様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 ２０ 日現在

機関番号：３２６３４
研究種目：挑戦的研究（萌芽）
研究期間：2020～2022
課題番号：２０Ｋ２０７０５

研究課題名（和文）AI-OCRを活用した英語初期印刷本の文字認識

研究課題名（英文）Character Recognition of English Early Printed Books with AI-OCR

研究代表者
松下　知紀（Matsushita, Tomonori）

専修大学・その他・名誉教授

研究者番号：５０１１５４２４

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：　英語の初期印刷本の文字認識は、将来中世英文学の研究に役割が期待される。本研究は、初期印刷本のうち、チョーサーの『カンタベリー物語：岸の物語』とウイリアム・ラングランドの『農夫ピアズ』を例にとり、研究を進めた。AI-文字認識プログラム開発技術が進み、成果が得られた。しかし、印刷の状態が悪く、カスレ文字、文字繋がりなどは誤認識となった。
AI-文字認識: トランスクリブス・プロジェクト，インスブルック大学、オーストリアに参加して、手稿写本の文字認識を行い、公開されているモデルに加えて、特定の写本のために創作されたモデルにより高い認識度を得ることができた。

研究成果の学術的意義や社会的意義
AI-OCRは現代社会においてDX，IoT分野と同様に社会的貢献を果たすことが可能である。従来、ほとんど研究対象にならなかった初期印刷本、手稿写本の特徴を捉え、比較検討することが可能になる。さらに、Digital Humanitiesという分野において、資料全体を文字イン指揮することにより、重要な研究資料を作成することになる。
さらに、それらを集積して、E-Librayを設立すれば、世界のどの場所にいても、世界で1つしかない資料を調査・研究が可能である。

研究成果の概要（英文）：Research on AI-OCR (Optical Character Recognition) will promote study of Medieval English literature.　The present study concerns Early English printed books such as Chaucer's The Canterbury Tales: The Knight's Tale and William Langland's Piers Plowman.　Progress in AI-OCR programs has gained a fruitful result.　However, there are still problems for recognition: slightly printed letters and ligatures (connected letters such as w = ea, w = he, etc.)
AI-OCR Project, University of Innsbruck, Austria has kindly given me a chance to investigate recognition of hand-written manuscripts of Canterbury Tales and Piers Plowman.　The Created Model for our MSS is quite suitable for recognition.

研究分野：人文情報学

キーワード：文字認識　初期印刷本　手稿写本　写本言語の比較

１．研究開始当初の背景

文字認識の研究は現代の英語などのアルファベットの活字を対象に行われてきた。現代のアルファベット活字は、画一的に作成され、変異の幅が極めて少ない。それに対して、グーテンベルクの初期印刷本（インキュナビュラ）等の場合、木版活字により印刷されている。木版活字は一定の変異差のある活字のため、認識率は低下する。20 年以上前から、アルファベットの文字認識に取り組んできたが、当時のプログラムでは十分な成果を得ることができなかった。

IT 技術の発展に伴い、現代のアルファベット印刷本の文字認識は次第に精度が向上した。しかし、現在でも、初期印刷本のアルファベットは、変異差が大きく、誤認識をする場合がある。本研究は、顔認証などの技術が発達し、画像解析が発達した現在、アルファベットの文字認識も精度を向上させられると感じた。また、AI には学習能力があるので、訓練を積むことにより、認識精度を向上察せられることが分かった。

日本では、郵便番号の読み取り作業が自動化し、効率を上げている。また、帳票を用いた手書き文字（漢字・平仮名）の文字認識も精度を上げている。日本語の場合、漢字の文字数が多く、誤認識の確率が高いが、記入する枠を設定することにより、精度を上げている。しかし、手書きの続き文字などは今後多くの学習過程が必要だろう。

英語などのアルファベットの場合、文字数が大文字・小文字と数字を合わせても、60‐70 文字に限られるので、研究成果が得られ易い,と考えた。

２．研究の目的

William Langland **の** *Piers Plowman: A-, B-, C-, Z- Texts* **の初期印刷本と手稿写本のデジタルテキストのデータベース化は、将来本格的に初期印刷本と手稿写本の文字認識研究が行われるために、必須条件である。そのため、以前作成した、デジタルテキストの再確認作業を行い、精度の高い研究の土台を形成する。**

