Vol.2019-CH-121 No.1
2019/8/1

研究文献目録の高次資源化と汎用化をめざして —国文学研究文献目録データベースを事例として

相田満[1]　栗城大地[2]　野本忠司[3]

概要：「国文学論文目録データベース」は、明治・大正・昭和・平成時代の雑誌・論集に収載される研究文献の目録をデータベース化したもので、令和元年となった 2019 年度には 60 万件を超える件数を公開している。網羅的データベースやリポジトリによる全文データベースなどの公開が普及する中で、目録の維持・メンテナンスに割かれてきた労力が今後も継続するに値するものであるか、その存在意義を問われつつあることも確かである。本データベースに於いても、その独自性を主張しえるだけの変容・変質を迫られている。そこで、本データベースにおいては、過去に蓄積された人力による 60 万件の論文データに付与された情報自体を資源として、新たに採録するためのエキスパートシステムの糧とするとめの方途を求めるための取り組みを進めることにした。本発表は現段階の準備状況を確認するとともに、その構想を実現の可能性と発展性について報告したい。

キーワード：文献目録データベース，機械学習，自然言語処理，オントロジ

Aiming for Higher-order Resources and Generalization of Research Bibliography -Case Study of the Database of Research Thesis in Japanese Literature

MITSURU AIDA[†1]　DAICH KURIKI[†2]
TADASHI NOMOTO[†3]

Abstract: The "Database of Reseach Thesis in Japanese Literature" is a database of research literature catalogs published in Meiji, Taisho, Showa, Heisei era magazines and journals.It has released more than 600,000 cases in the year 2019, which was the year of the First in Reiwa ERA. As public disclosure of full-text databases and the like by exhaustive databases and repositories becomes widespread, it is also being questioned whether their efforts devoted to the maintenance and maintenance of catalogs are worthy to continue in the future, or their significance. It is true that in this database, there is a need for transformation and alteration that can claim its uniqueness,too,
Therefore, in the present database, an effort to find a way to serve as a source of an expert system for newly collecting information, using as a resource the information itself added to the data of 600,000 articles by human power accumulated in the past Decided to advance. This presentation will confirm the current state of preparation and report on the feasibility and development of the concept.

Keywords: IPSJ Journal, MS-Word, Style files, "Dos and Dont's" list

1. はじめに

「国文学論文目録データベース」は、明治・大正・昭和・平成時代の雑誌・論集に収載される研究文献の目録をデータベース化し、令和元年となつた 2019 年度には 60 万件を超える件数を公開している。 採択文献の採録方法は、採録者が全文を通覧した上で分野別に類別を行い、作品・作者名などの論文に現れない関連語や、ヨミ、館蔵書については函架番号などの情報付加を行うことで、Cinii や国会図書館などから公開される網羅的データベースにはないエキスパート的知識が反映された専門特化型のデータベースとなっており、国文学研究者には不可欠な存在となっている。

しかしながら、網羅的データベースやリポジトリによる全文データベースなどの公開が普及する中で、目録の維持・メンテナンスに割かれてきた労力が今後も継続するに値するものであるか、その

1 国文学研究資料館
National Institute of Japanese Litrature
2 国文学研究資料館
National Institute of Japanese Litrature
3 国文学研究資料館
National Institute of Japanese Litrature
Kyoto University
†3 奈良先端大学院大学
Nara Institute of Science and Technology

存在意義を問われつつあることも確かで、本データベースに於いても、その独自性を主張しえるだけの変容・変質を迫られている。 そこで、本データベースにおいては、過去に蓄積された人力による60万件の論文データに付与された情報自体を資源として、新たに採録するためのエキスパートシステムの糧とするとめの方途を求めるための取り組みを進めることにした。 そのための方策としては、 ①過去の論文資源に付加された分類語・付加情報の有用性の確認 ②付加・整備されるべき情報資源（含オントロジ）の整備と不足する資源の蓄積 ③論文全文情報からの語彙抽出と分類 ④入力作業および情報付加作業の効率化あるいは自動化などが考えられる。本発表は現段階の準備状況を確認するとともに、その構想を実現の可能性と発展性について報告したい。

