的な談話表現を抽出した。田畑（2012a）、田畑（2012b）では Random Forests を用いて、Charles Dickens と Wilkie Collins の作品を識別する指標となる語彙を抽出し、それらを変数とした多変量解析を行い、共著作品と Dickens および Collins それぞれの作品の関係を明らかにした。この研究では、従来の研究におけるカイ 2 乗値や対数尤度比による特徴語抽出では、少数の作品に特徴的に生起する固有名詞などが、全体の特徴語として抽出されるという問題点について指摘され、この点をカバーしうる手法として、特定のテクストに生起が偏る語彙項目が判別マーカーリストに混入する問題を未然に抑止できる Random

Forests が提案されている。

2.3　コーパスを用いた、特徴語抽出に関する研究

語彙表の作成、特徴語抽出に関する研究は古くからおこなわれており、West（1953）の General Service List（GSL）は代表的である。コーパスを用いた語彙研究では、Kŭcera and Francis（1967）、Hofland and Johansson（1982）、Leech, Rayson and Wilson（2001）などが挙げられ、国内の研究では、園田（1996）の北大語彙表、大学英語教育学会基本語改定委員会（2003）の JACET8000 などが代表的である。一般学術語彙の研究では Coxhead（2000）の Academic Word List（AWL）が代表的であるが、分野別で、本研究と近い医学分野の語彙研究では、Chung and Nation（2003）が挙げられる。

本研究では、Random Forests を特徴語分析に応用しようとすることから、様々な統計指標を用い、語彙抽出を行った先行研究について代表的なもの挙げ る。Chujyo and Utiyama（2006）は British National Corpus の “commerce and finance” のデータを 9 種類（頻度、ダイス係数、対数尤度比、コサイン、マクネマー検定、カイ 2 乗値、イェーツ補正公式、自己相互情報量、補完類似度）の統計指標を用いて、分析した研究である。結果として作成された語彙表では、これらの指標により、レベルの異なった語の抽出が可能であることが明らかとなり、これらの指標の教育用語彙表作成への応用の有用性が示唆された。

これらの先行研究、及び1章で述べた研究背景を踏まえ、本研究では、以下の点について明らかにする。

1. Random Forests を用いて、理学療法以外の医療系論文を参照コーパスとした特徴語抽出では、どのような語が理学療法分野の特徴語として抽出されるか。
2. 抽出された語は、理学療法分野の特徴語として信頼性があり、それらを含むリストは妥当と言えるか。

3. 方　　法

3.1 使用データ

3.1.1 英語理学療法論文コーパス

分析対象のコーパスとして、筆者編纂の英語理学療法論文コーパス（非公開、以下 PT と表記）を用いた。本コーパスは宮本ほか（2007）、宮本ほか（2011）、宮本ほか（2012）と同様に、アメリカ物理医学リハビリテーションアカデミー発行の *Archives of Physical Medicine and Rehabilitation* とアメリカ理学療法士協会発行の *Physical Therapy* の 2 誌の2007年から2008年に発行されたものから、論文のみを抽出しコーパスを構築した。これら 2 誌を選択したのには、宮本ほか（2011）が示すように、これら 2 誌が理学療法分野のリーディング・ジャーナルである事、また、理学療法分野のコーパス研究が少ない中、先行研究の資料との関連性を持たせることで、当該分野の研究の発展に貢献できると考えたためである。本コーパスは、総語数1,364,005語、異なり語数24,437語のコーパスである。語数は R を用いて算出した。以下このデータの表記は PT とする。

3.1.2 Corpus of Contemporary American English

参照コーパスには、Corpus of Contemporary American English（COCA）の Full Text データを用いた。COCA は Brigham Young 大学の Mark Davis 氏を中心とした研究グループにより構築された総語数464,020,256語の現代アメリカ英語の大規模汎用コーパスある。本研究では2014年 3 月に公開された COCA Full Text 版のサブコーパス Academic の Medicine の中から論文のみを抽出し、分析対象とした。総語数は1,753,922語、異なり語数39,682語である。本研究使用分のデータの語数は R を用いて算出した。

