『養生訓』の自動形態素解析における辞書の影響

相良 かおる（西南女学院大学 保健福祉学部）

どのような文書も適切に語分割できる汎用的な形態素解析用の辞書は存在しない。本稿では、形態素解析器 MeCab 用のシステム辞書である、現代語を対象とした①UniDic、②IPA 辞書、古文の解析用に作成された③近代文語 UniDic、④近世口語 UniDic、そして、医療記録文書の解析用に作成した実践医療用語辞書 ComeJisyo の⑤IPA 辞書と併用可能なユーザ辞書と⑥UniDic と併用可能なユーザ辞書、全 6 種類の辞書を組み合わせ、江戸中期に書かれた「養生訓」の校訂版テキストデータと抄訳テキストデータを形態素解析した結果について述べる。

The Impact of Dictionaries in Automatic Morphological Analysis of Yojokun

Kaoru Sagara (Faculty of Health and Welfare, Seinan Jo Gakuin University)

No general-purpose dictionary for morphological analysis is currently capable of appropriately dividing words from all types of documents. This paper discusses the results of combining six types of MeCab morphological analysis system dictionaries—namely, 1) UniDic and 2) IPADIC, which are dictionaries for contemporary Japanese; 3) Kindai Bungo UniDic (UniDic-kindai) and 4) Kinsei Kogo UniDic (UniDic-kinsei), which are compiled for analysis of ancient writings; 5) ComeJisyo, a dictionary of practical medical terminology compiled for analysis of medical records, which is a user dictionary compatible with IPADIC; and 6) a user dictionary compatible with UniDic—and performing morphological analysis of text data of a revised edition and an abridged translation of Yojokun, which was written in the mid-Edo period.

1. まえがき

筆者は、医療施設で蓄積された医療記録情報の自然言語処理を支援することを目的に、2004 年より看護実践用語の収集を開始し、2008 年に、Shift_JIS コードで入力された医療記録データの自然言語処理を支援するために、形態素解析器 Mecab[1]のユーザ辞書として利用可能な分かち書き用辞書 ComeJisyoV1（登録語数 30,146 語）の無償公開を開始し、以後随時更新を続け 2013 年 11 月からは ComeJisyoV5-1（登録語数 77,760 語）を公開している[2][3][4][5][6]。また、Utf-8 版 ComeJisyo（75,831 語）を作成し、公開予定である。

これらの辞書を作成する中で、①医療施設で扱う医療記録は多種多様であり、看護記録、医師記録等の医療従事者によって、また内科と歯科等の診療科によって、使われる用語に違いがあること、②学術論文や契約書等のテキストデータと異なり、表現や用語の標準化がなされておらず、方言や業界用語のみならず、表記のゆれや誤字脱字等が含まれることが分かっている。そして、医療用語を収集し、言語学的に解析することの重要性に加えて、これらの医療記録文書を精度良く解析するための網羅性の高い辞書の構築は困難であると認識している。

一般に日本語の形態素解析は、辞書を用い、辞書に掲載の見出し語を形態素として①語分割を行い、②品詞を推定し、③語形変化の処理を行う。

医療従事者（人）が入力した、医療記録データ（自然言語）のテキストマイニング等を行う際の最初の処理は、「意味のある単位」に文字列を分割する語分割であり、語分割を目的に形態素解析器を利用するには、「意味のある単位」を見出し語とした辞書が必要であるが、現実には網羅性の高い辞書の構築は困難である。

従って、自動形態素解析した後、利用者自らが用途に見合った結果か否か確認しなければならず、判断するだけの知識がない場合は、関連する辞典や事典等で調べることになる。すなわち、辞書や事典で調べられる程度の意味を含む単位で解析されることが望ましい。

形態素解析器 MeCab は、言語や辞書等に依存しない、汎用的な設計となっており、MeCab 用の形態素辞書には、IPA 辞書、JUMAN 辞書、UniDic 辞書等があり、MeCab の公式サイトで推奨されているのは IPA 辞書である[1]。

一方、現代語コーパスを元に作成されている IPA 辞書に対し、UniDic は、国立国語研究所の規定した斉一な言語単位（短単位）を見出し語とし、『日本語話し言葉コーパス(CSJ)』を元に作成されたものと『現代日本語書き言葉均衡コーパス（BCCWJ）』を元にして作成されたものに加えて、古文の解析用に①旧仮名口語 UniDic、②近代文語 UniDic、③近世口語（洒落本）UniDic、④中世口語（狂言）UniDic、⑤中世文語（説話・随筆）UniDic、⑥中古和文 UniDic、⑦上代（万葉集）UniDic が作成され公開されている[7]。