2.「国文学研究文献目録データベース」の概要と位置づけ

日本文学研究の基盤を形成する研究文献の総合目録データベース「国文学論文目録データベース」は、国文学という概念が定着する以前の明治から大正・昭和・平成時代の雑誌および論集に収載される研究文献の目録をデータベース化したものである。2019 年 5 月 17 日現在の論文データの登録件数は 600,741 件、即応性には欠ける傷みはあるが、2-3 年遅れのスパンではありながら、毎年 11,000-12,000 件程度のデータを累加して現在に至っている。

データベースの利用状況も、たとえば 2018 年度の館全体の検索のべ数を示せば、

A.図書・雑誌所蔵目録（OPAC）： 710,613 件

B.日本古典籍総合目録データベース： 632,444 件

C.国文学論文目録データベース（CSV データ対応）：405,822 件

と、上位から数えて常に 3 位以内に入っており、ＰＶユニークユーザ数・ＰＶのべ数の等の視点に於いても同様に上位 3 位以内のアクセス数を誇る。

文献の採録方法は、採録者が全文を通覧した上で分野別に類別を行い、作品・作者名などの論文に現れない関連語や、ヨミ、館蔵書については函架番号などの情報付加を行うことで、CiNii や国会図書館などから公開される網羅的データベースにはないエキスパート的知識が反映された、いわば専門特化型のデータベース*1 となっており、国文学研究者には不可欠な存在となっている。

現状の文献目録データベースは、網羅的 DB と専門特化型 DB に大別され、現在は以下のようなカテゴリに分けることができるといえよう。

網羅的 DB

CiNii(国立情報学研究所)

JDreamⅢ（FUJITSU ジー・サーチ）［有料］

NDL ONLINE(国立国会図書館)

網羅的 DB （採録件数 10 万件以上のもの 他は割愛した）

国文学論文目録データベース（国文学研究資料館）

教育研究論文索引検索（国立教育政策研究所教育図書館）

日本語研究・日本語教育文献データベース（国立国語研究所）

東洋学文献類目検索（京都大学）

その他、研究者や研究テーマから文献にアクセスするこが可能となる Researchmap（科学技術振興機構）や、Google Scholar も著者と主題検索において、速報性と検出数において良好な結果を示してきており、上記網羅的 DB と連繋する機関リポジトリも原著の提供や収録件数において有効な検索結果を提供している。

3. 「国文学研究文献目録データベース」の特長と可能性

(1) 時間的推移と研究分野の消長

上記の内、「国文学研究文献目録データベース」の特長は、研究者の便宜をはかるために採録者が論文の原著にあたって、8 つの時代分類と 128 の分野に分けられ、タイトル中に現れないキーワードとして、作品・作者を加えて検索者の便宜を図っている点である。

作者・執筆者についてはそれぞれ読み方が付され、作品については正式名称に改められて採録され、しかも古典籍総合目録に採られている作品名については、その項目名に合わせて採録がなされている。

その意味で、表には見えない情報ながら、その情報の 2 次利用を図る際の辞書データとしてのポテンシャルには、大きな可能性がある。たとえば、古典籍総合目録には作品名に BIO（Biblio ID）が付されており、古典籍総合目録、さらには新日本古典籍総合目録データベースと連繋させることも可能である。著作者についても同様の措置が可能であるが、それについては今後の取り組みを待たねばならない。

また、冒頭でも述べた通り、明治～平成という長いスパンでデータが採録されいるので、学問分野の衰退・伸長を痛感することができる。たとえば、現在公開されるデータベースメニューには、各時

代分類と分野別に採録されているデータ総数が示されている。

それによると、以下の通り。

時代分類／分野

国文学一般　上代文学　中古文学　中世文学　近世文学　近代文学　国語　国語教育

国文学一般(54031)

