



京都大学 学術情報メディアセンター
早稲田大学 未来イノベーション研究所
産業技術総合研究所人工知能研究センター
美馬 秀樹

デジタル化の課題とデジタル化資料の活用について —情報学の観点から—

内容

**I : 岩波書店「思想」のデジタル化
プロジェクト**

II : デジタル化の課題

III : デジタル化資料の活用

岩波「思想」90年の構造化



× 構想

- + 日本を代表する思想・哲学ジャーナル『思想』（1921年創刊）の90年分、約16万ページ、論文本数約8600本について、デジタル化とその「知の構造化」を行う

× 目的

- + 『思想』という知の集積と20世紀日本の哲学・思想史を明らかにする
- + 文献のデジタル化に関する方法論を確立する
 - × OCR、文書構造認識、検索、知の構造化
- + 分析結果の教育での活用

岩波「思想」の構造化と問題点

× 対象

+ 1921年創刊 岩波「思想」90年分

× 1000号、約8600論文、16万ページ

× 86年まで活版印刷=14万ページ分はデジタルデータなし

× デジタル化の課題

+ 大量

+ 原本が古く、取り扱いに注意が必要

+ 1921～50年頃は旧字体が使われている

テキストデータの取得

- ✕ 印刷された文字をコンピュータで処理可能なデジタルテキストに変換
 - + 画像化: スキャン
 - + テキスト化: OCR(Optical Character Recognition)

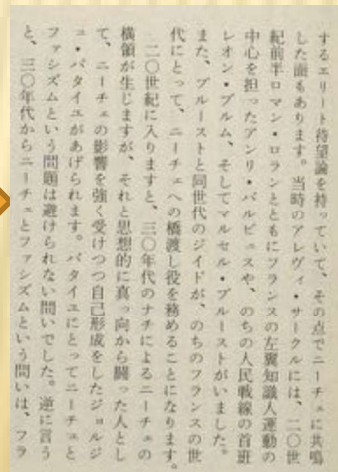
書籍・出版物



画像化



画像データ



テキスト化



テキストデータ

するエリート待望論を持っていて、その点でニーチェに共鳴した面もあります。当時のアレヴィ・サークルには、二〇世紀前半ロマン・ロランとともにフランスの左翼知識人運動の中心を担ったアンリ・バルビュスや、のちの人民戦線の首班レオン・ブルム、そしてマルセル・プルーストがいました。

デジタル化手法の検討

(4段階:◎, ○, △, ×)

	精度	コスト	時間	検討事項
人手入力	◎	×	△	・持出禁止あるいは取り扱いの難しい書籍では、コピーを取る必要がある
ブックスキャナ + OCR	○	○	◎	・誤認識をどう修正するか ・旧字体の認識精度
音声認識	△	△	×	・誤認識をどう修正するか ・旧字体の読み上げ

- ・コストと時間、精度のバランス
 - ・技術発展によりさらに精度向上、効率化が期待できる
- **ブックスキャナ + OCR(Optical Character Reader)** を選択

デジタル画像化

× 対象物のスキャンによりデジタルデータ化

+ スキャナの種類(スキャン形態による分類)

× フラットベッドスキャナ

- * 本をガラスの台に固定し、
下から光を当てて読み取る



× シートフィードスキャナ

- * 自動原稿送り装置で
原稿を送り、読み取る



× カメラスキャナ(スタンドスキャナ)

