一般口演
# 一般口演10
# 看護情報１（看護記録のＩＣＴ化とデータ活用）
2018年11月24日(土) 09:00 〜 11:00 C会場 (4F 411+412)

[3-C-1-7] UTF版実践医療用語辞書 ComeJisyo1.0の作成
○相良 かおる[1], 山崎 誠[2], 小野 正子[1] （1.西南女学院大学, 2.国立国語研究所）

本発表では、 Unicodeで正規化された医療情報の語分割用辞書 UtfComeJisyo1.0の概要を述べる。　筆者等は、Windows環境において Shift_JISコードで入力された医療情報を語分割するために、2008年に形態素解析器MeCabのユーザ辞書として利用可能な実践医療用語辞書 ComeJisyo（登録語数30,146語）を作成・公開し、以降随時更新し、2013年に ComeJisyoV5-1（登録語数77,760語）を公開している。なお、「実践医療用語」とは、市販の医療用語辞書ではカバーされていない隠語や略語を含む医療現場で使われている実践的な医療用語を言う。　今回発表する UtfComeJisyoV5は、 Utf-8（ BOM無し）環境での医療情報、主として看護領域の文書の語分割を可能とする。　登録語は、 ComeJisyoV5-1の登録語を対象に Unicodeの NFKC形式に正規化した75,089語を登録している。従って、半角カタカナ、全角英数字、そして機種依存文字は含まれない。また、属性として、 ComeJisyoV5-1に倣い、以下の属性を付加している。　・看護経過記録、プログレスノート、看護教育用模擬経過記録、模擬診療記録、医師経過記録における文書頻度　・看護師、助産師、管理栄養士の国家試験問題文における出現の有無　・看護師および管理栄養士養成校で採用する教科書の索引における出現状況 外国人受験者を考慮して2011年以降の看護師国家試験問題文に併記されている疾病名の英語

UTF 版実践医療用語辞書 ComeJisyo1.0 の作成

相良かおる[*1]、小野正子[*1]、山崎誠[*2]
*1 西南女学院大学、*2 国立国語研究所

Introducing ComeJisyo 1.0 - A UTF-8-based medical term dictionary

Kaoru Sagara[*1], Masako Ono[*1] ,Makoto Yamazaki[*2]
*1 Seinan Jo Gakuin University, *2 National Institute for Japanese Language and Linguistics,

This paper reports on the development of ComeJisyo 1.0, an updated medical dictionary based on UTF-8 encoding that adopts the MeCab open source part-of-speech morphological analyzer and text segmentation library. While numerous software designers have adopted UTF-8 based formats in recent years, the current ComeJisyo V5-1 medical dictionary is still based on Shift Japan Industrial Standards (Shift_JIS) Japanese character encoding, which cannot display a significant number of JIS third-level Chinese characters. This is an important issue because five of the standard printing fonts used in the tests for Japanese national nurse examinations, “癨”, “塡”, “剝”, “頬”, and “嚢”, contain such characters. The ComeJisyo 1.0 library contains 75,723 registered words, each accompanied by information as to whether or not it has appeared in questions of the Japanese national examinations for nurses, registered dietitians, and midwives within the past five years. Additionally, the dictionary provides document frequency information on five kinds of medical record data, as well as English language equivalent words and the English names of the diseases or conditions that are described in the nurse examination questionnaire.

Keywords: Medical terminology, Morphological analysis, Dictionary, Word segmentation, UTF-8.

1．はじめに

筆者らは、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、用語の分析[1)2)3)]と看護用語の標準化に関する調査研究[4)5)6)]を行った。また同年には看護支援システムの稼働状況を調査するために電子カルテシステムが稼働または一部稼働している施設 50 施設を訪問し、電子カルテシステムに詳しい看護師等に半構成的面接による調査を行い、入力環境を視察した[7)]。

その結果を踏まえ 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 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[8)9)10)11)12)]。