A. 初期印刷本のための AI-OCR

AI-OCR（Optical Character Recognition）**を作成し、研究の土台を準備する。現在** AI-OCR **プログラムは精密になり、高額のものになっているため、凸版印刷の** AI-OCR **を本研究の資料に適合するように、文字資料を切り出して研究を行う。さらに、プログラムの作動を行うためのマニュアルも準備する。**

B. 手稿写本のためのＡＩ－ＯＣＲ

ＡＩの分野では、画像認識研究が進展しており、顔認証が個人別に行われて、別の人物と識別が可能である。そのような背景を受けて、Innsbruck University Transkribus Project, Austria **では、精密な文字認識研究を大規模に展開しているので、この活動に参加して、**

成果を挙げる。Transkribus Project **は、単に文字認識研究を目標とするだけではなく、対象とする作品の著者の活動に焦点を当てて、幅広い活動を行っている。**

C. E-Library の設立準備

Digital Humanities **の分野における** AI-OCR **研究は、文字認識の研究を成果とするだけではなく、将来の人文分野の研究に役立つ資料作成に貢献する必要がある。そのため、作成する研究資料は、断片的なものではなく、全体の把握が前提となる。**

３．研究の方法

A. デジタル画像の作成：William Caxton は翻訳家・印刷業者であり、多くの本を出版している。初期印刷本の文字認識研究資料作成のために、Chaucer の The *Canterbury Tales* のデジタル画像資料を作成した。
B. 凸版印刷株式会社に同社の AI-OCR を The *Canterbury Tales* と William Langland の *Piers Plowman* の初期印刷本のために適合化作業を依頼した。
C. 凸版印刷に The *Canterbury Tales* の文字切り出し作業（５万字分）を依頼した。

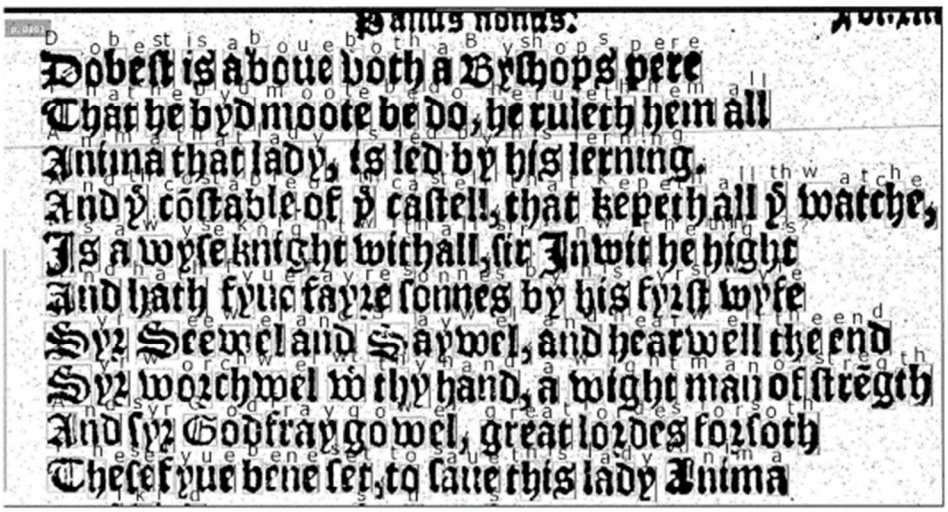
D. *The Canterbury Tales*：”The Knight’s Tale“と *Piers Plowman* の転写テキストを資料と照合しながら作成した。AI-OCR の出力結果と照合するため作業。