本稿では、これらの形態素解析用辞書を組み合わせ、江戸中期（日本史の歴史区分では「近世」）

に書かれた「養生訓」の中村学園大学校訂テキスト（以下、「校訂データ」という）[8]、並びに森下ジャーナル養生訓抄訳テキスト（以下、「抄訳データ」という）[9]を自動形態素解析した結果について報告する。

2. 調査データ

「養生訓」は、江戸中期の本草学者、儒学者である貝原益軒により書かれた全 8 巻から成る医学的教訓書であり、正徳 3 年（1713）に成立。和漢の事跡と体験に基づき、心身の健康と長寿を保つ養生法を通俗的に記したものである[10][11]。

本調査では、解析用データとして、Web 上で公開されている中村学園大学校訂テキスト[8]と、森下ジャーナル抄訳版[9]を用いる。

語分割の妥当性を調べるためにこれら 2 つの語分割データ（「妥当語分割データ」という）を作成する。なお、語分割の語の単位は、用途により異なる。今回の調査では、文法的規則を定めた厳密な正解データではなく、辞典・事典での辞書引きが可能な意味を持つ単位に分割する。

以下にこれらの概要を述べる。

2.1.1 中村学園大学校訂テキスト

概要：Web ページ[8]記載の概要を以下にまとめる。

益軒全集（明治 43 年）、有朋堂文庫本「益軒十訓」（大正 2 年）、貝原守一博士の校訂本、岩波文庫本および講談社学術文庫本を参考にして入力したものである。

漢字は現行の JIS 漢字に置き換えたが、仮名は旧仮名遣いのままにした。JIS 漢字で表せない漢字は、イメージとして付け加えた（図 1）。

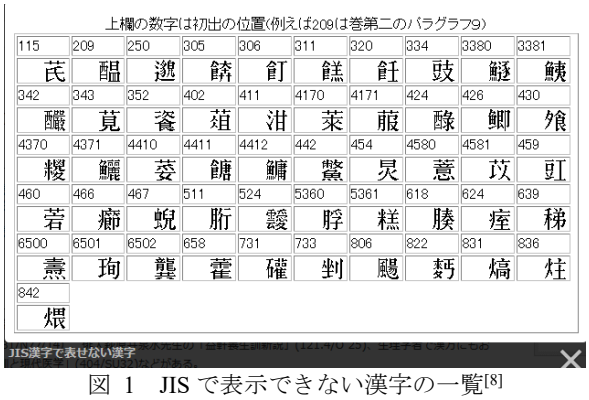

図 1　JIS で表示できない漢字の一覧[8]

送り仮名、読み、仮名書きの漢字化、言葉の説明など原文に無いものをカッコ内に挿入した。

検索や引用に便利なように各パラグラフに番号を付けたが、原文にない小見出しはつけなかった。全文を人手でキーボードによって入力した。

特徴：校訂データの特徴を以下に示す。

(1) 誤字や説明文、そして JIS 漢字で表せない漢字の一覧参照番号（図 1）等の挿入による非文が含まれる。

原文：べからず。、し(624)症(痙攣をおこす病気)となり、

正文：べからず。座症（痙攣をおこす病気）となり、

説明

・「座（JIS で表せない漢字）」を「し（624）」と表記。(624)は漢字の一覧表(図 1)の番号で、初出の位置（第 6 巻 24 段落目）を表す。

・「べからず。、」の「、」は誤入力

(2) 読み、意味、図 1 の番号等がカッコ内に含まれている。

原文：そうり(*理:肌のきめ)(618)いまだとちず。

正文：腠理いまだとちず。

説明

・「腠（JIS で表せない漢字）」を「そうり(*理:肌のきめ)（618）」と、平仮名で表し、加えて漢字の一覧表（図 1）の番号が記載されている。

(3) 段落前に半角数字で巻と段落が記載されている。

原文：[623]　熱食して汗いでば、風に当るべからず。

説明：[623]は、第 6 巻 23 段落目のこと。

(4) 平仮名の字数の割合は、55%と高く、次いで漢字の字数割合は、28%である（表 1）。

校訂データ：

ウェブブラウザを利用してテキストエディタにコピーし、文字コードを「UTF-8（BOM 無し）」、改行方法を「LF のみ（UNIX）」で保存したテキストファイルを用いる。

文字数：94,787 字（スペース含めない）

表 1　校訂データの字種割合

	文字数	割合
カタカナ	23	0.00
平仮名	51,913	0.55
漢字	26,928	0.28
その他	15,923	0.17
計	94,787	1.00

妥当語分割データ：

(1) 江戸中期（近世）の著作であることから、解析結果の異なり語数の多い（表 4）⑥近世口語 UniDic での解析結果を基に、形態素部分をテキストデータとして保存する。