一般(449)演劇・芸能(3079)沖縄文学(148)歌謡(867)芸能(461)古典文学(10477)詩歌(676)詩歌・歌謡(558)書評・紹介(3664)説話・昔話(4387)南島文学(2334)俳諧(1051)比較文学(2193)文学論(584)文学論・国文学論(8373)民俗学(7387)目録(345)目録・その他(4217)和歌(2781)

中古文学(62648)

一般(6808)歌謡(991)漢文学(2160)軍記(262)国語(2457)書評・紹介(2656)説話(3451)日記・随筆(7165)物語(24746)歴史物語(1859)和歌(10093)

近世文学(92849)

一般(23147)演劇・芸能(8371)演劇・芸能・芸能(3318)歌舞伎(136)漢文学・儒学(638)狂歌・狂文(1080)国学・和歌(8989)国語(2570)儒学・漢文学(5622)書評・紹介(3466)小説(15495)浄瑠璃(414)川柳・狂歌(347)川柳・雑俳(2178)俳諧(4363)連歌・俳諧(11967)和歌(130)和歌・和文(618)

国語(60199)

一般(6025)一般及び雑(300)音声・音韻(288)音声・音韻・アクセント(2450)敬語(955)言語生活(4049)語彙・意味(6710)辞書・資料(646)辞書・資料・訓点語(1552)書評・紹介(2326)対照研究(3685)日本語(371)日本語教育(6879)文字・表記(3290)文体・文章(2004)文法(12130)方言(6539)

上代文学(37856)

一般(6484)歌謡(1171)漢文学(419)古事記・日本書紀(4625)国語(1659)祝詞・宣命(325)書評・紹介(1831)神話(2309)風土記(1136)万葉集(17897)

中世文学(72302)

キリシタン文学・語学(1036)一般(8822)演劇・芸能(8654)演劇・芸能・芸能(1498)歌謡(905)漢文学(949)軍記物語(7758)国語(2490)書評・紹介(3066)唱導・縁起(310)小説(522)随筆(402)説話(1130)説話・唱導・縁起(4407)日記・紀行・随筆(3860)能(797)能・狂言(415)仏教文学(1107)仏教文学・神道(4861)物語・小説(3671)歴史物語・史論(534)連歌(2733)和歌(12375)

近代文学(182851)

一般(27705)演劇・芸能(4163)近代詩(5193)国語(588)作家別(23840)詩(2518)児童文学(4261)時評・展望(260)書評・紹介(10897)小説(24857)大衆文学(436)短歌(9398)著作家別(61230)俳句(4764)評論(2741)

国語教育(36564)

ことば(846)一般(16569)言語事項(1194)古典（古文・漢文）(319)国語教育（古典）(479)作文(926)作文・書写(298)書くこと(1407)書写・書道(365)書評・紹介(1622)読むこと(4798)読解・読書(2259)表現(1219)理解(3150)話すこと・聞くこと(1113)

上記の内、大分類にあたる「時代分類」は、国文学一般（54031）／上代文学（37856）／中古文学（62648）／中世文学（72302）／近世文学（92826）／近代文学（182851）／国語（60199）／国語教育（36564）

と、近代文学（含む現代文学）以外の古典文学の研究領域では、近世文学が最も数が多く、上代文学が最も少ないことがわかる。

さらにいうならば、こうした時間的推移に着目すると、国文学研究の質的変化を発表論文の件数という観点から俯瞰することができる。

この傾向は、経年変化という時間軸に基づいた分析を示すとさらに顕著に分かる。たとえば、図 2 は 1963 年（昭和 38）から 1995 年（平成 7）、における採録分野別の推移をグラフに示したものだが、先の新元号が『万葉集』巻 5「梅花之歌三十二首并序文」から採られたことは、そうした退潮傾向にあった上代文学の研究分野が蘇る契機を得たという点でも、画期的なことであったことがうかがえる。