一方、電子カルテシステムの普及に伴い、医療分野においてもテキストマイニングを用いた研究が成されるようになってきた。

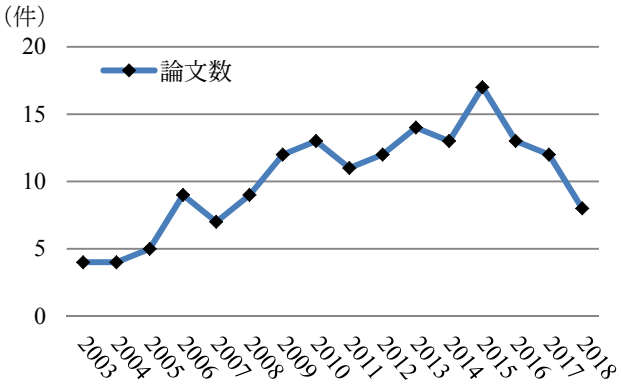

図 1 CiNii 論文検索結果（2018 年 8 月 1 日現在）

図 1 は、CiNii の論文検索において「テキストマイニング」と「医療」の AND 検索の結果 163 件の分布である。

近年、テキストマイニングや統計ソフト等のソフトウェアの多くが前提とする符号化文字集合は、Unicode であり、その符号化形式は、OS に依存しない UTF コードである。

医療記録データをこれらのソフトで処理するための語分割に、Shift_JIS コードで作成された ComeJisyoV5-1 を UTF コードに変換して利用する利用者も現れてきた。

しかしながら、Shift_JIS で扱える文字全てが Unicode で扱える訳ではなく、ComeJisyoV5-1 の登録語に含まれる半角カタカナや機種依存文字を UTF コードに変換すると文字化けが生じ、文字コードの変換だけで利用することは困難である。

そこで、我々は、UTF 版の辞書 ComeJisyo1.0（以下、本辞書という）を作成し公開することとした。

本発表では、本辞書の概要と精度実験の結果について述べる。

2. 用語の定義

本稿で用いる用語の定義を以下に示す。

JIS X 0208（JIS 基本漢字）: 日本で初めて規定された、漢字を含む日本語の符号化文字集合。JIS 第 1 水準 2,965 字、JIS 第 2 水準 3,390 字、英数字・カタカナ・符号など非漢字 524 字、合計 6,879 文字が収録されている[13)14)]。

JIS X 0221（国際符号化文字集合（UCS））: Unicode の内、日本語でよく使われる文字を集めた部分集合。JIS 第 1 水準から第 4 水準迄の漢字が全て含まれる[13)14)]。

Shift_JIS: 昭和 57 年（1982 年）にマイクロソフト、アスキー、三菱電機などが共同で開発した日本語の符号化文字集合で、JIS X 0208（6,879 文字）を利用した文字符号化方式であり、パソコン用の標準的な文字コードとして広く普及している。英数字や片仮名、そして句読点等の1バイト文字 191 字を含む 7,070 文字が収録されている。①～⑳や㈱、㎏等の環境依存文字を独自に実装した変種が多くある[13)14)]。

Unicode: 全世界共通で使えるように、世界中の文字に単一の符号を対応付けようと開発された符号化文字集合。最新の Unicode10.0.0 の収録文字数は、136,690 字である[15)]。

UTF: Unicode の符号化方式[13)14)]。

UTF-8: ASCII と互換性がある 8 ビット単位の Unicode の符号化方式[13)14)]。

BOM: ファイルの先頭に付けるバイト順マークであり、Byte Oder Mark の頭文字である[13)14)]。

NFKC 形式: Unicode の正規化方式の一つで、NFKC は、

Normalization Form Compatibility Compositionの頭文字である。正規化対象となるものには、Shift_JIS 等との互換用に導入された全角・半角がある。「全角A」は「半角 A」に、「半角ｱ」は「全角ア」に、ローマ数字の「Ⅱ」は大文字半角英字「I」の 2 文字「II」に、丸数字「①」は半角数字「1」に変換される。また、三点リーダ「…」はピリオド「. . .」となる。

出現登録語数（重なり有）: 対象テキストデータの中に含まれる本辞書の登録語数で、以下のように求める。

例えば対象テキストデータに「橈骨動脈触知良好」が含まれる場合、本辞書には「橈骨動脈触知良好」と「触知良好」が登録されていることから、出現登録語数（重なり有）は、「橈骨動脈触知良好」と「触知良好」の 2 語となる。