分析に先立ち次のような下処理を行った。分析対象データの抽出には COCA の発行元より提供されている、“genre”（academic、spoken など）、

“source”（テクストの掲載されている学術誌名など）、“title”（テクスト名）などが収録された“source”を参照した。Medicine に該当するファイルから、理学療法雑誌のファイルを除き、“source” を参照しても元データにアクセスできなかった 3 誌（*Creative Nursing* 2005–2006 のみ、*Practice Nurse*、*Occupational Health*）も分析対象外とした。その後、総語数2,000語に満たないファイルを除外した[2]。続いて、個々の文章の掲載雑誌の原本にあたり、Table of Contents から、Article, Original Article など、明らかに論文であることが記載されているものを抽出した。記載のないものに関しては、Introduction, Method, Result, Discussion の文章構造をもつものであるか否かを、本文にあたり確認し分類を行った[3]。なお各文章の冒頭につけられた “##” で始まる通し番号、改行を表すタグ、Conclusion より後に出現する、謝辞や参考文献なども取り除いた。記号やタグを除いたのは、それらが有効な文字として認識され、特徴語の 1 語として抽出されるのを避けるためであり、また Conclusion より後を削除したのは、PT に謝辞、参考文献が含まれておらず統一を図るためである。COCA では結論部より後に書誌情報に関する URL など不必要な情報が多く含まれており、これらが抽出されるのを避けるため除外した。この一連の処理を行い、本研究で用いた資料を以下 CM と表記する。

3.2　主なツール

本研究では、フリーソフトウエアの統計言語およびその実行環境である R（Ver. 3.1.2）で COCA の Full Text 版からのデータの切り出し、特徴語抽出及び頻度表作成、この頻度情報をもとにした分析を行った。Random Forests による分析では、R のパッケージ randomForest 4.6-10 をダウンロードして用いた。

3.3　分析方法

はじめに両コーパスに含まれる語を基本形に集約（レマ化）した。次に各

コーパスに含まれる論文ごとの単語頻度表（以下「ファイル別頻度表」と表記する）を作成し、1000語あたりの調整頻度を求め、この表から、数字や数を表す語（one, two など）、アルファベット１文字（“’s” の “s” など。I は一人称の代名詞の可能性があるため除外していない。また、不定冠詞の a も除外していない。）を除外した。続いて、ファイル別頻度表を用いて、各単語を変数とし、変数1000で10回 Random Forests を試行し、判別精度を確認した。次に Random Forests を50回試行し、判別に寄与した語の順位上位１位から1000位までに、50回すべての Random Forests 試行回に共通して抽出される語がどの程度含まれるかを調査した。（判別寄与語は Gini 係数の降順ソートで求めた。）この調査で共通して抽出された語を特徴語とし、現役理学療法士[4]による検証を行った。この検証では、協力者の理学療法士に、理学療法分野の臨床または文献購読で有用な語を選択してもらうことにより、抽出された各語の信頼性を検証し、信頼性のある語を多く含むかどうかにより作成された語彙リストの妥当性を検証した。また、特徴語に関しては、これに加えて、抽出された語に対しての見解を自由回答で述べてもらった。さらに抽出語を JACET8000 のレベルに分類し、抽出を目指す語のレベルが適切であるかどうかについても検証した。

4. 結果と考察

4.1　判別精度の確認

Random Forests 10回の試行の平均判別精度は97.04％であり、すべての試行回で95％を超える高い判別精度が得られた。このことから PT、CM の両資料の間には語彙の使用傾向に明らかに差がある事が確認できた。