(2) テキストエディタの置換機能により、改行

コードを空白に置換後、“EOS(end of String)”を改行コードに置換、更に空行を削除する。
(3) 目視で、辞書引きが可能な程度の意味を持つ単位に語分割する。
延べ語数： 63,585 語
異なり語数： 6,640 語

2.1.2 養生訓抄訳テキスト

抄訳データ：

森下ジャーナルの Web ページ[9]よりウェブブラウザを利用してテキストエディタにコピーし、文字コードを「UTF-8（BOM 無し）」、改行方法を「LF のみ（UNIX）」で保存したテキストファイルを用いる。

文字数：73,843 字（スペース含めない）

表 2 抄訳データの字種割合

	文字数	割合
カタカナ	464	0.01
平仮名	44,048	0.60
カタカナ	464	0.01
漢字	21,207	0.29
その他	7,660	0.10
計	73,843	1.00

妥当語分割データ：

(1) 現代語で書かれていることから、③IPA 辞書による解析結果の形態素部分をテキストデータとして保存する。
(2) テキストエディタの置換機能を用い、改行コードを空白に置換後、“EOS(end of String)”を改行コードに置換し、更に空行を削除する。
(3) 目視により、辞書引きが可能な程度の意味を持つ単位に語分割する。
延べ語数： 49,953 語
異なり語数： 4,493 語

2.2.3 解析用辞書

本調査で用いる形態素解析器 mecab-0.996[1]用の辞書を以下に示す。

【システム辞書】
(1) IPA 辞書（392,126 語）[1]
(2) 現代書き言葉 UniDic（書字形 327,670 語）[12]
(3) 近代文語 UniDic（書字形 349,964 語）[12]
(4) 近世口語 UniDic（書字形 339,505 語）[12]

【ユーザ辞書】
(5) ComeJisyoUTF8-1.0（75,831 語）
(6) UniDic 用 ComeJisyo（75,089 語）

3. 方 法

調査 1：校訂データと抄訳データの 2 種類について以下の①から⑥の 6 パターンの辞書により形態素解析した結果を調べる。
① 現代書き言葉 UniDic
② 現代書き言葉 UniDic ＆ UniDic 用 ComeJisyo
③ IPA 辞書
④ IPA 辞書 ＆ ComeJisyoUTF8-1.0
⑤ 近代文語 UniDic
⑥ 近世口語 UniDic

調査 2：妥当語分割データと照合する。

調査 3：実践医療用語辞書 ComeJisyo の影響を②および④より調べる。

調査 4：校訂データに対応する抄訳データの解析結果について考察する。

4. 結 果

表 3 は、それぞれの解析結果の延べ語数をまとめたものである。校訂データでは、斉一な短単位を見出し語とする①現代書き込言葉 UniDic での解析が最も細かく語分割され、現代語で書かれた抄訳データにおいても UniDic 系の辞書を用いた、①と②、⑤と⑥の延べ語数が多かった。

表 3 形態素解析結果（延べ語数）

辞書	校訂データ	割合	抄訳データ	割合
①	**68,413**	1.08	**49,352**	1.03
②	**68,349**	1.07	**49,157**	1.03
③	67,789	1.07	48,828	1.02
④	67,697	1.06	48,533	1.01
⑤	66,107	1.04	**49,305**	1.03
⑥	65,878	1.04	**49,374**	1.03
妥当語分割データ（再掲）	63,585	1.00	47,952	1.00

表 4 形態素結果（異なり語数）

辞書	校訂データ	割合	抄訳データ	割合
①	5,475	0.82	4,293	0.96
②	5,484	0.83	4,356	0.97
③	5,755	0.87	4,328	0.96
④	5,793	0.87	**4,400**	0.98
⑤	5,957	0.90	4,361	0.97
⑥	**6,131**	0.92	4,384	0.98
妥当データ（再掲）	6,640	1.00	4,493	1.00

表 4 は、それぞれの解析結果の異なり語数をまとめたものである。校訂データでは⑥近世口語 UniDic の、抄訳データでは④IPA 辞書＆ComeJisyoUTF8-1.0 の解析結果の語数が多かった。

表 5 妥当語分割データと一致する校訂データ

	平仮名語	割合	混種語	割合	漢字語	割合
①	1,498	0.66	671	0.65	2,308	0.75
②	1,498	0.66	672	0.65	2,321	0.76
③	1,298	**0.57**	596	**0.58**	2,060	0.67
④	1,299	**0.58**	597	**0.58**	2,093	0.68
⑤	1,823	**0.81**	906	0.88	2,366	0.77
⑥	1,936	**0.86**	953	0.93	2,387	0.78
妥当校訂	2,259	1.00	1,029	1.00	3,070	1.00