出現登録語数（重なり無）: 対象テキストデータの中に含まれる本辞書の登録語数で、以下のように求めている。

例えば対象テキストデータに「橈骨動脈触知良好」が含まれる場合、出現登録語数（重なり無）は、文字長の長い「橈骨動脈触知良好」1語となる。

形態素解析: 言語学的な意味での「形態素」に分割することではなく、電子化辞書を用いてソフトウェアで自動解析すること。

解析登録語数: 形態素解析結果に含まれる本辞書の登録語数。

3. UTF 版 ComeJisyo1.0 の概要

本辞書は、Utf-8（BOM 無し）環境において、医療情報、主として看護領域の文書の語分割のために、形態素解析器 MeCab のユーザ辞書としての利用を想定して作成した辞書である。

以下にその概要を述べる。

登録語数:

Shift_JIS 版 ComeJisyoV5-1 の登録語 77,760 語を、Unicode 正規化形式 NFKC 形式に変換して得られた 75,089 語と JIS 第 3 水準の漢字である印刷標準字体「掻」、「塡」、「剝」、「頬」、「嚢」を含む 634 語の計 75,723 語を登録している。

扱える文字の制約:

本辞書が扱える文字は、英数字は半角のみ、カタカナは全角のみである。また、ローマ数字 （Ⅲ ⅲ Ⅳ ⅳ等）、丸数字 （① ⑫ 等）、機種依存文字 （㎏ ㌘ ㈱ ㍻ ㌘等）は扱えない。

一方、JIS X 0221 の利用により、JIS 第 1 水準から第4水準の全漢字を扱うことが可能となり、看護師国家試験の冊子体で使われる、Shift_JIS では扱えない第 3 水準の漢字「剝」、「掻」、「頬」、「嚢」、「塡」は扱うことができる[16]。

付加情報:

本辞書独自の付加情報としては、ComeJisyoV5-1 に倣い、以下の属性を付加している。

1) 看護経過記録、プログレスノート、看護教育用模擬経過記録、模擬診療記録、医師経過記録における文書頻度
2) 看護師、助産師、管理栄養士 5 年分の国家試験問題文（2013 年-2017 年）における出現の有無
3) 看護師および管理栄養士養成校で採用している教科書 54 冊の索引における出現状況
4) 外国人受験者を考慮して 2011 年以降の看護師国家試験問題文に併記されている疾病名の英語

4.形態素解析精度に関する実験

今回、現場で入力された医療記録データと、看護師国家試験問題を実験データとし、以下の実験を行った。

4.1 実験 1:医療記録の解析

Shift_JIS の環境において他職種により入力された個人情報を含まないプログレスノート3,000 行（94,036 文字）をNFKC 形式に正規化後、UTF-8（BOM 無し）に変換したものを対象テキストデータとし、本辞書を用いて解析した。

以下に実験の手順を示す。

1) 実験データに含まれる英数字を半角文字に変換
2) 同様に半角カタカナを全角文字に変換
3) 形態素解析器として Mecab0.996 を利用し、本辞書をユーザ辞書とし、IPA 辞書をシステム辞書として解析
4) 対象テキストデータに含まれる出現登録語数（重なり有）を求める
5) 対象テキストデータに含まれる出現登録語数（重なり無）を求める

4.2 実験 2:看護師国家試験問題の解析

看護師国家試験問題 5 年分（2013 年-2017 年）のテキストデータ（7,578 行、173,622 字）を対象とし、以下、看護国試データという、前述の 1）～5）の処理を行った。

5. 結果

5.1 医療記録の形態素解析

表 1、表 2 は、解析結果の語数を求めたものである。

本辞書に登録されている語は、名詞であることから、名詞に占める割合を求めている。

表 1 医療記録の解析結果の概要

	延べ語数	異なり語数
全解析語数（EOS 除く）	56,563	6,018
名詞の数	24,931	4,879
名詞内の解析登録語数	4,599	1,979
出現登録語数（重なり有）	6,199	2,383
出現登録語数（重なり無）	4,730	2,022

全解析語数の名詞の占める割合は延べ語数で 44％（=24,931 語／56,563 語）、異なり語数で 81％（4,879 語／6,018 語）であった。

表 2 名詞に占める登録語数の割合（医療記録）

	名詞に占める割合	
	延べ語数	異なり語数
名詞の数	100.0%	100.0%
名詞内の解析登録語数	18.4%	40.5%
出現登録語数（重なり有）	24.9%	48.8%
出現登録語数（重なり無）	19.0%	41.4%