4.2　抽出語の安定性の調査

Random Forests 試行50回に対しての判別寄与語上位１位から1000位までの共通抽出語の割合を調査した（図１）。図１から、判別寄与順位１位から

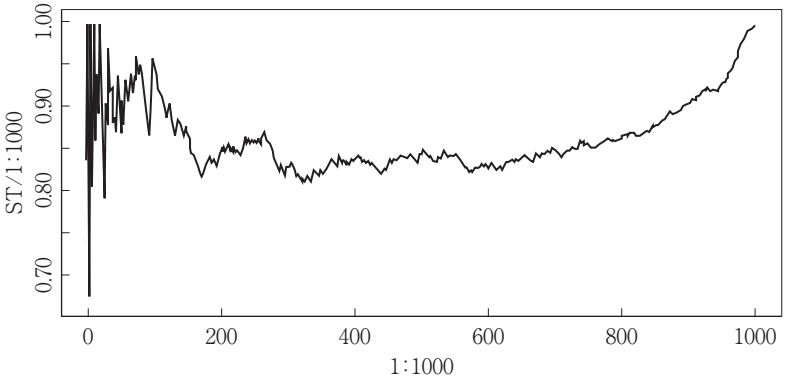

図 1　判別寄与語の順位に対する共通抽出語の割合（寄与度 1 位から1000位）

100位あたりまでの間で、高い割合で安定的に共通語が抽出されていることがわかる。表 1 に、判別寄与語の順位による共通抽出語割合を数値化したものを示す。

表 1 から、判別寄与語の順位 1 、 2 、 4 、 7 、12、21位に100％の割合で共通語が抽出されているのがわかる。しかし、この中で仮に 21 までの語を特徴語としたとしても、教育目的の語彙としては実用的な語数が抽出できたとは言い難い。そのため、実用的な語数をそなえた、95％前後の割合で共通語を抽出できる判別寄与語の順位に着目すると、100位から105位がそれにあ

表 1　判別寄与語の順位と共通抽出語の割合（抜粋）

判別寄与語の順位	1	2	4	7	12	21	34
共通語抽出の割合（％）	100	100	100	100	100	100	97.06
判別寄与語の順位	76	101	100	22	104	103	82
共通語抽出の割合（％）	96.05	96.04	96	95.45	95.19	95.15	95.12
判別寄与語の順位	102	79	78	77	74	36	35
共通語抽出の割合（％）	95.1	94.94	94.87	94.81	94.59	94.44	94.29
判別寄与語の順位	105	17	68	84	67	83	66
共通語抽出の割合（％）	94.29	94.12	94.12	94.05	94.03	93.98	93.94

てはまる。この点は Random Forests の抽出語の安定性を調査した八野(2015)[5]の結果とも一致し、当該資料における Random Forests による特徴語抽出では、Random Forests を複数回（本研究では50回）試行し、100位程度の共通抽出語を特徴語とすることが妥当であるといえる。次節では判別寄与順位105位における共通抽出語を特徴語として検討する。

4.3 特 徴 語

表2に特徴語を示す。表中の太字は PT の特徴語を表す[6]。表2を概観すると太字の割合が多いが、これは PT が理学療法分野単独のテクスト群で

表2　判別寄与語上位105位の共通抽出語

1 **rehabilitation**	26 **spinal**	51 **compare**	76 child
2 **fig**	27 **measure**	52 **speed**	77 **able**
3 **eg**	28 **participant**	53 **maximal**	78 **validity**
4 **functional**	29 **injury**	54 **shoulder**	79 **wheelchair**
5 **therapist**	30 **leg**	55 **recruit**	80 **criterion**
6 **subject**	31 **flexion**	56 the	81 us
7 **ie**	32 **gait**	57 inform	82 **force**
8 **motor**	33 **arm**	58 **sit**	83 **ankle**
9 **physical**	34 **knee**	59 **position**	84 **fatigue**
10 **muscle**	35 **pain**	60 difference	85 lt
11 **limb**	36 **score**	61 use	86 **visual**
12 **impairment**	37 **reliability**	62 non	87 **body**
13 **limitation**	38 **people**	63 with	88 study
14 **mobility**	39 **stroke**	64 **performance**	89 **velocity**
15 **extremity**	40 **task**	65 material	90 school
16 **movement**	41 **because**	66 **stand**	91 **strength**
17 **disability**	42 gender	67 **exercise**	92 drug
18 al	43 **sex**	68 between	93 analysis
19 **walk**	44 **joint**	69 **clinical**	94 **activity**
20 **perform**	45 **sci**	70 **motion**	95 **inclusion**
21 I	46 **hip**	71 figure	96 we
22 **conclusion**	47 **ability**	72 patient	97 risk
23 **trial**	48 **cognitive**	73 area	98 **test**
24 **function.**	49 public	74 scale	99 a
25 **consent**	50 exposure	75 **measurement**	