表 5 は、妥当語分割データに一致する校訂データの異なり語数である。「平仮名のみの語（以下、「平仮名語」という）」では、⑤近代文語 UniDic および⑥近世口語 UniDic での解析において、一致する異なり語数の割合は 81%と 86%、「混種語」では 88%と 93%と高く、現代語文を対象とする IPA 辞書を用いた解析③と④においては、平仮名語の一致語の割合は 57%と 58%、混種語では、共に 58%と低かった。

表 6 妥当語分割データに一致する抄訳データ

	平仮名語	割合	混種語	割合	漢字語	割合
①	832	0.87	1,132	0.90	1,839	0.85
②	833	0.87	1,131	0.90	1,874	**0.87**
③	917	**0.96**	1,194	**0.95**	1,871	0.86
④	918	**0.96**	1,193	**0.95**	1,913	**0.88**
⑤	818	0.86	1,100	0.88	1,839	0.85
⑥	807	0.85	1,109	0.88	1,844	0.85
妥当語分割データ（再掲）	954	1.00	1,256	1.00	2,166	1.00

表 6 は、妥当語分割データに一致する抄訳データの異なり語数である。平仮名語および混種語では、IPA 辞書を用いる③と④の割合が 96%と 95%と高く、「漢字のみの語（以下、「漢字語」という）」の割合は①～④まで約 85%と同程度であり、ComeJisyo を併用する②と④においては、87%と 88%で若干高くなっていた。

表 7 は、解析結果における ComeJisyo の登録語数である。現代語で書かれた抄訳データにおける語数が多かった。

表 7 解析結果②④に含まれる ComeJisyo の登録語

	延べ語数	異なり語数
② 校訂	89	38
② 抄訳	212	117
④ 校訂	254	65
④ 抄訳	317	143

表 8 解析結果②④に含まれる ComeJisyo 登録語例

	校訂・抄訳	校訂	抄訳
1	下血	気虚	こむらがえり
2	外邪	気血	ふくらはぎ
3	香蘇散	気滞	胃もたれ
4	両肩	香蘇散	運動不足
5	両膝	清暑益気湯	栄養不足
6	肋骨部	津液	温泉療法
7	脾	不食	気分転換
8	脾虚	補中益気湯	滋養強壮
9	茯苓	老人病	食欲不振
10		臍下	肥満防止

表 8 は一致した登録語の一例である。校訂データと抄訳データに共通する登録語が 9 語（「下血」「外邪」「香蘇散」「両肩」「両膝」「肋骨部」「脾」「脾虚」「茯苓」）あった。なお、④で出力された校訂データの解析結果に含まれる ComeJisyo の登録語 65 語の内 44 語は、⑥近世口語 UniDic による解析結果の中に含まれていた。

表 9 ①～⑥共通の不一致語数（異なり語）

	校訂	抄訳
平仮名語	**262**	11
混種語	70	39
漢字語	**581**	164
計	913	214

表 9 は、妥当語分割データと一致しない「平仮名語」「混種語」「漢字語」の異なり語数の内、①～⑥全てに共通の語数の一覧である。ここでの「一致しない」とは、辞書の見出し語が細かく過分割される場合（「熱　症（熱症）」）と、連接する異なる見出し語を含む境界の誤り（「其　物にく　は　ふ　べき（其物に　くわふべき）」）を指す。校訂データの解析結果において、平仮名語および漢字語が多かった。

表 10 平仮名語の共通不一致語例

	校訂・抄訳	校訂	抄訳
1	いわたけ	うったい	あわもり
2	ちょうろぎ	あえしお	いさごまい
3		かゞむ	うたかぐさ
4		こゝろみ	じゅつだま
5		しろうり	どれだけ
6		とゞこほり	ぬりこむ
7		にえばな	まくわうり
8		とうちんこう	まなす
9		のんど	もちごめ
10		ほしゐまゝ	

表 10 は、表 9 における平仮名語の例である。校訂データと抄訳データに共通の不一致語は、「いわ　た　け（いわたけ」と「ちょう　ろ　ぎ（ちょうろぎ）」の 2 語であった。

抄訳データの UniDic を用いた①と②の解析結果では、「しまいこむ」「飲みこむ」「ぬり　こむ」となっていた。一方、IPA 辞書を用いた③と④の解析結果では、「ぬり　こむ」同様に「しまい　こむ」と 2 語になっていた。