これらの名詞に含まれる本辞書の登録語の割合は、約 41％（=1,977 語／4,879 語）であった。

図 2 は、それぞれの語数の関係を表したものである。

出現登録語数（重なり無）における解析登録語数の割合は 97％であった。解析されなかった残り 3％（62 語）は、29 語がカタカナおよび平仮名のみからなる語であった。また 62 語の殆どが、以下に示すように複数の単語が臨時的に結びついた臨時一語の一部分であった。

例：「上部消化管出血疑い」の解析に対して、本辞書には

「消化管出血疑い」と「上部消化管出血」が登録されており、今回の実験では、「上部」と「消化管出血疑い」に分ち書きされたため、「上部消化管出血」が未解析語となった。

なお、「消化管出血疑い」は、出現登録語数（重なり無）に含まれない登録語 19 語（図 2）の中に含まれている。

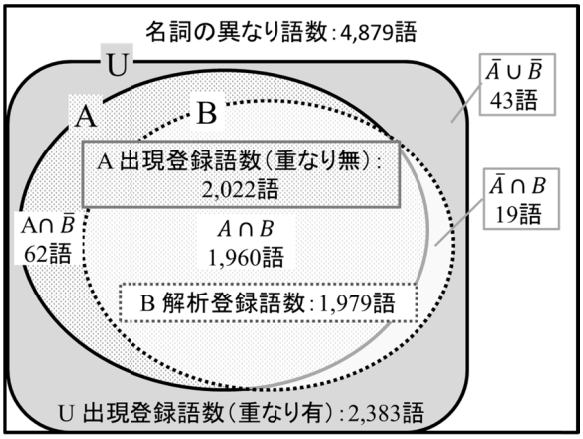

図 2　解析結果の概要（医療記録）

$$R=\frac{A\cap B}{A}=\frac{1{,}960\text{ 語}}{2{,}022\text{ 語}}\times 100=97\%\quad \ldots(1)$$

5.2 看護師国家試験問題文の解析

看護国試データの解析結果の概要は、表 3、表 4 の通りである。

全解析語数に占める名詞の割合は、延べ語数で 53.0%（=51,376 語／96,943 語）、異なり語数で 85%（=8,031 語／9,461 語）であった。

表 3 看護師国家試験データの解析結果の概要

	延べ語数	異なり語数
全解析語数（EOS 除く）	96,943	9,461
名詞の数	51,376	8,031
名詞内の解析登録語数	7,489	3,244
出現登録語数（重なり有）	11,602	4,156
出現登録語数（重なり無）	7,932	3,344

これらの名詞に含まれる本辞書の登録語の割合は、約 40%（=3,244 語／8,031 語）であった。

図 3 は、それぞれの語数の関係を表したものである。

表 4　名詞に占める登録語数の割合（看護国試）

	名詞に占める割合	
	延べ語数	異なり語数
名詞の数	100.00%	100.00%
名詞内の解析登録語数	14.58%	40.39%
出現登録語数（重なり有）	22.58%	51.75%
出現登録語数（重なり無）	15.44%	41.64%

出現登録語数（重なり無）における解析登録語数の割合は医療記録のそれと比べて若干低く 96%であった。解析されなかった残り 4%（131 語）の文字長の中央値は 2 文字、平均は 3 文字であり、医療記録データの実験同様に複数の単語が臨時的に結びついた本辞書に未登録の臨時一語の一部分であった。

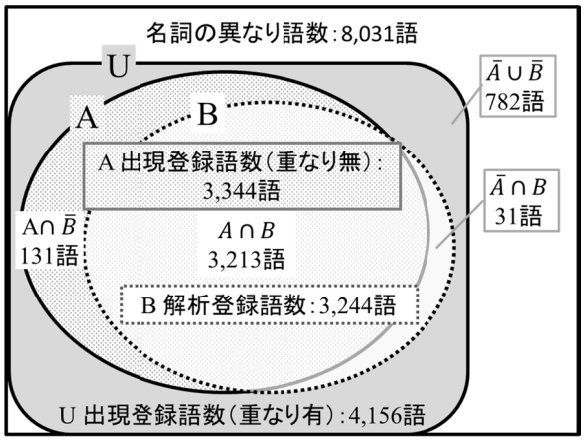

図 3 解析結果の概要（看護国試）

$$R=\frac{A\cap B}{A}=\frac{3{,}213\text{ 語}}{3{,}344\text{ 語}}\times 100=96\%\quad \ldots(2)$$