表 3　PT の特徴語

1	(1)	**ability**	25	(2)	**function**	49	(4)	**limb**
2	(1)	**able**	26	(2)	**injury**	50	(4)	**limitation**
3	(1)	**activity**	27	(2)	**knee**	51	(4)	**measurement**
4	(1)	**arm**	28	(2)	**muscle**	52	(4)	participant
5	(1)	because	29	(2)	**perform**	53	(4)	recruit
6	(1)	**body**	30	(2)	**physical**	54	(5)	**ankle**
7	(1)	**compare**	31	(2)	**scale**	55	(5)	**cognitive**
8	(1)	**exercise**	32	(2)	score	56	(5)	inclusion
9	(1)	**force**	33	(2)	sex	57	(5)	**mobility**
10	(1)	**leg**	34	(2)	**strength**	58	(5)	**validity**
11	(1)	**measure**	35	(2)	**task**	59	(5)	**velocity**
12	(1)	**movement**	36	(2)	**trial**	60	(6)	**rehabilitation**
13	(1)	**pain**	37	(3)	**disability**	61	(6)	**reliability**
14	(1)	people	38	(3)	**motion**	62	(6)	**therapist**
15	(1)	**performance**	39	(3)	**motor**	63	(6)	**wheelchair**
16	(1)	**position**	40	(3)	**stroke**	64	(7)	**fatigue**
17	(1)	**shoulder**	41	(3)	**visual**	65	(8)	**spinal**
18	(1)	**sit**	42	(4)	**clinical**	66	(9)	**extremity**
19	(1)	**speed**	43	(4)	**consent**	67	(9)	**flexion**
20	(1)	**stand**	44	(4)	criterion	68	(9)	**gait**
21	(1)	**subject**	45	(4)	fig	69	(9)	**impairment**
22	(1)	test	46	(4)	**functional**	70	(9)	**maximal**
23	(1)	**walk**	47	(4)	**hip**	71	(9)	**sci**
24	(2)	**conclusion**	48	(4)	**joint**			

あるのに対し、CM が医学・医療に関する様々な分野が混在するテクスト群であるため、単独分野の PT の特徴語が CM よりも多く抽出されたと考えられる。また同じ理由から、CM に the、a などの一般的な語が特徴語として抽出されたと考えられる。eg、ie に関して、PT では eg、ie と表記されるが、CM ではこれらの語は e.g.、i.e と表記され、それらがファイル別頻度表作成時に e と g、i と e に分割されたと考えられる。そのため、アルファベット1語をファイル別頻度表から除外した際に取り除かれ PT の特徴語として抽出されたといえる。このような背景から、この2語は特徴語とはしない。fig、figure に関しては、コンコーダンスを確認したところ略表記の fig が PT で顕著に使用される傾向が確認できた。次に表 3 に PT の特徴語を

示す[7]。表3では表2に含まれていたCMの特徴語、eg、ieを取り除き、また JACET8000 のレベル順に語を並べ替えた。（ ）内の数字は JACET8000 のレベルを表す[8]。

また以下に表3について、協力者の理学療法士から得られた見解を箇条書きで示す。

1. 語彙表には、馴染みのある言葉が多かった。
2. 語彙表の語は、（協力者が）理学療法学生になる以前なら意味が分からなかったであろう。
3. 語彙表の語は臨床に出てから日常的に使っている。
4. walk と gait では、とりわけ gait をよく使う。
5. walk は、“6 minuits walk test” などの表現では呼吸器疾患の術後や心疾患の患者に関する内容で使う。
6. 概観して、「歩行分析」と「脊椎損傷」に関連した語がやや多い。
7. 「歩行（分析）」は、理学療法分野内で基本的であるが、「脊椎損傷」に関しては分野内でもやや専門性が増す。表3の語の番号では63から71までに「脊椎損傷」に関連のあるものがやや多い。しかしながら臨床に出てからのことを考えると63から71までの語も役立つ語である。