- * 原稿を専用の台に置き、カメラで撮影する
- * 本の非破壊スキャンが可能

カメラスキャナ

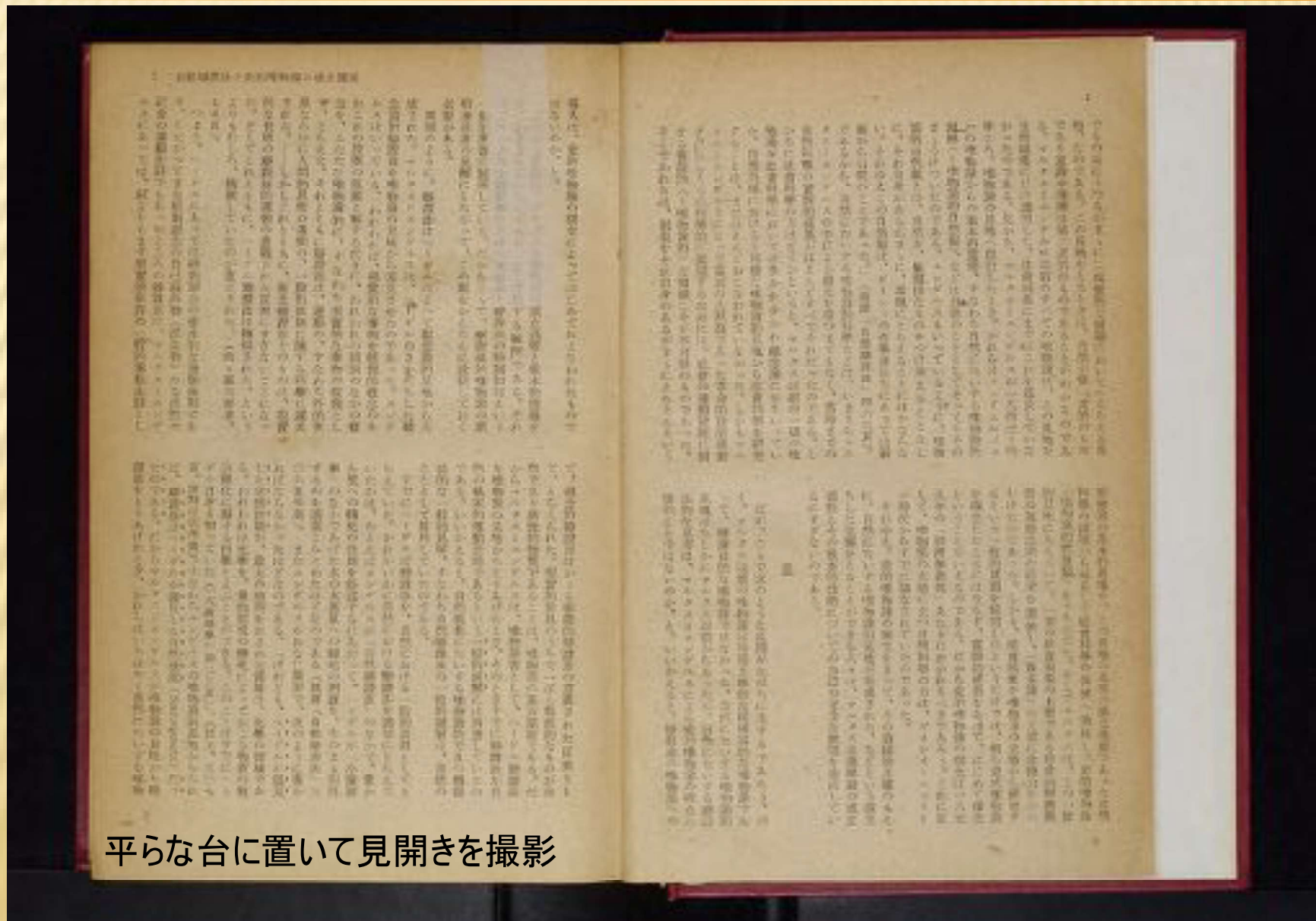


自動的に本をめくり
上にある2台のカメラで撮影

本を痛めることなく
自動でスキャンが可能

キルタス社製BookScanner (APT BookScan
2400RA)

カメラスキャナによるスキャン例

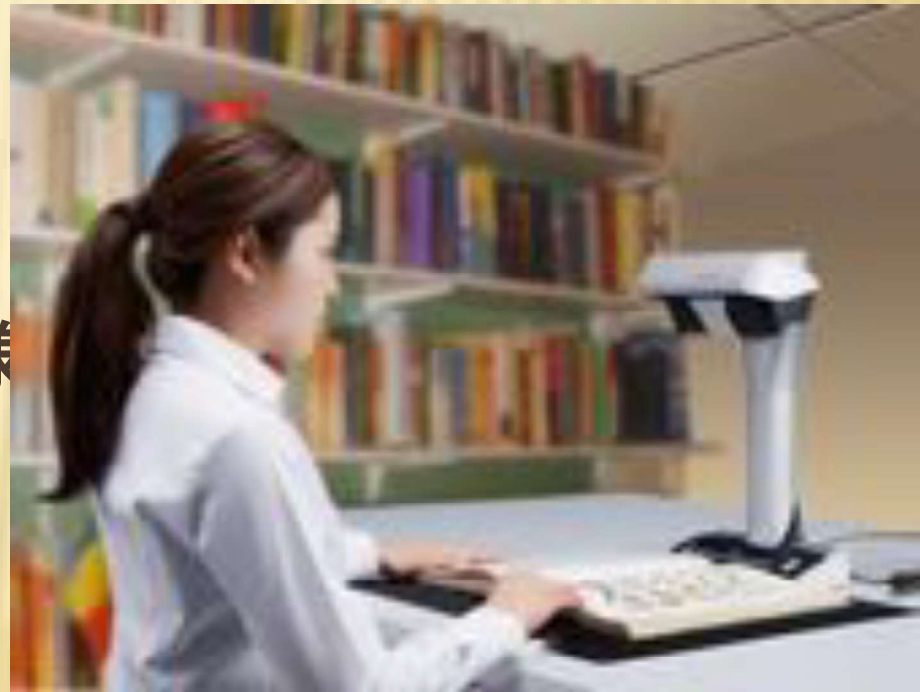


平らな台に置いて見開きを撮影

カメラスキャナ？

✕ Scansnap SV600

- + 実はカメラではなくラインセンサーでスキャン
- + ラインセンサーとライトが首を振る
- + 機能としてはカメラカメラスキャナと同様



BOOK TURNER

CASIO

大切な本やノートを電子書籍化

電子書籍化支援システム

BOOK TURNER

ブックターナー



[裁断不要] [ページめくりをアシスト]

BT-100

購入前のご注意

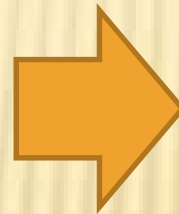
- ・本製品は、完全自動で書籍を電子化するブックスキャナーではありません。使用者の補助が必要ですので、最初にトレーニングしてから、大切な本などを撮影することを推奨します。
- ・本やノートの紙質、厚さ、製本状態によっては、本製品のページめくり動作がうまくできない場合があります。その場合は、半自動モードもしくは手動モードで撮影してください。
- ・本製品を使用する場合は、著作権法を遵守してください。
- ・本製品を使用するには、タブレットまたはスマートフォンと専用アプリが必要です。
- ・本製品には、タブレットやスマートフォンは含まれません。

デジタルテキスト化: CHARACTER RECOGNITION)

OCR(OPTICAL

× デジタル画像から文字を認識する

するエリート待望論を持っていて、その点でニーチェに共鳴した面もあります。当時のアレヴィ・サークルには、二〇世紀前半ロマン・ロランとともにフランスの左翼知識人運動の中心を担ったアンリ・バルビュスや、のちの人民戦線の首班レオン・ブルム、そしてマルセル・ブルーストがいました。また、ブルーストと同世代のジイドが、のちのフランスの世代にとって、ニーチェへの橋渡し役を務めることになりました。二〇世紀に入りますと、三〇年代のナチによるニーチェの横領が生じますが、それと思想的に真っ向から闘った人として、ニーチェの影響を強く受けつつ自己形成をしたジョルジュ・バタイユがあげられます。バタイユにとってニーチェとファシズムという問題は避けられない問いでした。逆に言う



するエリート待望論を持っていて、その点でニーチェに共鳴した面もあります。当時のアレヴィ・サークルには、二〇世紀前半ロマン・ロランとともにフランスの左翼知識人運動の中心を担ったアンリ・バルビュスや、のちの人民戦線の首班レオン・ブルム、そしてマルセル・ブルーストがいました。また、ブルーストと同世代のジイドが、のちのフランスの世代にとって、ニーチェへの橋渡し役を務めることになりました。二〇世紀に入りますと、三〇年代のナチによるニーチェの横領が生じますが、それと思想的に真っ向から闘った人として、ニーチェの影響を強く受けつつ自己形成をしたジョルジュ・バタイユがあげられます。バタイユにとってニーチェとファシズムという問題は避けられない問いでした。逆に言う

OCRシステム

× 市販ソフト

- + メディアドライブ WinReaderPRO、 e.Typist
- + ABBYY FineReader
- + パナソニック 読取革命
- ...

× サービス

- + Google Cloud Vision API
- + Evernote API
- ...