表 5　出力結果の例

解析語	付加情報
血液検査	名詞,サ変接続,*,*,*,*,血液検査,ケツエキケンサ,ケツエキケンサ ,◆,：,,,；,,,：,看助栄教,：,5,：,12846
と	助詞,並立助詞,*,*,*,*,と,ト,ト
尿検	名詞,サ変接続,*,*,*,*,尿検,ニョウケン,ニョウケン ,◆,：,,,；,,,：,看教,：,3,：,24337
査	名詞,一般,*,*,*,*,*
の	助詞,連体化,*,*,*,*,の,ノ,ノ
結果	名詞,副詞可能,*,*,*,*,結果,ケッカ,ケッカ
長男	名詞,一般,*,*,*,*,長男,チョウナン,チョーナン
は	助詞,係助詞,*,*,*,*,は,ハ,ワ
胃瘻	名詞,一般,*,*,*,*,胃瘻,イロウ,イロウ ,◆,：,gastric fistula,；,,,：,看栄教,：,4,：,7853
の	助詞,連体化,*,*,*,*,の,ノ,ノ
造設	名詞,サ変接続,*,*,*,*,造設,ゾウセツ,ゾウセツ ,◆,：,,,；,,,：,看栄教,：,4,：,20798
を	助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
希望	名詞,サ変接続,*,*,*,*,希望,キボウ,キボー
せ	動詞,自立,*,*,サ変・スル,未然ヌ接続,する,セ,セ
ず	助動詞,*,*,*,特殊・ヌ,連用ニ接続,ぬ,ズ,ズ
防護用	名詞,接尾,一般,*,*,*,防護用,ボウゴヨウ,ボウゴヨウ, ,◆,：,,,；,,,：,看教,：,1,：,44406
具	名詞,接尾,一般,*,*,*,具,グ,グ
は	助詞,係助詞,*,*,*,*,は,ハ,ワ
どれ	名詞,代名詞,一般,*,*,*,どれ,ドレ,ドレ
か	助詞,副助詞／並立助詞／終助詞,*,*,*,*,か,カ,カ

表 5 に形態素解析結果の一例を示す。「尿検査」が「尿検」と「査」に過分割されている。これは医療記録で使われる略語の「尿検」と、「尿検査」の両方が本辞書に

登録されており、「尿検」が優先されて生じた過分割の例である。

表 5 の「◆」以降の情報は、本辞書独自の情報である。外国人看護師向けの教育等、教育での利用を想定し、看護師国家試験に付加される英語、看護師、助産師、管理栄養士国家試験設問文での出現情報、医療記録での文書頻度を付加している。例えば、「胃瘻」に付加された情報は、英語訳は“gastric fistula”であり、看護師および管理栄養士の国家試験問題に出現し、また養成校で利用する教科書の索引に出現していることを表している。なお、最後の数字“7,853”は、登録語を維持・管理するためのものである。

6. 考察とまとめ

今回、UTF コード版の実践医療用語辞書 ComeJisyo1.0 を作成し、解析精度実験の結果を示した。

形態素解析の精度を示す指標として、再現率と適合率、そしてこれらを用いたF値が用いられる[17]。本稿では、適合率は求めていない。これは、適合率が解析結果を利用する上の本辞書の適合の度合いを表す訳ではないためである。また前述の再現率 97%と 96%も本辞書の解析精度の評価指数として適切とは言い難い。なぜなら、本辞書が、対象データに含まれる全ての医療用語を網羅していないためである。そして、前述の例にあるように「上部消化管出血疑い」を構成する「上部消化管出血」と「消化管出血疑い」が共に本辞書に登録されており、これらには共通部分「消化管出血」があり、再現率も適合率も 100%になることはないためである。