上記の検証結果から、表3にはおおむね理学療法分野の特徴を表し、かつ基本的な語が抽出できているといえるが63から71までの語は分野内でやや専門性を備えた語であるため、導入の際には注意が必要である。

さらに、表3の太字の語（63から71除く）が含まれる PT の特徴表現を検討することで、表3に示した特徴語の有用性を検討したい。特徴表現抽出では、PT、CM 両コーパスから、表3に太字で示した語が含まれる4語の単語連鎖（4-gram）のうち PT、CM 両コーパスに合計で10回以上出現のある表現の抽出を行い、抽出された表現と頻度から G-score[9]を求めた。抽出された表現数は482であった。この資料を G-score をもとに降順ソートし、上位のものを PT の特徴表現とし、同資料を昇順ソートし上位のもの

を CM の特徴表現とした。PT の特徴度 1 位の表現は “activities of daily living” で G-score は 122.92 であった。CM の特徴度 1 位の表現は “in a body cast” で G-score は −27.62 であった。抽出された 482 表現のうち、有意水準 0.01 の表現を特徴表現とした。PT では 1 位の “activities of daily living” から375位の “good test retest reliability”（G＝6.75 p＝0.0094）までの 375 表現を PT の特徴表現とした。また CM 1 位の “in a body cast” から14位 “of inspiratory muscle training”（G＝−9.41 p＝0.0022）までの 14 表現[10]を CM の特徴表現とした。紙幅の都合上、表 4 には両コーパス特徴表現上位20項目（CM は14項目）のみを示す。

表 4　特徴表現

	PT	CM
1	activities of daily living	in a body cast
2	the subjects in the	child in a body
3	subjects with balance or	compared with non hispanic
4	freely chosen gait speed	clinical probability of pe
5	with acquired brain injury	to increase physical activity
6	subjects were instructed to	children in body casts
7	subjects were asked to	to prevent recurrent injury
8	the se of measurement	chronic conditions and disabilities
9	muscle strength and power	of low physical activity
10	in persons with sci	social and community activities
11	adults with mobility limitations	frequent poor physical health
12	of the gh joint	high performance liquid chromatography
13	in people with sci	interventions to increase physical
14	oa of the knee	of inspiratory muscle training
15	passive leg cycle exercise	—
16	compared with control subjects	—
17	the time of injury	—
18	persons with incomplete sci	—
19	test retest reliability of	—
20	in the sci population	—

この特徴表現は、PT の特徴語の有用性について検討するための資料であるため、PT の特徴表現についてのみ考察する。PT の特徴表現の 1 位を見

ると activities of daily living がきており、この表現は特徴語の協力者の理学療法士によると、理学療法分野で有用な表現であるとのことである。

さらに同様に協力者の理学療法士から信頼性について確認できた表現では部分的なものも含めて gait speed、the se of measurement、muscle strength and power、mobility limitations、oa of the knee などが挙げられる。これらの中には、JACET8000 ではレベル 1 に分類される一般的な語を含む物もあるが、理学療法分野では、そのような語の中にもその他の語との組み合わせにより、さらに活用の可能性が広がる語がある事が示唆された。今後は本研究で抽出された特徴語の周辺の語句、たとえば、上記の 4-gram 等を詳細にわたり分析することで、さらなる知見が得られることが期待できる。