E. AI-OCR の出力を転写テキストと照合し、異同の問題点を整理した。

F. AI-OCR: Transkribus Project, Innsbruck, Austria に参加して、手稿写本の文字認識を行った。先ず、Public Model により出力し、転写テキストと比較を行い、異同を調査した。

G. AI-OCR: Transkribus Project の操作手順を点検するために、マニュアル作成をケイオス社に依頼して作成した。

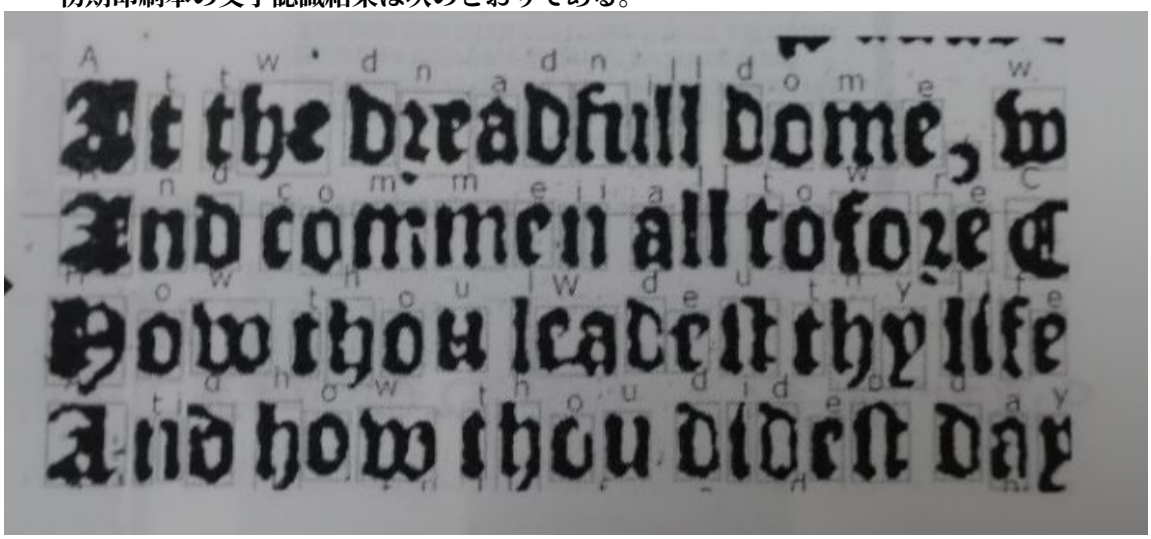
Piers Plowman: The B-Text, 1561 Crowley (Cr2)