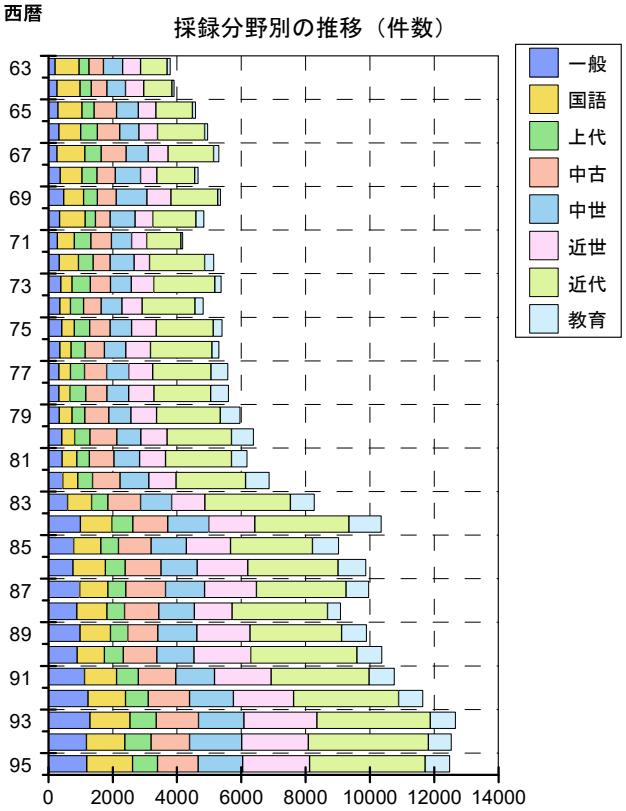


図 1　採録分野別の推移

(2)主題・観点から見た分野内の傾向と消長

このことは、別の視点からもうかがえる。具体例として、近年研究の関心が増えつつある地名・地理に関する研究成果の観点か示してみる。

この分野への国文学研究者側の関心は、総じて積極的とは言い難かった。このことを端的に示す事例としては、2010年10月2日開催の中古文学会秋季大会シンポジウム「平安文学と地理」の趣意文（加納重文氏）が象徴的だろう。

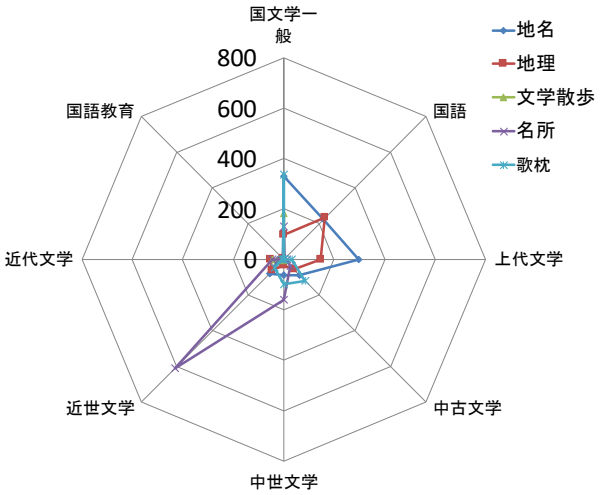
> 作家と作品の内面を規定するものとして、それが成立してきた環境は第一義的な意味を持つと思われるけれど、そのことを意識の中心にした研究に、なぜかはなはだ成果が乏しい。特に、地理的環境である平安京に関わる文学的考察が、不思議なほどに少ない。管見の範囲内でいえば、長谷章久（『古典文学の風土』）・角田文衛（『紫式部の周辺』『紫式部伝』）がその双璧であるけれど、孤高に聳えるのみで、これに続く山脈を望見しない。まことに寂寥の思いを抱く。（秋季大会シンポジウム要旨、『中古文学』87、2011年5月）

しかし、実態は逆で、表に示すように「歌枕」という和歌の題材とされた日本の名所旧跡のことをさしていう文学的に由緒深い事項については、網羅的に集積されたCiNiiよりも「国文学論文目録データベース」の方が多くの結果が集められていることが分かる。