5. おわりに

本研究では Random Forests を用い、医療系英語論文の大規模汎用コーパスである、COCA Full-Text 版 Academic Medicine を参照コーパスとして、英語理学療法論文の特徴語の調査を行った。理学療法士による検証では、臨床または文献の購読で有用な語が多く抽出され、抽出語の信頼性と作成された語彙リストの妥当性が検証された。このことから、Random Forests を用いることで、専門英語の語彙抽出でもテクスト群に一貫して生起し、かつそのテクスト群の特徴を表す語が抽出できることが示唆された。さらに抽出された語は、JACET8000 のレベル 1 から 4 程度の語が顕著であり、これらは英検 3 級程度から大学 1 年生程度の範囲に収まり、抽出を目指した語のレベルと一致した。これらの語は英検 3 級レベルの英語から ESP への橋渡しのための語彙としては適切であると考えられ、語のレベルも妥当であったといえる。またこれらの語は一般語としての側面も持ち合わせるため、一般的な文章にも出現すると考えられる。教育への応用では、そのような文章、たとえば学術論文ではない一般向けの医療系文章などを用いて、これらの語に

触れる読解を行った後に、読解の対象を論文にシフトしていくことで、段階的に当該分野の英語論文読解へ応用してゆくことができると考えられる。このほか、表４に示した語で JACET8000 のレベル５以上に分類された語、とりわけ表３の中の63番以後の語は、理学療法士による検証でもやや専門性が増すとされたことから、橋渡しのための語彙からは除き学習者の習熟度に応じて提示する必要がある事も示唆された。

しかしながら、本研究では妥当性と信頼性の検証に関しては限界があり、１名の理学療法士に依頼できたのみであった。今後の課題としては、複数の理学療法士による検証を行いたい。さらに、今後の研究では、本研究において抽出された語の周辺の語句を詳細にわたり検討し、英検３級程度の英語から理学療法分野の ESP への橋渡しに役立つ表現の抽出を行いたい。

註

1. 引用中の Bagging に関して、上田（2005：12）では「バギング法は単純アンサンブル学習法とまったく同じであるが、学習データを分割するのではなく、学習データから復元抽出して、K 組の学習データを作成し、ある識別器を k 通りの方法で独立に学習させ、最後にそれらを統合する方法である。」とされている。統計における学習とは、データに基づき判別基準を得ることである。
2. 論文抽出前に、ランダムに複数のファイルを確認したところ、総語数2000語未満のファイルに、論文以外の文章が複数見られたため。
3. PT 含まれるファイルはこの形式のため、統一を図った。
4. 臨床経験11年の現役理学療法士。理学療法分野では、勤務する病院の標榜する診療科により、理学療法士が担当する専門分野が分かれているが（“呼吸器”、“整形”など）、主要な専門分野のほとんどを担当した経験がある。また、日常的に理学療法分野の文献（英語を含む）を購読したり、定期的に専門に関する勉強会等に参加したりしている。
5. 八野（2015）では基本形への集約の際に、使用したレマリストにない語がレマ化されなかった場合などは、当該の語に対して加工は行っていないが、本研究では手動で修正を行った。
6. 表中の太字は PT の特徴語を表し、太字でない語は CM の特徴語を表す。
7. 表中の太字は理学療法士による妥当性と信頼性の検証において理学療法分野で

有用であると判断された語を表す。また、太字以外の語は理学療法分野においても用いられるが、太字の語ほど理学療法分野の特性を持たない語を表す。

8. JACET8000 のレベル分けには清水（2003）JACET8000 分析プログラム（v8an.pl）を使用した。（ ）内の数字は JACET 8000 のレベル指標（1〜8）を表す。(9) は JACET8000 の 8 段階に分類されない語を表す。

9. 表現抽出において Random Forests を行わなかったのは、PT-CM 間に10回以上出現する 4 語の単語連鎖では頻度が下がり、0 を含む表現が多く出てきたためであり、Random Forests では判別精度が80％台まで下がり、判別ができているとは言えない結果となったためである。そのため従来の特徴表現抽出で用いられている G-score を用いた。

10. PT と CM で抽出された特徴表現数に大きな差があるのは、この特徴表現の抽出法が、PT の特徴語を含む物であるためであり、CM の特徴語で特徴表現を抽出すると、CM の特徴表現が PT よりも多く抽出されることが予測される。なお本研究は、PT の特徴語の検証のための表現抽出であるため、CM の特徴語に関しては抽出を行わない。

参考文献

Breiman, L. (2001) “Random Forests.” *Machine Learning* 45 : 5-23.