OCR精度の向上による 言語処理結果の向上

× OCR精度

95% から 99.85% へ

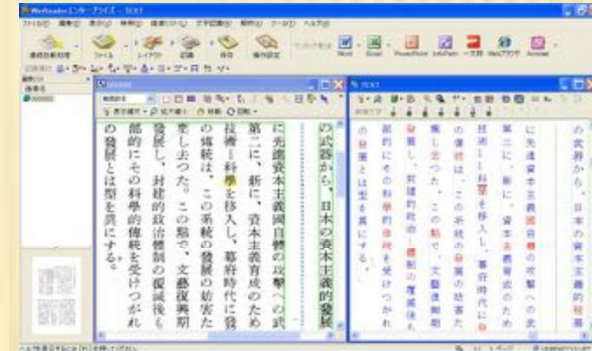
OCR精度95%での用語抽出結果

	1930年	1931年	1932年	1933年	1934年	1935年	1936年	1937年	1938年	1939年
1	自然科学	日本	日本	道徳	日本	自己自身	西田哲学	道徳	アメリカ	自己同一
2	資本主義	カント	ギルヘルム	フランス	宗教	自己限定	日本	自己自身	英国	日本
3	日本	アメリカ	ブルジョア	自然科学	フランス	道徳	自己自身	日本	日本	道徳
4	フランス	自己自身	カント	カント	資本主義	日本	自己限定	アントニ	自己同一	プラトン
5	道徳	ブルジョア	宗教	日本	ブルジョア	根柢	寺田先生	主体	フランス	宗教
6	ブルジョア	自己限定	自然科学	ブルジョア	昭和八	自然科学	フランス	職業生活	イギリス	ピュタゴラス
7	回顧	宗教	ファウスト	アジア	道徳	フランス	自己矛盾	フランス	ヨーロッパ	自己自身
8	エンゲルス	フランス	資本主義	スペンサー	昭和九	独逸	ノエシス	表現活動	自己自身	支那
9	唯物弁証法	エンゲルス	自己自身	ドイツ	ドイツ	宗教	道徳	ベルグソン	自由主義	絶対矛盾
10	宗教	メーリング	絶対精神	ゲシヒテ	自然科学	老子	カント	自己否定	自由主義	吉利

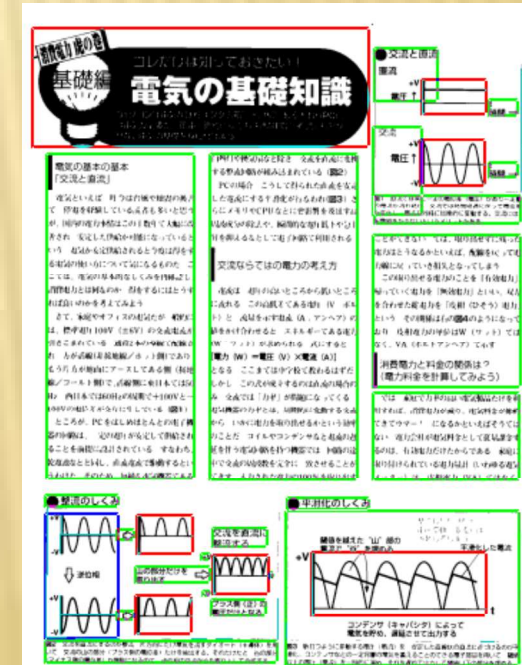
道徳、宗教、ブルジョア等の認識間違いが多数みられる

OCRによる電子化と課題

- ✖ 文字認識精度が低い
 - + 旧字体の認識精度の問題
 - + 強調部分、ルビ、外来語部分の認識
 - + レイアウト解析に失敗し文字化けする→近代文語論文の特徴抽出を行い精度を上げる



- ✖ 人手による作業コストの増大
 - + レイアウト解析が不十分
 - ✖ タイトル、著者、脚注部分等のメタ情報を識別しなければならない
 - ✖ 論文の区切りを識別（ページの途中で別論文が始まる等）
 - + 誤認識の人手による訂正と学習→自動化・システム化を行い、作業コストを軽減する