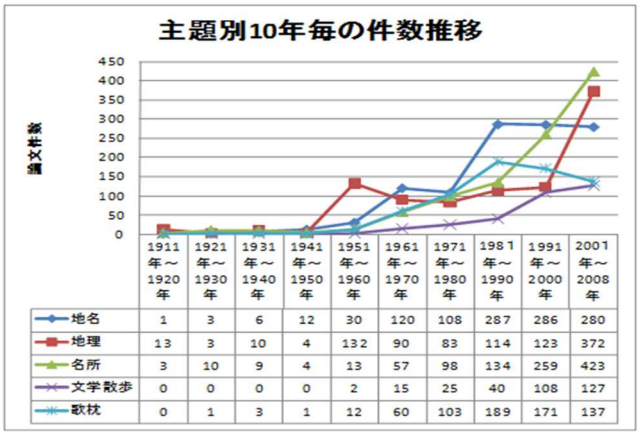
検索語	Cinii	国文研
地名	4268	1511
地理	117852	968
文学散歩	378	305
名所	2000	1759
歌枕	545	817
総計	125043	5360
2019/7/9調査		

さらに分析すると、この発言は日本文学の中の「中古文学」いう限られた分野のみに有効な発言であったことが、国文学の分野における分布を見ると判然とするのである。

なお、論文誌ジャーナルおよびJIPの査読のシングルブラインド制への移行に伴い[1]、2014年2月1日以降、投稿用原稿にも、著者名・所属・謝辞を表示することになっている。



さらに、これらの観点を時系列で分析したものが、次のグラフである。



年間1万件～1万2千件の間で推移する収録文献は、現在は館に寄贈される文献を中心に採録されるもので、他の専門特化型のデータベースと同様、採録者は一々の論文の現物にあたり、タイトルには現れない付加キーワードと、細分化された分類体系の類別作業を行ってきた。上記の分析は、そうした成果の上に成り立つものといえる。特に、大分類の時代分類は、他の網羅的データベースサービスではなし得ないものといえよう。

4. 文献目録作成という作業

先行研究の情報収集作業は、研究者にとっては基礎中の基礎ともいえるものである。かつては、若手研究者や大学院生を中心に文献目録の作成が数多く行われ、研究時評と併せて、少なからぬ研究雑誌において、その作成は修行同然のように組み込まれていた。そのようにして

国文学論文目録データベースは、そうした個別の取り組みを総括的に扱うものに位置づけられる。しかしながら、8種の時代分類に大別し、さらにそれを細かな分野に細分化する記述を養成するためには、少なからぬ時間を要する。すなわち、各時代毎に配された採録者がアルバイトとして

最低でも 2-3 ヶ月の養成期間を要し、スキルを上げることによって、資料整理補助員担当として採録・分類作業を重ねることによって構築される本データベースは、ある意味、知的労働者の人的集約作業によって成り立つといっても過言ではない。

しかし、実際は運営はデータベース科研という外部資金に頼らなくてはならず、内部資金のみで恒常的にこのデータベースを運営することには限界があることが指摘されてきていたことは否めず、かといってアクセス数だけが事業成果を保証するものではないということは、本事業の存続が問われる事自体が、証していると言わざるを得ないだろう。

一旦取りやめてしまうと事業の再開は人材的な面でもほぼ不可能と予想される。折しもシステム更新予定される時期にあたることから、今後の事業展開を考える上でも、事業の継続化のために抜本的な対策が求められることとなった。

5. 作業見直しの指針

まず予算不足を解消するために、現状の作業を抜本的に見直すことで、付加価値のあるデータ形成をめざすこととした。具体的には、既存のデータベース資源に付された分類やキーワードを学習データとして、以下のことに取り組む。当然のことながら、内製化による自助努力だけでは限度もあるため、外部資金を取り込みつつ根本的なエキスパートシステムを構築することとなる。そのための視点は以下の 4 点となろう。

①論文の内容を元にキーワード・カテゴリ特定機能自動化
②論文における引用元論文情報の付加価値
③より深い研究の提案
④機械学習用データ収集を通じて既存のデータ更新の人件費確保