表 11 混種語の共通不一致語例

	校訂・抄訳	校訂	抄訳
1	たき香	かな書	イ草科
2	ひねり艾	げん醋	うす醤油
3	黄ぎ	なめ味噌	さし身
4	孫思ばく	ねぶり臥す事	ちらし薬
5	虹げい	の玉へり	ひゆ菜
6	艾ちゅう	ひねり艾	塩から
7		わか死に	食べ過ぎ
8		筋引つり	食中たり
9		砂かん	生ねぎ
10		食すゝみ	腹鳴り

表 11 は、表 9 における混種語の例である。校訂データと抄訳データに共通の不一致語は 6 語（「たき香」「ひねり艾」「黄ぎ（黄耆）」「孫思ばく（孫思邈）」「虹げい（虹蜺）」「艾ちゅう」）であった。

表 12 漢字語の共通不一致語例

	校訂・抄訳	校訂	抄訳
1	海菜	医統正脈	胃腸病
2	掛香	飲茶烟草附	虚弱体質
3	灸瘡	陰血	魚鮓
4	強壮剤	衛生宝鑑	古井戸
5	局方発揮	黄栢	香茶餅
6	禁灸	牙根	自然死
7	鼓子花	外台秘要	生大根
8	五宜	各致余論	節飲
9	振薬	寒症	鎮嘔剤
10	生葱	軒岐救生論	頤生輯要

表 12 は、表 9 における漢字語の例である。漢方薬の名前や薬草名、「局方発揮」や「衛生宝鑑」等の中国や朝鮮の医書名が多く見られた。

表 13 は、校訂データの解析例である。⑥近世口語 UniDic の解析結果では、「ほしゐ」と「こらゑ」、「外邪」と「七情」が、正しく切り出されている。

表 14 は、現代語で書かれた抄訳データの解析例である。古語が含まれていないため、UniDic 系の辞書を用いた場合の解析結果に相違はみられなかった。

5. 考察とまとめ

今回の調査の目的は、「どのような文書も適切に語分割できる汎用的な形態素解析用の辞書は存在しない」という立場から、以下の 5 点とした。

(1) 江戸中期に書かれた『養生訓』の校訂データの形態素解析において、同時代の資料を対象とした「近代文語 UniDic」および「近世口語 UniDic」の有用性を調べること。

(2)古文に含まれる医療用語の抽出における ComeJisyo の有用性および課題を調べること。

(3)現代語で書かれた抄訳データの形態素解析における現代書き言葉 UniDic および IPA 辞書の有用性を調べること。

(4)『養生訓』の解釈における、分野および時代の異なるテキストを対象とする形態素解析用辞書を用いた解析結果（分ち書き）の有用性を調べること。

(5)校訂データと抄訳データの解析による古語と現代語の対応付けの可能性を調べること。

目的(1)：表 5 より、現代書き言葉 UniDic による解析①と②における平仮名語の一致割合 66%に比べて、⑤近代文語 UniDic では 81%、⑥近世口語 UniDic では 86 %と高く、混種語においては、①と②の一致割合 65%に対して⑤は 88%、⑥は 93%と高かった。なお、「のゝしる」「はゞ」「ほしゐまゝ」等、仮名一字を単位とする踊り字「ゝ」「ゝ」「ゞ」を含む 24 語が、過分割されており、これらについては、解析前に対応する文字に変換する前処理を行うか、「Web 茶まめ」の前処理機能を使って解析することで、⑤および⑥の平仮名語の一致割合は更に向上すると考えられる[13]。

また、平均的な日本語の平仮名の割合 52.9%と漢字の割合 28%に対し[14]、校訂データの割合は 55%と 28%であり、平仮名の割合が若干高いことから、校訂データの和語の解析において、⑥近世口語 UniDic が有用であると考えられる。

目的(2)：⑥における「漢字語」の一致割合は 78%と低く、漢方薬の名前や中国、朝鮮で書かれた医書名、そして症状や病名等、校訂データの特徴語が正しく解析されなかった。加えて表 7 より解析結果に含まれる現代語の ComeJisyo の登録語数は、②で 38 語、④で 65 語と少なく、また④の 65 語の内 44 語は、⑤近代文語 UniDic および⑥近世口語 UniDic の解析結果に含まれていた。

また校訂データに含まれる医療用語で現在も使われているものは、漢方薬名程度である（表 12）。従って、江戸時代以前に書かれた医学的な文書から医療用語を抽出するための ComeJisyo