精度低下の原因

× 異体字・旧字体

例: 教⇔教 学⇔學

× フォントの違い

「と」:一画目が斜め、二画目の入りが大きい

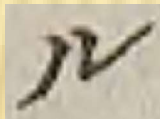


→「ご」や「ざ」と誤る



「感」:「心」の部分が小さい

→「戚」や「咸」と誤る



「ル」:右側が上がっている

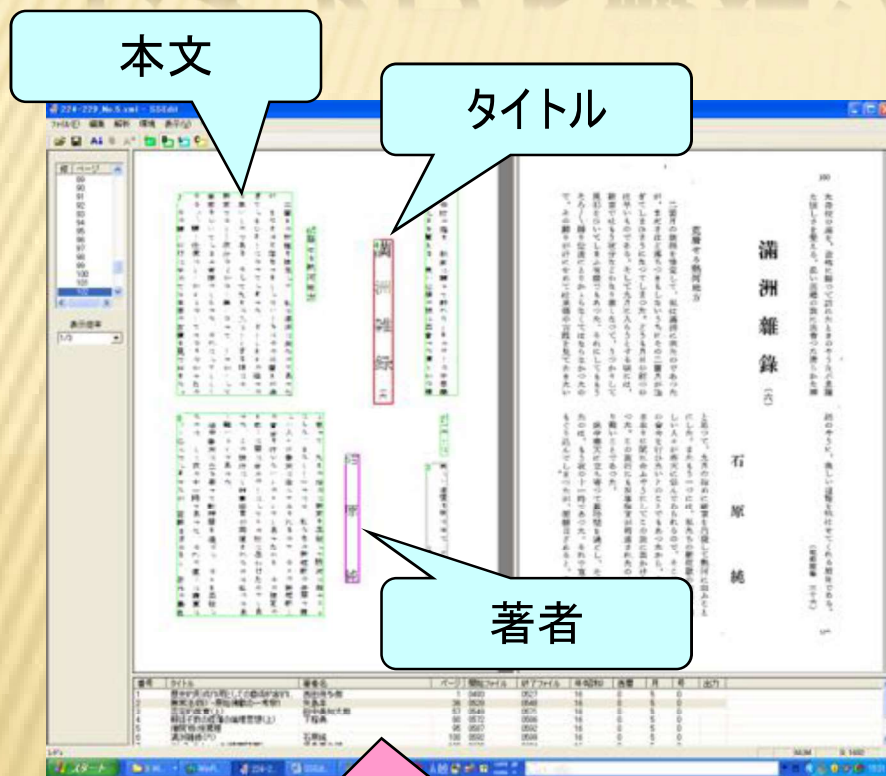
→「ル」や「ル」と誤る

精度99.85%での用語抽出結果

	1930年	1931年	1932年	1933年	1934年	1935年	1936年	1937年	1938年	1939年
1	自然 科学	ヘーゲル	ヘーゲル	自然 科学	ヘーゲル	自己 自身	ヒューマニズ ム	アントニ	アメリカ	自己 同一
2	マルクス 主 義	マルクス 主 義	ゲーテ	唯物 論	マルクス	自己 限定	西田 哲学	ヘーゲル	ヘーゲル	ピュタゴラス
3	唯物 論	弁証 法	マルクス 主 義	マルクス	フランス	ヘーゲル	ヘーゲル	自己 自身	自己 同一	プラトン
4	ヘーゲル	カント	ギルヘルム	カント	キリスト	アリストテレ ス	自己 自身	職業 生活	自由 主義	自己 自身
5	エンゲルス	自己 自身	形而 上学	スペンサー	アリストテレ ス	自然 科学	自己 限定	表現 活動	イギリス	ディオニューソ ス
6	プロレタリ アート	プロレタリ アート	弁証 法	存在 論	日本 精神	ロゴス	弁証 法	ベルグソン	自己 自身	キリスト
7	資本 主義	ジャーナリズ ム	エンゲルス	アリストテレ ス	自然 科学	プラトン	ノエシス	自己 否定	フランス	モラル
8	トーキー	形而 上学	ファウスト	イデオロ ギー	エンゲルス	エネルギー	プラトン	プラトン	ヨーロッパ	ロゴス
9	コント	マルクス	ディルタイ	ディルタイ	自己 自身	ノエシス	アリストテレ ス	弁証 法	ロゴス	エートス
10	観念 論	イデオロギー	カント	形而 上学	カント	ベルグソン	ソクラテス	デカルト	アントニ	ソクラテス

上位には間違いがほとんど見られない

これまでの取り組み レイアウト解析ソフトウェアの開発



論文の区切りを認識して
CSV形式にまとめる

- × タイトル・著者名の抽出
- × ルビ・強調文字と単語(文字)の対応付け
- × ページ番号、注の検出
- × 発行年・月・号の取得
- × 新旧字体変換の作成
- × CSVファイルへの出力
- × 誤認識文字の置換(候補から選択、手動入力)
- × レイアウトの修正(ブロックの切り出し、順番の変更)

機械学習を用いた レイアウト属性付け

レイアウト属性の正解データ

ページ番号

柱

本文

タイトル

著者

機械学習

特徴量による
分類器

特徴量

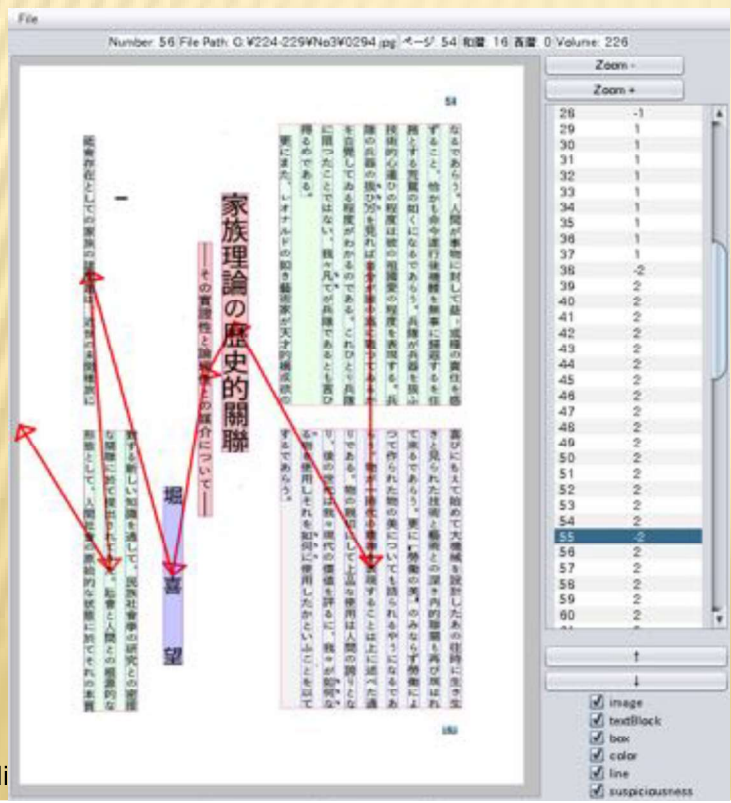
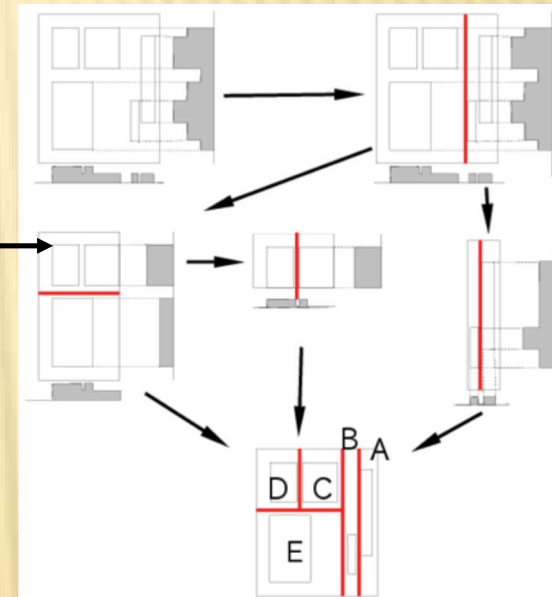
視覚的特徴量
文字サイズ
ブロックサイズ
ブロック位置
余白長さ
言語的特徴
「名詞」割合
「人名」割合

精度(F値)	ルール	機械学習
タイトル	0.913	0.960
著者	0.966	0.985
柱	0.974	0.979
ページ番号	0.994	0.995
本文	0.985	0.992