そのために数点の実験を試みたが、現段階では細かな点で克服すべき点も見えてきており、その為に必要な積算を開示することによって、問題定義と指針を問う結果となってしまったというのが正直な所である。

①キーワード・カテゴリの自動化付与

本件については、1992 年（平成 4）～1996 年、1996 年から 2000 年（平成 12）の間、それぞれ汎用機（日立製作研 HITAC860/60、HITAC860/60K）により国文学論文目録データベースのオンライン公開がなされていた。当時は、搭載データベースシステムでは中間一致検索ができなかったため、検索効率を上げるために論文タイトルと副タイトルからそれぞれキーワードの切り出しを行い、そのキーワトーに付されたヨミと併せて検索の効率化を果たす HAPPINESS（（株）平和情報）データベースによる自動キーワード付加処理が行われた上でのデータベース公開がなされていた。今回の計画では、まず手付された上で蓄積・公開されてきた時代分類・キーワード（作品・作者）・執筆者・執筆者よみのデータを正当データとして、タイトルさらには論文本文から得られたキーワードをさらに自動分類に資することをめざした。

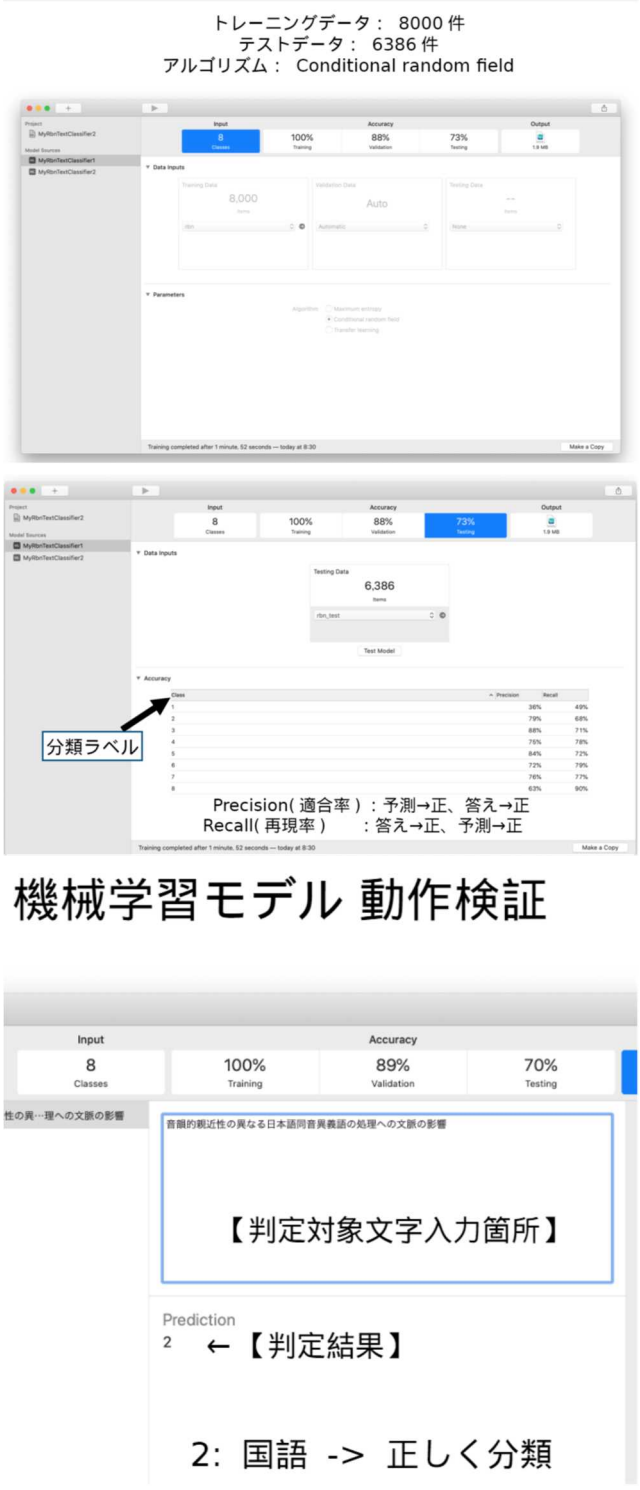
試験を行うに際して、
自然言語処理を利用した国文学関係論文の自動分類に関する資源状況レポートは以下の通りである。