の利用は有効ではないと考える。

目的(3)：表 6 より、現代書き言葉 UniDic を用いた解析結果①と②における平仮名語の一致割合 87%に比べて、IPA 辞書を用いた③と④では 96%と高く、混種語では、①と②の 90%に対して、③と④では 95%と高かった。また、漢字語では、IPA 辞書を用いた③の 86%に対し、UniDic 系の解析①および⑤と⑥は 85%と同じ値であった。これは、現代書き言葉 UniDic に見出し語を追加して古文用辞書を作成しているためである[15]。更に ComeJisyo を併用した②は 87%、④は 88%と一致割合は向上し、解析結果に含まれる ComeJisyo の登録語数は②では 117 語、④では 143 語であった(表 7)。以上のことから、抄訳データの分かち書きにおいては、④IPA 辞書＆ComeJisyoUTF8-1.0 が有用であると考えられる。

なお、表 11 および表 12 から、抄訳データにおける不一致語（未知語）には、「うす醤油」や「さし身」等の食べ物の名前や「胃腸病」、「自然死」等、医療記録に含まれる用語も含まれ、ComeJisyo を更新する際には、これらを追加する予定である。

目的(4)：平仮名の多い文章は、現代語文であっても単語の境界が分かり難く読み辛い。また、辞書を使った自動形態素解析においても「て」「に」「を」「は」等の助詞の認定誤りによる過分割が生じる。そして今回の校訂データおよび抄訳データは平仮名が多く、語の境界が分かり難い。

校訂データには、難解な語に送り仮名、仮名書きの漢字化、言葉の説明等がカッコ付きで付加されており、解読することが目的である場合、分野および時代の異なる辞書を用いて分ち書きする必要はないかもしれない。しかし古文を学ぶ初心者にとっては、解析誤りが多くあったとしても、分ち書きされている方が原文よりも読みやすいのではないだろうか（表 13、表 14）。

目的(5)：今回、抄訳データを用いたため、解析結果から古語と現代語の対応表を作成することは不可能であった。

今回の調査で、短単位を見出し語とする UniDic において、「しまいこむ」「飲み込む」は 1 語に、「ぬり　こむ」は 2 語に解析されることが、IPA 辞書においては、「飲み込む」は 1 語に、「しまいこむ」「ぬり　こむ」は 2 語に解析されることが明らかとなった。

このことから、意味に重きを置く場合、「溜めこむ」「のぞき　こむ」は「溜める」「のぞく」から意味を調べることができ、2 語でも構わないが、「ひっこむ」「しこむ」「ふれこむ」を 2 語にするのは都合が悪い。用途によっては、文法規則に則った単位が適切ではない場合があることが示唆された。

今後、語単位に配慮した実用的な実践医療用語辞書 ComeJisyo を目指し、医療施設で蓄積される医療記録データの利用方法等、現場のニーズを把握する予定である。

また、水上茂樹先生[16]が貝原益軒アーカイブの作成に着手された 1995 年から 20 年余りが経ち、図 1 の文字は UTF コードで表示できるようになった。JIS で表せない文字を原文の漢字に置き換えた改訂版の校訂データを作成し、言語資源として公開したいと考えている。

謝辞

本研究は、2018 年度西南女学院大学共同研究費の助成を受けています。

養生訓の校訂版を作成し、公開して下さった水上茂樹先生に感謝致します。

参考文献

[1] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://taku910.github.io/mecab/ (参照 2018-10-24)

[2] 相良かおる,浅原正幸,小野正子,小作浩美：形態素解析器 MeCab 用看護用語ユーザ辞書の作成と公開,第 28 回医療情報学連合大会論文集, p.938-939, 2008

[3] 相良かおる,浅原正幸,小野正子,外山健二：形態素エンジン MeCab 用辞書 ComeJisyoV２および看護教育支援用かな漢字変換辞書の作成と公開, 第 29 回医療情報学連合大会論文集,p.983-984, 2009

[4] 相良かおる,小野正子,小木曽智信,小作浩美: 電子医療記録の分ち書き用ユーザ辞書 ComeJisyo の紹介と単語生起コスト,言語処理学会 第 18 回年次大会 発表論文集,p. 621-624, 2012

[5] 相良かおる,小野正子,小作浩美,鈴木隆弘,高崎光浩,嶋田元：分かち書き用辞書 ComeJisyo の評価,医療情報学 第 32 巻 第 6 号,p.301-307,2012

[6] 相良かおる,小野正子：実践医療用語辞書 ComeJisyo の紹介,第 33 回医療情報学連合大会論文集,p.828-830,2013

[7] UniDic: http://unidic.ninjal.ac.jp/ (参照 2018-10-24)

[8] 養生訓 中村学園大学校訂テキスト：http://www.nakamura-u.ac.jp/library/kaibara/archive03/text01.html (参照 2018-10-24)

[9] 養生訓（抄訳）森下ジャーナル：http://home.att.ne.jp/theta/mo/you/index.html (参照 2018-10-24)

[10] 広辞苑 第 5 版,岩波書店.