人手で作成したルールに比べ高精度

ブロック読み順の推定

OCR結果のブロック間の順番を自動で推定
自動読み上げ
テキスト検索の精度向上
機械学習による推定



認識誤差: 0.04

← 人手による修正インターフェース
最終的な人手チェックの労力を削減
半分程度の時間で修正可能

言語モデルを用いた OCR文字誤り訂正

OCR出力

誤り検出

- 言語モデル(文字Trigram)を用いてOCRの文字誤りを自動的に検出

訂正文字
候補生成

- 誤りとされた文字に対し訂正候補文字を文字Trigramを用いて生成
- OCRシステムが出力する候補文字と合わせて訂正文字候補とする

訂正文字
候補選択

- 生成された文字候補中から言語モデル(単語Trigram)および文字類似度を用いて最も確率の高い文字候補を選択する

訂正後文字列

OCR実験:「思想」年代別文字認識精度

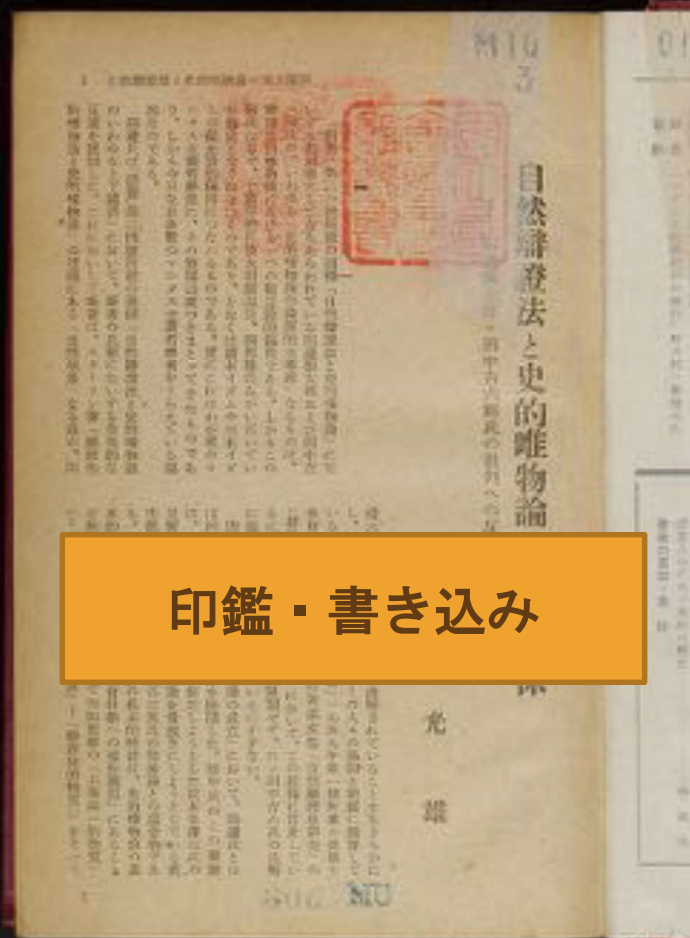
各年代1論文を手手でチェック

年	データ番号	文字数	誤認識数	認識精度	
2000	0010-1	15998	31	99.81%	
1990	0010-1	41638	215	99.48%	
1980	0010-1	32229	83	99.74%	
1970	0010-1	29476	846	97.13%	参考文献に横倒し文字多数
1960	0010-1	3090	61	98.02%	横倒し文字
1950	0010-1	6066	170	97.19%	原本の状態が悪い
1940	0010-1	8316	52	99.37%	
1930	0010-1	4294	348	91.90%	横倒し文字
1921	0010-1	9218	1078	88.31%	横倒し文字、原本の状態、フォントの違い

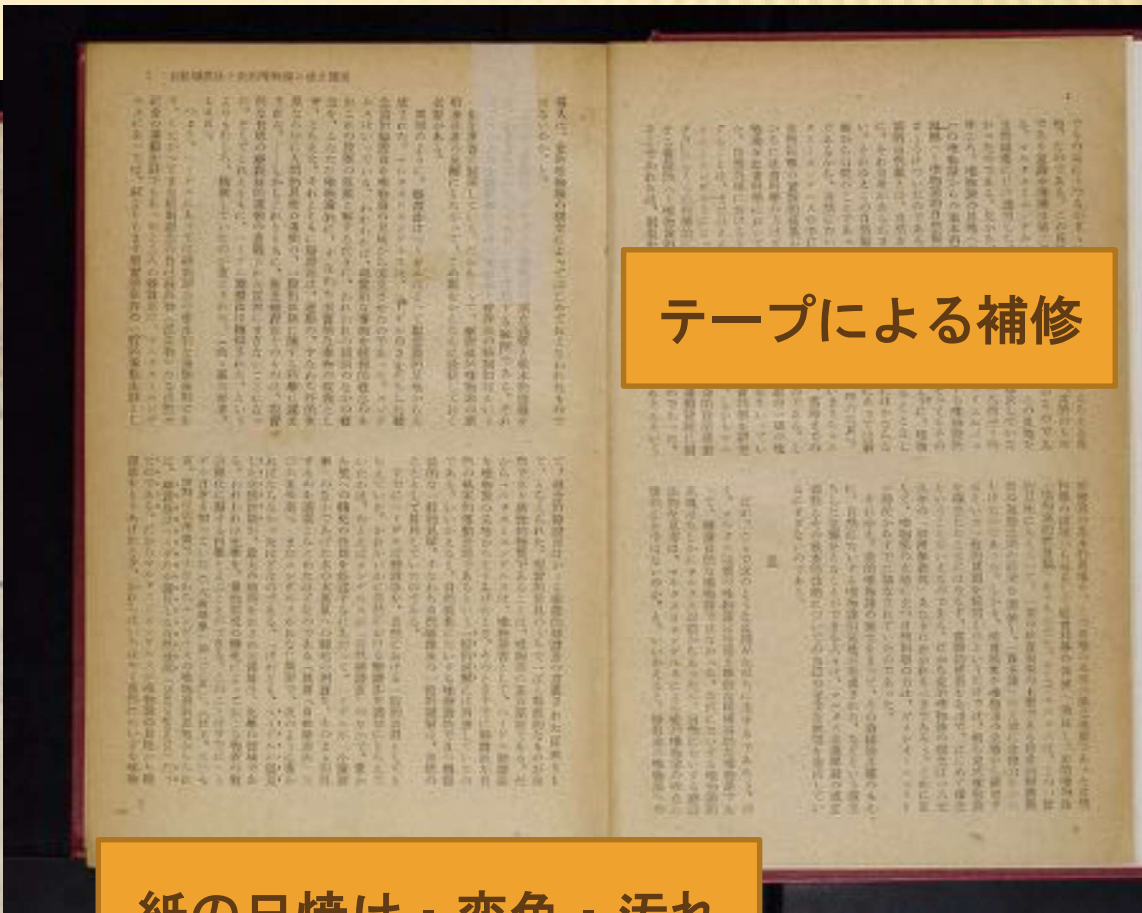
古い年代ほど原本の状態などにより精度が低下

その他、OCRの課題

原本の状態による認識の困難さ



印鑑・書き込み



紙の日焼け・変色・汚れ

テープによる補修



広告の一行が
本文に取り込まれ
る

要員の取組

マリー・キュリー

ドイツ(東部)のシュテルン紙によ
って、大衆マリーが躍如する
新資料を駆使して筆力のある作
が躍り上げた。東大文藝部学友の
新しい日記、田中夏子訳、5年1200

毛沢東伝

東洋及主編、19世紀中国を編纂
とした原基の製生から中華人民
共和国成立まで、米が原資料多数を
駆使して編「赤字版」新、村松
正樹、全1巻、村松・真樹訳、5年1200