また、利用者の望む解析結果が「上部｜消化管出血疑い」なのか、「上部消化管出血｜疑い」なのか、それとも「上部｜消化管出血｜疑い」なのかも分からない、すなわち唯一の正解を定めることができないためである。

このような語の係り受けの問題は、複数の単語が連結している複合語等を登録語とする辞書の持つ問題である。

また、IPA辞書のみで解析すれば「防護｜用具」となるところを本辞書の登録語に「防護用」があるために、これが優先され、「防護用｜具」のように、不適切な場所で分割される場合もある(表 5)。

このような誤解析を減らすためには、登録語を拡充し、網羅性を高める必要がある。医療記録に含まれる全ての語を登録した辞書の作成は困難であるが、小児科の看護記録用というように利用目的が限られていれば、網羅性の高い辞書の作成は可能である。

今回作成した辞書は、JIS 第 1 水準から JIS 第 4 水準までの漢字が扱え、冊子体の看護師国家試験問題で使われる印刷標準字体「掻」、「塡」、「剝」、「頬」、「嚢」を含む語の解析が可能である。

一方、実際の医療施設では Shift_JIS を使用している場合が少なくないことから、半角カタカナやローマ数字、丸数字等を含む用語の解析が出来ない本辞書は、「実践医療用語辞書」というよりは、教育・研究目的での利用に適している。

臨床の場で働く医療従事者に、2 次利用で用いられる統計ソフト等の解析ツールが使用する文字コードを意識した入力、例えば「丸数字やローマ数字、“㎎”や“㎏”等の機種依存文字、または半角カタカナを入力しない」というような負担を強いることを筆者等は望んではいない。従来の Shift_JIS 版の実践医療用語辞書 ComeJisyo の改善・登録語の拡充等は継続する予定である。

従って、利用目的により本辞書と従来の辞書を使い分け、活用して頂ければと思う。

謝辞

本研究は、西南女学院大学共同研究費の助成を、並びに一部は、科学研究費補助金（18H03499）による補助を受けています。

参考文献

1) 相良かおる: 看護記録に含まれる文書の統語構造,日本医療情報学会 第 5 回看護情報研究会論文集,pp.85-88. 2004
2) 相良かおる, 小野正子, 鈴木隆弘, 嶋田元, 小作浩美: 看護記録文の計量的用語調査, 人文科学とコンピュータシンポジウム, p.103-110, 2010
3) 小木曽智信, 相良かおる: 医療分野で使われる複合語の語種構成, 第 29 回社会言語科学会研究大会発表論文集, p.158-161, 2012
4) 相良かおる,小作浩美,小暮潔: 標準看護実践用語の特徴 , 第 6 回看護情報研究会論文集, P.73-75. 2005
5) Kaoru Sagara, Akinori Abe, Hiromi itoh Ozaku, Noriaki Kuwahara, and Kiyoshi Kogure: Features of Standardized Nursing Terminology Sets in Japan, In Proceedings of the 9th on Nursing Informatics (NI2006) , p.471-475, 2006
6) 相良かおる,小作浩美,小暮潔,納谷太,桑教則彰: 看護文書の意味解析用辞書の構築における ICNP®と「分類語彙表」の活用可能性, 医療情報学 第 24 巻 第 6 号, p.657-665, 2005
7) 相良かおる, 黒田裕子, 小田正枝, 岡崎寿美子, 山勢博彰, 城戸茂里, 平尾百合子, 棚橋泰之, 林みよ子, 脇坂浩, 中木高夫: 看護支援システムの稼動状況 予備的研究としての半構成的面接調査報告, 看護診断学会 第 11 巻第 1 号 pp.18-28 2006.
8) 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
9) 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
10) 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
11) 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
12) 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
13) 深沢千尋著. 文字コード超研究. ラトルス.2009.
14) 矢野啓介. 文字コード技術入門. 技術評論社. 2010.
15) unicode® 10.0.0 2017 June 20(Announcement) (https://www.unicode.org/versions/Unicode10.0.0/)
16) 相良かおる,橋本直幸,小野正子. 看護師・助産師・管理栄養士国家試験に含まれる漢字調査. 第 19 回日本医療情報学会看護学術大会論文集（2018.7）. P.137-140. 2018.
17) 情報処理学会編. 言語処理学事典. 共立出版株式会社. 2009