[11] 講談社カラー版　日本語大辞典　第 2 版,講談社,1995.

[12] 鴻野知暁：『日本語歴史コーパス』に出現した新規語の『UniDic』への登録について: https://pj.ninjal.ac.jp/corpus_center/lrw/LRW2016-tkouno-slides.pdf (参照 2018-10-24)

[13] 小木曽智,近代語テキストの形態素解析 - 国立国語研究所：https://pj.ninjal.ac.jp/corpus_center/cmj/doc/05ogiso.

pdf(参照 2018-10-24)
[14] 刀山 将大 ,佐藤 理史,近藤 秀,吉田 達平: 日本語の文の平均像を体現した文を探す (1) 文の特徴量の抽出, 第 13 回情報科学技術フォーラム,第 2 分冊,p.217-218,2014
[15] 小木曽 智信, 小町 守, 松本 裕治 : 歴史的日本語資料を対象とした形態素解析, 自然言語処理,20 巻 5 号,p.727-748,2013
[16] 水上茂樹 : https://www.aozora.gr.jp/index_pages/person1494.html(参照 2018-10-24)

表 13 校訂データの解析結果例

辞書	語数	分かち書き
	原文	養生の術は、先(ず)わが身をそこなふ物を去べし。身をそこなふ物は、内慾と外邪となり。内慾とは飲食の慾、好色の慾、睡の慾、言語をほしゐままにするの慾と、喜・怒・憂・思・悲・恐・驚の七情の慾を云。外邪とは天の四気なり。風・寒・暑・湿を云。内慾をこらゑて、すくなくし、外邪をおそれてふせぐ、是を以(て)、元気をそこなはず、病なくして天年を永くたもつべし。
①	138	養生の術は、先(ず)わが身をそこなふ物を去べし。身をそこなふ物は、内慾と外邪となり。内慾とは飲食の慾、好色の慾、睡の慾、言語をほしゐままにするの慾と、喜・怒・憂・思・悲・恐・驚の七情の慾を云。外邪とは天の四気なり。風・寒・暑・湿を云。内慾をこらゑて、すくなくし、外邪をおそれてふせぐ、是を以(て)、元気をそこなはず、病なくして天年を永くたもつべし。
②	135	養生の術は、先(ず)わが身をそこなふ物を去べし。身をそこなふ物は、内慾と外邪となり。内慾とは飲食の慾、好色の慾、睡の慾、言語をほしゐままにするの慾と、喜・怒・憂・思・悲・恐・驚の七情の慾を云。外邪とは天の四気なり。風・寒・暑・湿を云。内慾をこらゑて、すくなくし、外邪をおそれてふせぐ、是を以(て)、元気をそこなはず、病なくして天年を永くたもつべし。
③	142	養生の術は、先(ず)わが身をそこなふ物を去べし。身をそこなふ物は、内慾と外邪となり。内慾とは飲食の慾、好色の慾、睡の慾、言語をほしゐままにするの慾と、喜・怒・憂・思・悲・恐・驚の七情の慾を云。外邪とは天の四気なり。風・寒・暑・湿を云。内慾をこらゑて、すくなくし、外邪をおそれてふせぐ、是を以(て)、元気をそこなはず、病なくして天年を永くたもつべし。
④	139	養生の術は、先(ず)わが身をそこなふ物を去べし。身をそこなふ物は、内慾と外邪となり。内慾とは飲食の慾、好色の慾、睡の慾、言語をほしゐままにするの慾と、喜・怒・憂・思・悲・恐・驚の七情の慾を云。外邪とは天の四気なり。風・寒・暑・湿を云。内慾をこらゑて、すくなくし、外邪をおそれてふせぐ、是を以(て)、元気をそこなはず、病なくして天年を永くたもつべし。
⑤	134	養生の術は、先(ず)わが身をそこなふ物を去べし。身をそこなふ物は、内慾と外邪となり。内慾とは飲食の慾、好色の慾、睡の慾、言語をほしゐままにするの慾と、喜・怒・憂・思・悲・恐・驚の七情の慾を云。外邪とは天の四気なり。風・寒・暑・湿を云。内慾をこらゑて、すくなくし、外邪をおそれてふせぐ、是を以(て)、元気をそこなはず、病なくして天年を永くたもつべし。
⑥	133	養生の術は、先(ず)わが身をそこなふ物を去べし。身をそこなふ物は、内慾と外邪となり。内慾とは飲食の慾、好色の慾、睡の慾、言語をほしゐままにするの慾と、喜・怒・憂・思・悲・恐・驚の七情の慾を云。外邪とは天の四気なり。風・寒・暑・湿を云。内慾をこらゑて、すくなくし、外邪をおそれてふせぐ、是を以(て)、元気をそこなはず、病なくして天年を永くたもつべし。