翻梁啓超

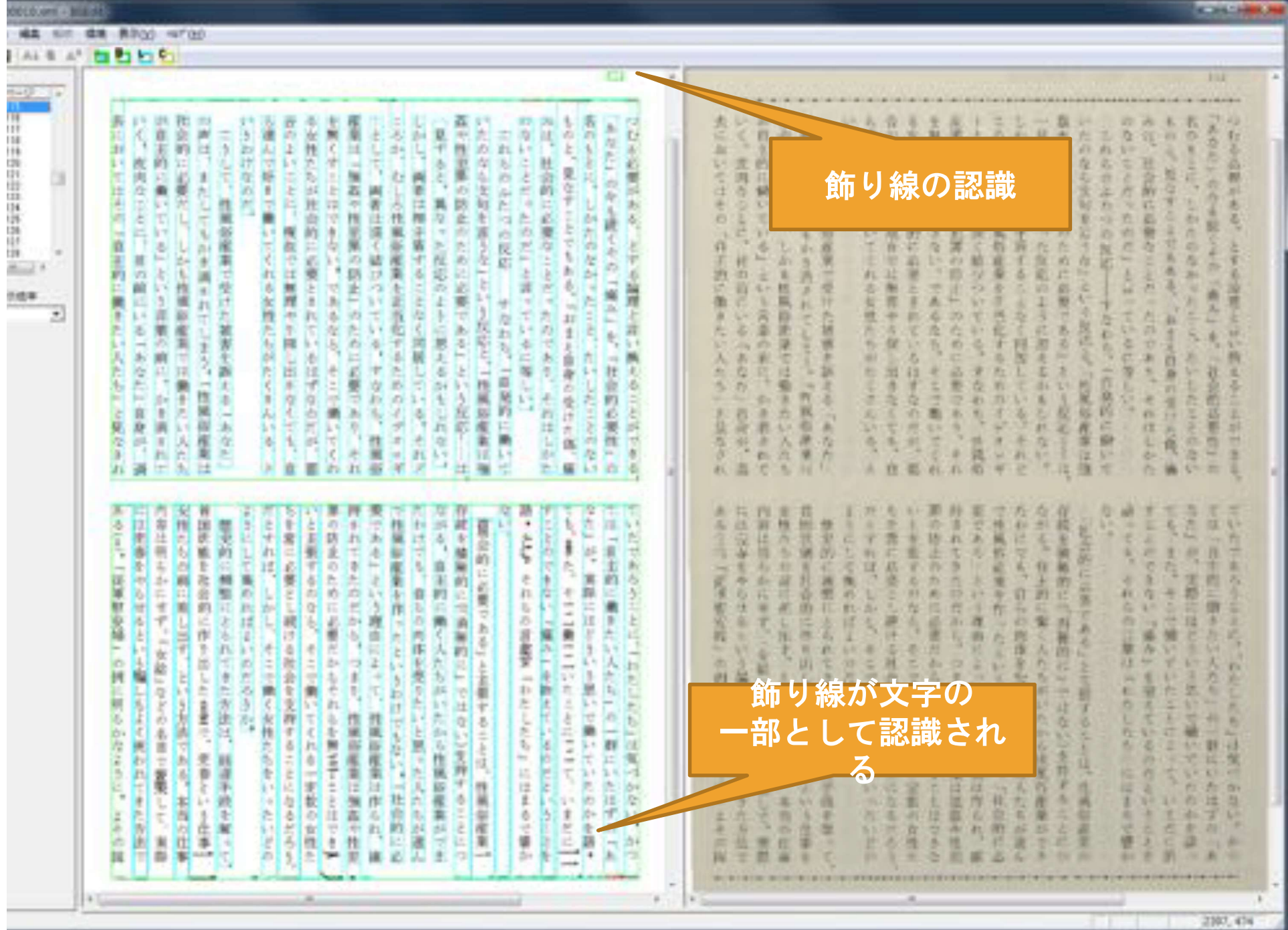
西学を伝授する支那の第一人
梁啓超編、東アジアの支那史的
転換期に於て行われた。近代中
国の思想史の史的探求の終を遂げ
し、その過程と神話を明かす、東大
文芸部、最新の花巻集、5年1200

セザンヌの面

19世紀後半の近代美術を以
てするの芸術性へ、セザンヌの芸術
の展開と進化を、各ジャンルに對

者、あるいは「強制不妊手術の被害者」という地点で終わ
てしまうことの問題性もわかりますし、また、私のような人
間が被害者の方々の痛みを「*ハートブレイク*」するな
らば、それはまさに「*ハートブレイク*」政治そのもの
なのだと言わねばならない。しかし、他方で、補償が無くていい
かと言え、決してそうではないし、また補償の実現に向け
て、被害当事者の方々を固りの人間がサポートしていくこと
も必要なんじゃないか。
先ほどふれた集会の翌々日に、「優生手術に対する謝罪を
求める会」の面々で、被害者の女性と一緒に、厚生省に出向
き、交渉をしました。私は同席できなかったんですが、厚生
省は「当時としては合法だったんだから、問題はない」の一
点張りだったそうです。この交渉には、サチスの優生政策の
被害者に対する戦後補償の実現に取り組み、「〇月の集会で
も講演をしてくれた、クリスティーネ・テラーさんというド
イツの精神科医(C・テラー)／野田正彰／小茂和一郎(医学・精