H. AI-OCR: Transkribus Project の Created Model を作成し、手稿写本に適用した。その結果、Public Model より高い精度の成果が得られた。

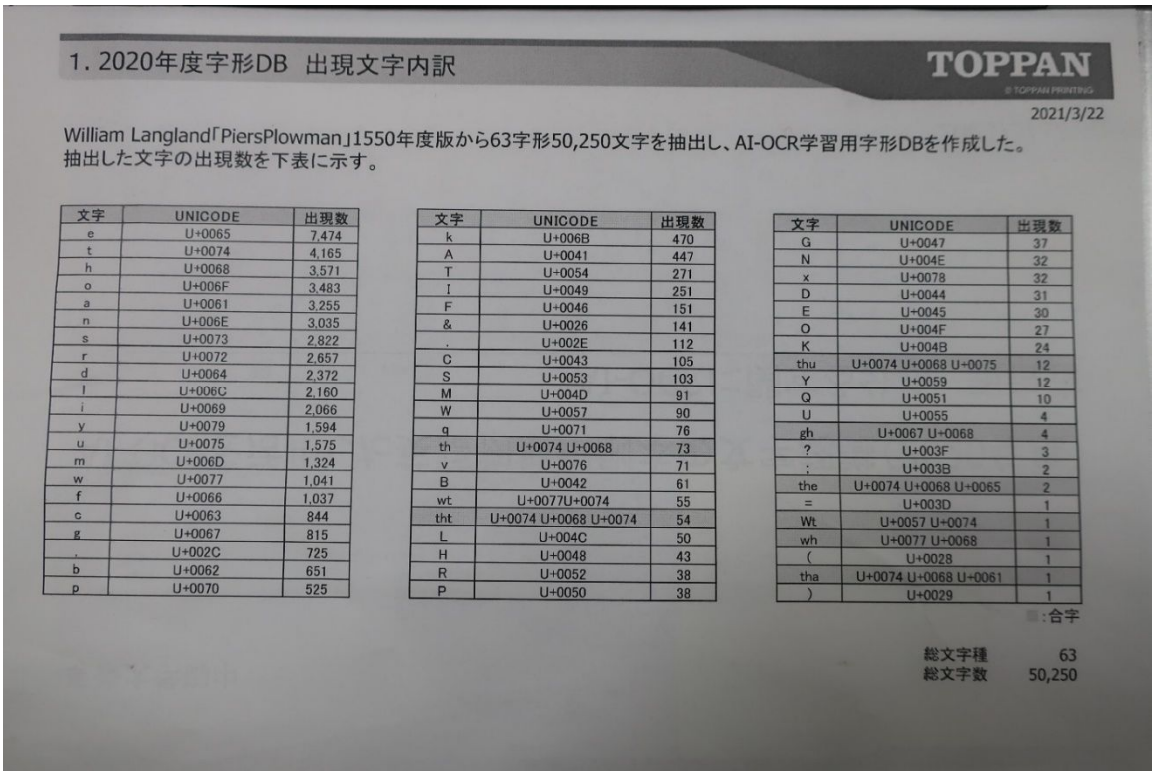
４．研究成果

初期印刷本の文字認識結果は次のとおりである。



At tw dnadnll dome w / And commueii all towre C / How thou lwdeu thy life / Atd how thou dideo day と認識された。 The > tw, dreadfull > dnadnll, commen > commeii, tofore > towre, leadest > lwdeu, didest > dideo と認識されている。これらの誤認識文字は印刷状態が悪く、独立文字とみなされず、接合しているとプログラムが判断している。今後学習により、

正しい判定力を装備する必要がある。初期印刷本の切り出し文字の出現数は、e，t，h，o，a，n が 3000 を超えるのに対して、大文字の O，K，Y，U の出現数は極めて少ない。



1. 2020年度字形DB 出現文字内訳

TOPPAN

2021/3/22

William Langland「PiersPlowman」1550年度版から63字形50,250文字を抽出し、AI-OCR学習用字形DBを作成した。
抽出した文字の出現数を下表に示す。

文字	UNICODE	出現数
e	U+0065	7,474
t	U+0074	4,165
h	U+0068	3,571
o	U+006F	3,483
a	U+0061	3,255
n	U+006E	3,035
s	U+0073	2,822
r	U+0072	2,657
d	U+0064	2,372
l	U+006C	2,160
i	U+0069	2,066
y	U+0079	1,594
u	U+0075	1,575
m	U+006D	1,324
w	U+0077	1,041
f	U+0066	1,037
c	U+0063	844
g	U+0067	815
,	U+002C	725
b	U+0062	651
p	U+0070	525

文字	UNICODE	出現数
k	U+006B	470
A	U+0041	447
T	U+0054	271
I	U+0049	251
F	U+0046	151
&	U+0026	141
.	U+002E	112
C	U+0043	105
S	U+0053	103
M	U+004D	91
W	U+0057	90
q	U+0071	76
th	U+0074 U+0068	73
v	U+0076	71
B	U+0042	61
wt	U+0077U+0074	55
tht	U+0074 U+0068 U+0074	54
L	U+004C	50
H	U+0048	43
R	U+0052	38
P	U+0050	38

文字	UNICODE	出現数
G	U+0047	37
N	U+004E	32
x	U+0078	32
D	U+0044	31
E	U+0045	30
O	U+004F	27
K	U+004B	24
thu	U+0074 U+0068 U+0075	12
Y	U+0059	12
Q	U+0051	10
U	U+0055	4
gh	U+0067 U+0068	4
?	U+003F	3
;	U+003B	2
the	U+0074 U+0068 U+0065	2
=	U+003D	1
Wt	U+0057 U+0074	1
wh	U+0077 U+0068	1
(	U+0028	1
tha	U+0074 U+0068 U+0061	1
)	U+0029	1