Chujyo, K. and M, Utiyama. (2006) “Selecting level-specific specialized vocabulary using statistical measures” *System* 34 : 255-269.

Chung, T. M. and P. Nation (2003) “Technical vocabulary in a specialized texts.” *Reading in a Foreign Language* 15 2 : 103-116.

Coxhead, A. (2000) “A new academic word list.” *TESOL Quarterly* 34, 2 : 213-238.

大学英語教育学会基本語改定委員会（2003）『大学英語教育学会基本語リスト JACET List of 8000 Basic Words』社団法人大学英語教育学会.

藤原まみ（2011）「食物栄養学科・理学療法学科・作業療法学科・保育学科におけるキャリア教育としての英語教育――学生の実態とニーズ分析――」『九州 栄養福祉大学研究紀要」8：195-233.

波部　斉（2012）「ランダムフォレスト」『情報処理学会研究報告、コンピュータビジョンとイメージメディア』31：1-8.

Hofland, K. and S. Johansson (1982) *Word frequencies in British and American English*. Bergen : The Norweighan Computing Center for the Humanities.

八野幸子（2015）「Random Forests による特徴語抽出――抽出語の安定性の調査――」『統計数理研究所共同研究リポート 345　人文データのテキストマイニ

ングⅡ」41-51.
金　明哲・村上征勝（2007）「Random Forests 法による文章の書き手の同定」『統計数理』第55巻 2 号：255-268.
小林雄一郎・田中省作・冨浦洋一（2011）「Random Forests を用いた英語科学論文の分類と評価」、『情報処理学会研究報告、人文科学とコンピュータ研究会報告』第90巻 6：53-68.
Kŭcera, H. and W. H. Francis (1967) *Computational analysis of present-day American English*. Providence. RI: Brown University Press.
Leech, G. Rayson, P. and A. Wilson (2001) *Word frequencies in written and spoken English based on the British National Corpus* Harlow, UK Pearson Education Limited.
Mitsuda, S. (2009). "Developing ESP reading material for physical therapist." *On Language and Language Education* 2: 71-98.
宮本祥子・宮本謙三・宅間豊・井上佳和・竹林秀晃・岡部孝生・滝本幸治（2007）「理学療法教育における英語文献読解のための教育語彙選定：独自のコーパス分析を通して」『理学療法学』第34巻 6 号：260-266.
宮本祥子・五百蔵高浩・宮本謙三・宅間豊・井上佳和・竹林秀晃・岡部孝生・滝本幸治（2011）「理学療法分野における英語専門語彙（ESP 語彙）の抽出とその特性」『理学療法学』第38巻 6 号：421-435.
宮本祥子・五百蔵高浩・宮本謙三・宅間豊・井上佳和・竹林秀晃・岡部孝生・滝本幸治（2012）「理学療法分野における英語表現――共起パターンの特性――」『理学療法学』第39巻 2 号：90-101.
園田勝英（1996）『大学生用英語語彙表のための基礎的研究』言語文化部研究報告叢書 7　北海道大学.
田畑智司（2012a）「Dickens と Collins の共著作品への文体統計学的アプローチ」『情報処理学会研究報告、人文科学とコンピュータ研究会報告』第93巻 3 号：1-7.
田畑智司（2012b）「テクストマイニングからテクスト分析へ：Collins との共著作品における Dickens の文体」『電子化言語資料分析研究 2011-2012』3-17.
上田修功（2005）「アンサンブル学習」『情報処理学会論文誌、コンピュータビジョンとイメージメディア」46：11-20.
West, M. (1953) *A General Service List of English Words*, London : Longman Green and Co.
Corpus of Contemporary American English Full-Text（2015年 3 月30日閲覧）URL:http://corpus.byu.edu/full-text