神医学と)
くれましたし
保護法の
いること
を正すの
諦であり
「たとえ
たんです
本の事情
傷つけ
の被害者
そのなん
んどの仕
じ、ない
くもかゆ
栗原



飾り線の認識

飾り線が文字の一部として認識される

はむる必要がある。とすも論理上許し難い。換言すれば、
「あなた」の命を殺すの「一歩」も、「社会的必要性」の
名のもとに、しかたのないこと、たいていこのことのない
ものと、異なることでもあり、おまゝ自身の受けかた、種
別は、社会的に必要なことと、たつたのである。それはしかた
のないことと、たつたこととを言っているに等しい。
これらもまたこの反駁。すなわち、一層具体的に論じて
いたのなる支那を導くという反駁と、「性風俗産業は種
別や性風俗の防止のために必要である」という反駁——は
「見せよ」と、真実を反駁の上で思えるかも知れない。
しかし、両者は押す事なくなく論議している。それはア
ニムが、むしろ性風俗産業を正気化するためのイデオロギ
ーとして、両者は深く結びついている。すなわち、性風俗
産業は「種別や性風俗の防止」のために必要であり、それ
を無くすることはできない。であるなら、そこを論じてく
る女性たちが社会的に必要とされているはずなのである。要
害のよいか、種別では無理やき理し出すまでもなく、自
ら選んで種別で働いてくれる女性たちがたくさんいる。そ
れらの方針である。

二つとして、性風俗産業を受けた被害者を救済する（あなた）
の責は、またしてもかき置かれてしまふ。「性風俗産業は
社会的に必要だし、しかも性風俗産業で稼働したい人たちが
自主的に働いている」という言葉の前は、かき置かれてし
ゆく、皮肉なこと。その前に「あなた」自身は「種
別においてはその自主的に働きたい人たちが」と見なされ
ていたのである。この「あなた」は「種別」か、か
つては「自主的に働きたい人たちが」の一群にいたはず
だが、言葉にはどうも思っていないかを疑
う。また、その「種別」は「あなた」が「種別」
のことではない。「種別」を指しているのだから、
「あなた」も、それらの言葉「あなた」にはまるま
り。
「社会的に必要である」と主張すること、性風俗産業は「
種別を種別的に必要とする」ではないと支持することにつ
いて、自主的に働く人たちがいたから性風俗産業が主
たわけでも、自らの責任を受けたかと思ふ。たまたまが選ん
で性風俗産業を作ったというわけでもない。「社会的に必
要である」という理由によつて、性風俗産業は作られ、種
別されてきたのだから、つまり、性風俗産業は種別や性風
俗の防止のために必要だからとされるを論じて、これはでき
ないと思ふのである。そこで論じてくれる「あなた」の女性た
ちを救う必要として種別を社会を支持することになるだろう。
可なりれば、しかし、そこで働く女性たちをいかに救うか
、さして種別を救うか、いかに救うか。
種別の種類にあられたる古法は、種別手段を減らして、
種別手段を社会的に作り出した。受容という仕事——
女性たちの種別は種別し出す、という方法である。本職の仕事
内容は種別をしない。女性などの名目で受容して、種別
には種別をやらせるとも種別しよと種別されてきた古法は
ある。種別手段の「種別」の種別に種別をいかに救うか、

のむる必要がある。とすも論理上許し難い。換言すれば、
「あなた」の命を殺すの「一歩」も、「社会的必要性」の
名のもとに、しかたのないこと、たいていこのことのない
ものと、異なることでもあり、おまゝ自身の受けかた、種
別は、社会的に必要なことと、たつたのである。それはしかた
のないことと、たつたこととを言っているに等しい。
これらもまたこの反駁。すなわち、一層具体的に論じて
いたのなる支那を導くという反駁と、「性風俗産業は種
別や性風俗の防止のために必要である」という反駁——は
「見せよ」と、真実を反駁の上で思えるかも知れない。
しかし、両者は押す事なくなく論議している。それはア
ニムが、むしろ性風俗産業を正気化するためのイデオロギ
ーとして、両者は深く結びついている。すなわち、性風俗
産業は「種別や性風俗の防止」のために必要であり、それ
を無くすることはできない。であるなら、そこを論じてく
る女性たちが社会的に必要とされているはずなのである。要
害のよいか、種別では無理やき理し出すまでもなく、自
ら選んで種別で働いてくれる女性たちがたくさんいる。そ
れらの方針である。

二つとして、性風俗産業を受けた被害者を救済する（あなた）
の責は、またしてもかき置かれてしまふ。「性風俗産業は
社会的に必要だし、しかも性風俗産業で稼働したい人たちが
自主的に働いている」という言葉の前は、かき置かれてし
ゆく、皮肉なこと。その前に「あなた」自身は「種
別においてはその自主的に働きたい人たちが」と見なされ
ていたのである。この「あなた」は「種別」か、か
つては「自主的に働きたい人たちが」の一群にいたはず
だが、言葉にはどうも思っていないかを疑
う。また、その「種別」は「あなた」が「種別」
のことではない。「種別」を指しているのだから、
「あなた」も、それらの言葉「あなた」にはまるま
り。
「社会的に必要である」と主張すること、性風俗産業は「
種別を種別的に必要とする」ではないと支持することにつ
いて、自主的に働く人たちがいたから性風俗産業が主
たわけでも、自らの責任を受けたかと思ふ。たまたまが選ん
で性風俗産業を作ったというわけでもない。「社会的に必
要である」という理由によつて、性風俗産業は作られ、種
別されてきたのだから、つまり、性風俗産業は種別や性風
俗の防止のために必要だからとされるを論じて、これはでき
ないと思ふのである。そこで論じてくれる「あなた」の女性た
ちを救う必要として種別を社会を支持することになるだろう。
可なりれば、しかし、そこで働く女性たちをいかに救うか
、さして種別を救うか、いかに救うか。
種別の種類にあられたる古法は、種別手段を減らして、
種別手段を社会的に作り出した。受容という仕事——
女性たちの種別は種別し出す、という方法である。本職の仕事
内容は種別をしない。女性などの名目で受容して、種別
には種別をやらせるとも種別しよと種別されてきた古法は
ある。種別手段の「種別」の種別に種別をいかに救うか、

「思想」 文字認識誤りの訂正

× 実験評価

+ データ: 1940年1月号最初の論文

- × ページ数 11 (うち1ページは行認識誤りのため不使用)

- × 文字数 7368

- × 誤り文字数 50 (0.67%、新字旧字の違いは誤りとししない)

+ 手法

- × 誤り検出法: WRP出力の確信度 or Ngram

- × 訂正文字候補: WRP出力の候補のみ

or WRP候補 + Ngram候補

「思想」 文字誤り訂正

× 文字誤り検出結果 (誤り文字数50)

	検出数	正解数	誤検出	Precision	Recall	F値
確信度(閾値80)	212	20	192	0.094	0.400	0.153
確信度(閾値90)	1421	28	1393	0.019	0.560	0.038
Ngram	357	42	315	0.118	0.840	0.206