まず、以下は MeCab と Python を利用しての分類結果、冒頭の数字が分野を示している。

総レコード数 596,403　（2019 年 1 月時点）

1,文学反響 Ｌｉｔｅｒａｒｙ　Ｅｃｈｏｉｎｇ　垣内松三
4,源氏物語のモデル 源氏物語 紫式部 手塚昇
5,室町時代の小歌と閑吟集 閑吟集 志田義秀
4,大鏡に関する考察 大鏡 藤村作
2,ノリといふ語 芳賀矢一
1,浦島伝説 久松潜一
8,古典講読用書様式の一提案 藤村作
8,聴方教授について 田中末広
1,国語研究室焼失主要書目録 目録 橋本進吉
1,生の象徴としての短詩 岩城準太郎
2,日本語教育　政策について 保科孝一 国語政策について
4,源氏物語のモデル 承前　源氏物語 紫式部 手塚昇
1,古典の本文整定 山岸徳平
6,契沖の文学批判 契沖 久松潜一
4,大鏡に関する考察 承前　大鏡 藤村作
5,室町時代の小歌と閑吟集 承前　閑吟集 志田義秀

さらに、「時代分類」の状況については、現時点では辞書が間に合わなかったため、MeCab と既存の公開辞書を使用した実験のほかに、トレーニングデータ：8000 件テストデータ：6386 件にて、アルゴリズム：Conditional random field を使用した分類付与実験と 2 通りで実現可能性の検証実験を行った。



動作検証を行ったところでは、
「音韻的真菌性の異なる日本語同音異義語の処理への文脈の影響」
の判定は
2:国語　→　正しく分類
となり、同様に、
「天石屋戸神話の成育過程」は、
3：上代文学　→　正しく分類
「小特集　今日､文学を諛むとは？　普遍性と観和性一古典文学を字ぷこと」
は、
1：国文学一般　→　正しく分類
という結果を得てはいるが、
学習データを変えると Precision(適合率)や Recll 率が変動することがわかり、
最後のものについては、
8:国語教育　→　間違った分類
を得ることにもなった。

まだまだ大括りの検証だが、それをさらに全文に及ぼす迄に至るには､現行の PDF 作成時の読み込みでは精度が悪すぎ、縦書き、2 段組みに対応した OCR の導入が求められることがわかつた。現時点では民生品 2 種類のものが実用に耐えることが分かってきており、今後さらに文字数を増やしての検証がどこまで可能か、駄目な場合は、どこを採ればいいのか検証を重ねることが課題となった。

また、自然言語処理を行う際にも、MeCab のネイティブ辞書を使用したところ、たとえば、「紫式部」が「紫,式部」と分かれて認識されてしまうために、国文学研究に特化した辞書を使用した検証を進める必要がある。60 万件に付されたデータを学習データとする場合と、　先に紹介した汎用機時代の HAPINESS によるキーワード自動切り出しの効率を上げるために別に形成していた語彙辞書の蓄積があるため、それらを使用した辞書の切り替えと併せた実験は今後も精度を高めていかなくてはならないだろう。

6. おわりに

最後に、今回はテスト用の論文データとして相田が事前に用意した 288 件の抜き刷りや論文データを基に作業を進めようとしたが、結果は、縦書き・組版の所で処理が止まってしまった。

こうした OCR 処理の問題もさておき､PDF 公開されているデータではあってもコピープロテクトがかかっているデータも少なくない。その意味で OCR を利用しての効率化と、認識結果の自動分類の実現のためには、どのような辞書を整えなくてはならないかなど、細かな点で解決しなくてはならない点は多いが、出来る所から確実に解決していくことが、重要だという認識は一致している。