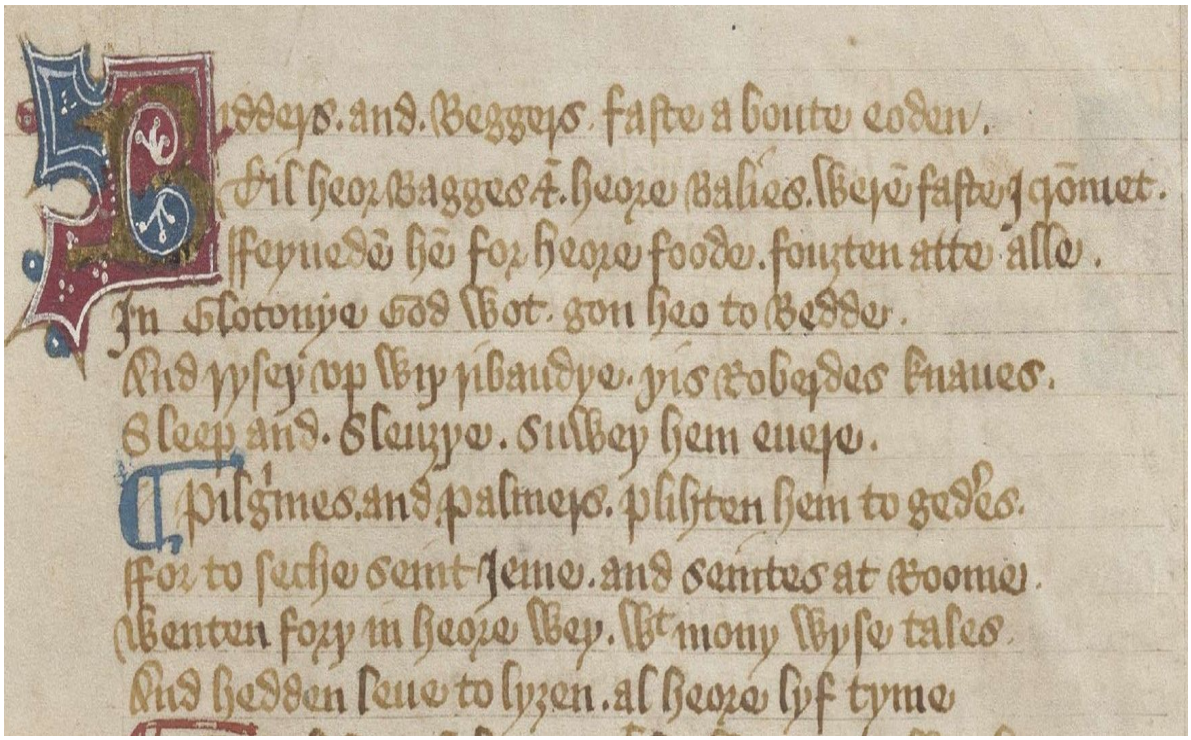
:合字

総文字種 63
総文字数 50,250

B．手稿写本の文字認識

AI-OCR: Transkribus

The Transkribus Project produced a program to transcribe hand-written MSS.



Piers Plowman: A, Prologue 40-49. Vernon MS. Eng Poetry a. 1, fol. 395ra. Bodl Lib, Oxford.

AI-OCR：Transkribus－ Public Model **の文字に意識結果は、**77.5%~90.3%**であり、平均**84.9%**だった。**