表 14　抄訳データの解析例

辞書	語数	分かち書き
	原文	健康法の第一は、体を損なう原因をはぶくことにある。その原因は体の内にあるものと外から入ってくるものがある。※体の内にあるものは、自分自身の欲望を押さえられないことによるものがある。外から入ってくるものは、環境によるものである。※自分の欲望のまま生活しないことや、環境の変化にたいして常に注意していれば、健康で元気に暮らせ、病気にかかることもなく寿命をまっとうできる。※
①	118	健康 法 の 第 一 は 、 体 を 損なう 原因 を はぶく こと に ある 。 その 原因 は 体 の 内 に ある もの と 外 から 入っ て くる もの が ある 。 ※ 体 の 内 に ある もの は 、 自分 自身 の 欲望 を 押さえ られ ない こと に よる もの が ある 。 外 から 入っ て くる もの は 、 環境 に よる もの で ある 。 ※ 自分 の 欲望 の まま 生活 し ない こと や 、 環境 の 変化 に たいして 常 に 注意 し て いれ ば 、 健康 で 元気 に 暮らせ 、 病気 に かかる こと も なく 寿命 を まっとう できる 。 ※
②	117	健康 法 の 第 一 は 、 体 を 損なう 原因 を はぶく こと に ある 。 その 原因 は 体 の 内 に ある もの と 外 から 入っ て くる もの が ある 。 ※ 体 の 内 に ある もの は 、 自分自身 の 欲望 を 押さえ られ ない こと に よる もの が ある 。 外 から 入っ て くる もの は 、 環境 に よる もの で ある 。 ※ 自分 の 欲望 の まま 生活 し ない こと や 、 環境 の 変化 に たいして 常 に 注意 し て いれ ば 、 健康 で 元気 に 暮らせ 、 病気 に かかる こと も なく 寿命 を まっとう できる 。 ※
③	114	健康 法 の 第 一 は 、 体 を 損なう 原因 を はぶく こと に ある 。 その 原因 は 体 の 内 に ある もの と 外 から 入っ て くる もの が ある 。 ※ 体 の 内 に ある もの は 、 自分 自身 の 欲望 を 押さえ られ ない こと による もの が ある 。 外 から 入っ て くる もの は 、 環境 による もの で ある 。 ※ 自分 の 欲望 の まま 生活 し ない こと や 、 環境 の 変化 にたいして 常に 注意 し て いれ ば 、 健康 で 元気 に 暮らせ 、 病気 に かかる こと も なく 寿命 を まっとう できる 。 ※
④	112	健康 法 の 第一 は 、 体 を 損なう 原因 を はぶく こと に ある 。 その 原因 は 体 の 内 に ある もの と 外 から 入っ て くる もの が ある 。 ※ 体 の 内 に ある もの は 、 自分自身 の 欲望 を 押さえ られ ない こと による もの が ある 。 外 から 入っ て くる もの は 、 環境 による もの で ある 。 ※ 自分 の 欲望 の まま 生活 し ない こと や 、 環境 の 変化 にたいして 常に 注意 し て いれ ば 、 健康 で 元気 に 暮らせ 、 病気 に かかる こと も なく 寿命 を まっとう できる 。 ※
⑤	118	健康 法 の 第 一 は 、 体 を 損なう 原因 を はぶく こと に ある 。 その 原因 は 体 の 内 に ある もの と 外 から 入っ て くる もの が ある 。 ※ 体 の 内 に ある もの は 、 自分 自身 の 欲望 を 押さえ られ ない こと に よる もの が ある 。 外 から 入っ て くる もの は 、 環境 に よる もの で ある 。 ※ 自分 の 欲望 の まま 生活 し ない こと や 、 環境 の 変化 に たいして 常 に 注意 し て いれ ば 、 健康 で 元気 に 暮らせ 、 病気 に かかる こと も なく 寿命 を まっとう できる 。 ※
⑥	119	健康 法 の 第 一 は 、 体 を 損なう 原因 を はぶく こと に ある 。 その 原因 は 体 の 内 に ある もの と 外 から 入っ て くる もの が ある 。 ※ 体 の 内 に ある もの は 、 自分 自身 の 欲望 を 押さえ られ ない こと に よる もの が ある 。 外 から 入っ て くる もの は 、 環境 に よる もの で ある 。 ※ 自分 の 欲望 の まま 生活 し ない こと や 、 環境 の 変化 に たいし て 常 に 注意 し て いれ ば 、 健康 で 元気 に 暮らせ 、 病気 に かかる こと も なく 寿命 を まっとう できる 。 ※