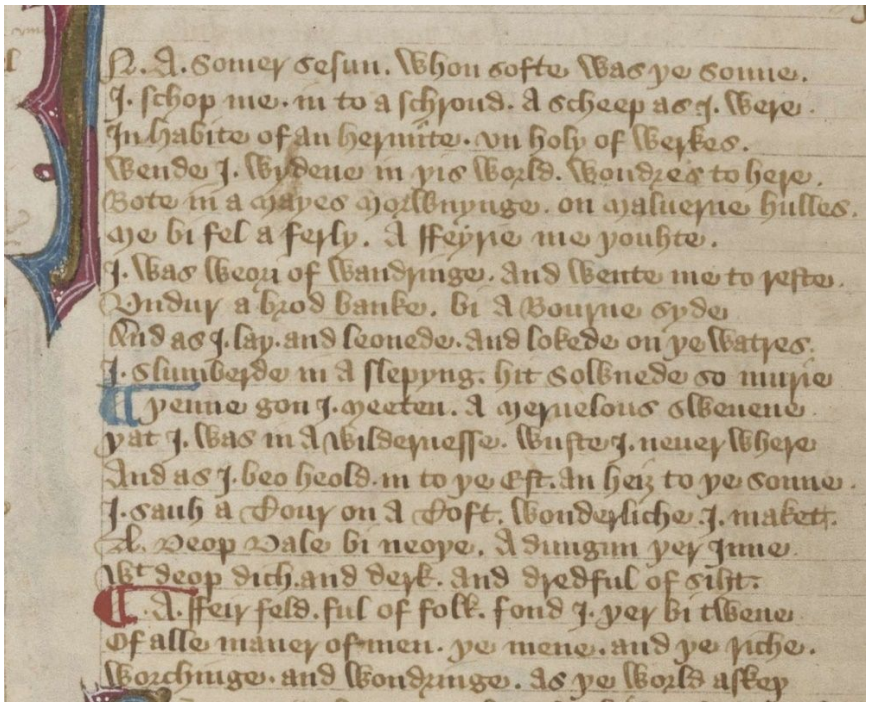
- P,40 **_ddeis. and. Begge_is. faste a boute eoden** (3/31, 28/31) 90.3%
  P,41 **Dil heorsagges et heore valies. Wese_ faste iomniet** (9/40, 31/40)
  77.5%... (omit) …
  Correct Recognition Rate (299/352, 84.9%) Misrecognition：P,40-P,49: 53

誤認識の例として、次のようなものがある。

P,40 i>r / P,41 T>D; B>s; &>et; B>v; w>W; r>s; Icr>i; m>ni /P,42 ff>F; e>o; □>z / P,43 I>i; G>g; /…(omit)…

２．TRANSKRIBUS: CREATED MODEL

Created Model による文字認識の場合、誤認識がほとんどなく、精度の高いプログラムであることが分かる。



Piers Plowman: A, Prol 1¯19, Vernon MS. Eng Poetry a. 1

</TextLine> <TextEquiv> <Unicode>

- In a somer sesun . whon softe was ŷe sonne,
- I schop me in-to a schroud . A scheep as I were;
- In Habite of an Hermite . vn-holy of werkes,
 （中略）
- Ŷat I was in A Wildernesse . wuste I neuer where,
- And as I beo-heold in-to ŷe Est . an-heiĝ to ŷe sonne,
- I sauh a Tour on A Toft . wonderliche I-maket;
- A Deop Dale bi-neoŷe . A dungun ŷer-Inne,
- With deop dich and derk . and dredful of siht.
- A Feir feld ful of folk . fond I ŷer bi-twene,
- Of alle maner of men . ŷe mene and ŷe riche,

  Worchinge and wondringe as ŷe world askeŷ.

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 松下知紀	４．巻 110
２．論文標題 ラングランド『農夫ピアズ：A, B, C& Z Versions』パラレル・テキスト・第二歌	５．発行年 2022年
３．雑誌名 専修人文論集	６．最初と最後の頁 47，101
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 松下知紀	４．巻 50
２．論文標題 『農夫ピアズ A, B, C, Z Versions』パラレル・テキスト・プロローグ	５．発行年 2020年
３．雑誌名 『人文科学年報』	６．最初と最後の頁 213，241
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Tomonori Matsushita	４．巻 106
２．論文標題 William Langland's Piers Plowman C-Version, V manuscript -- Trinity College Dublin MS 212 -- I	５．発行年 2020年
３．雑誌名 『専修人文論集』	６．最初と最後の頁 147，171
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 Tomonori Matsushita
２．発表標題 An IT Approach to Producing Diplomatic Texts of Piers Plowman and The Canterbury Tales Manuscripts
３．学会等名 New Chaucer Society（国際学会）
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関