× OCR確信度と文字誤りの関係

確信度	100	90-99	80-89	70-79	60-69	50-59	1-49	0
総文字数	4780	1138	1209	161	26	11	14	28
誤り文字数	4	9	8	14	1	1	4	9
誤り率(%)	0.08	0.79	0.66	8.70	3.85	9.09	28.57	32.14

確信度0 ≡ 横倒し文字として認識

「思想」 文字誤り訂正

× 文字誤り訂正結果

誤り 検出法	文字候補 生成法	正解 訂 正数	誤 訂正数	文書精度	改善 率	改善 数
確信度 (閾値80)	OCR	15/ 20	11/ 192	99.32→99.38	+0.05	+4
	+Trigram	13/ 20	20/ 192	99.32→99.23	-0.10	-7
Trigram	OCR	23/ 42	20/ 315	99.32→99.36	+0.04	+3
	+Trigram	22/ 42	73/ 315	99.32→98.63	-0.69	-51

デジタル化に向けての課題

× デジタル画像化

+ いかにか質のよい大量のデジタル画像データを取得するか

× OCR

+ レイアウト解析

× 文字の位置やつながり写真等との切り分け

+ 文字認識

× 古いフォント

× るび

× 言語モデル

機械学習(ディープラーニング)の活用

最近の話題：DNNによるレイアウト認識

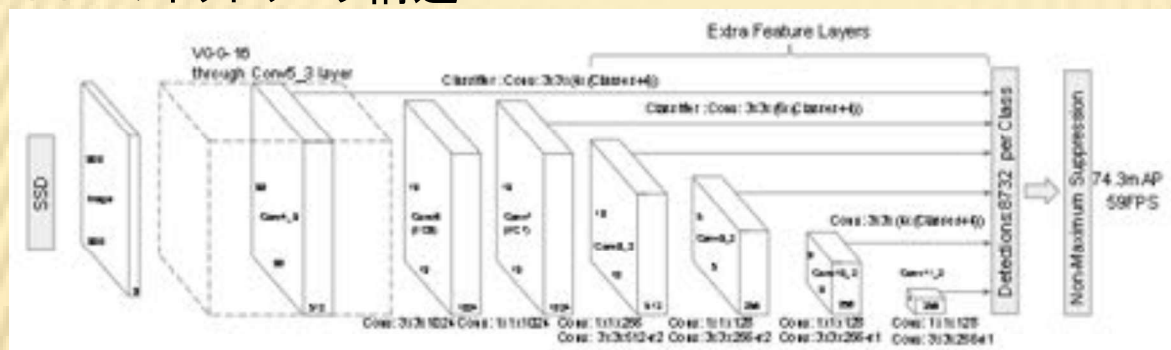
DNNによるレイアウト認識実験

- × 人文学オープンデータ 共同利用センター-N2I
プロジェクト + 東京大学 + 国立国会図書館 +
NIIの共同研究
 - + 「思想」「国民之友」「東洋学芸雑誌」のレイ
アウト認識正解データ約1000ページ
 - + SSD(Single Shot Multi Detector)
 - × kerasによる実装(https://github.com/rykov8/ssd_keras)
の書き換え

SSD

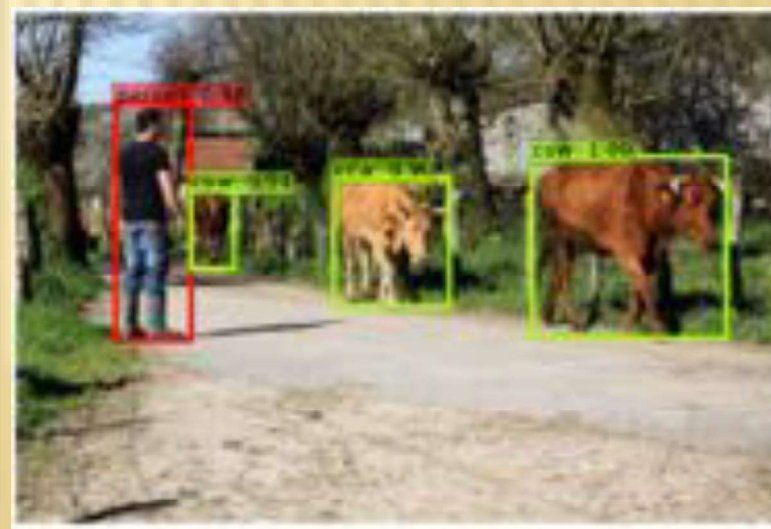
物体認識問題に対して高い性能を発揮していたVGG-16[1]と呼ばれるネットワーク構造に手を加え、物体検出の問題に応用できるようにした手法

SSDのネットワーク構造



(原著論文(<https://arxiv.org/pdf/1512.02325.pdf>)から引用)

[1] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014). (<https://arxiv.org/pdf/1409.1556.pdf>)

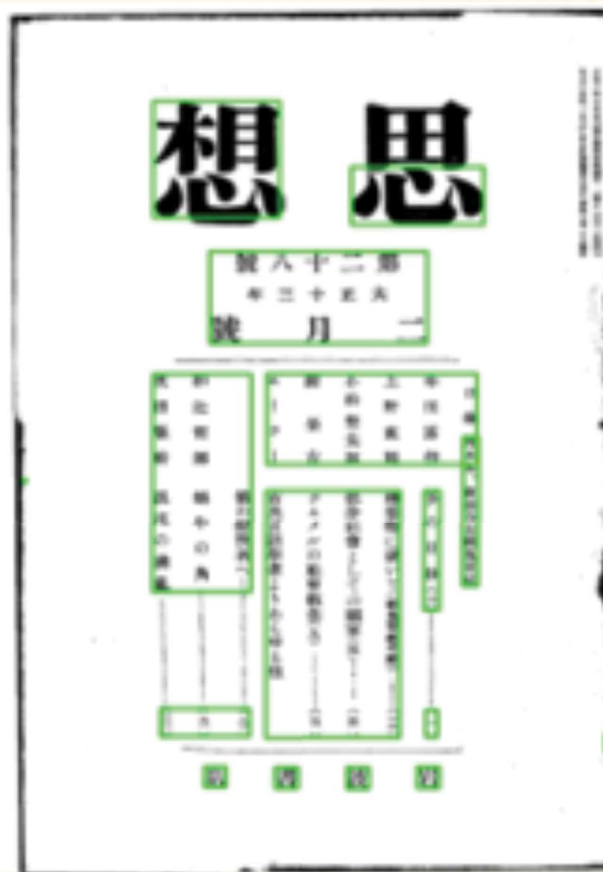


現状の結果の一部（1）

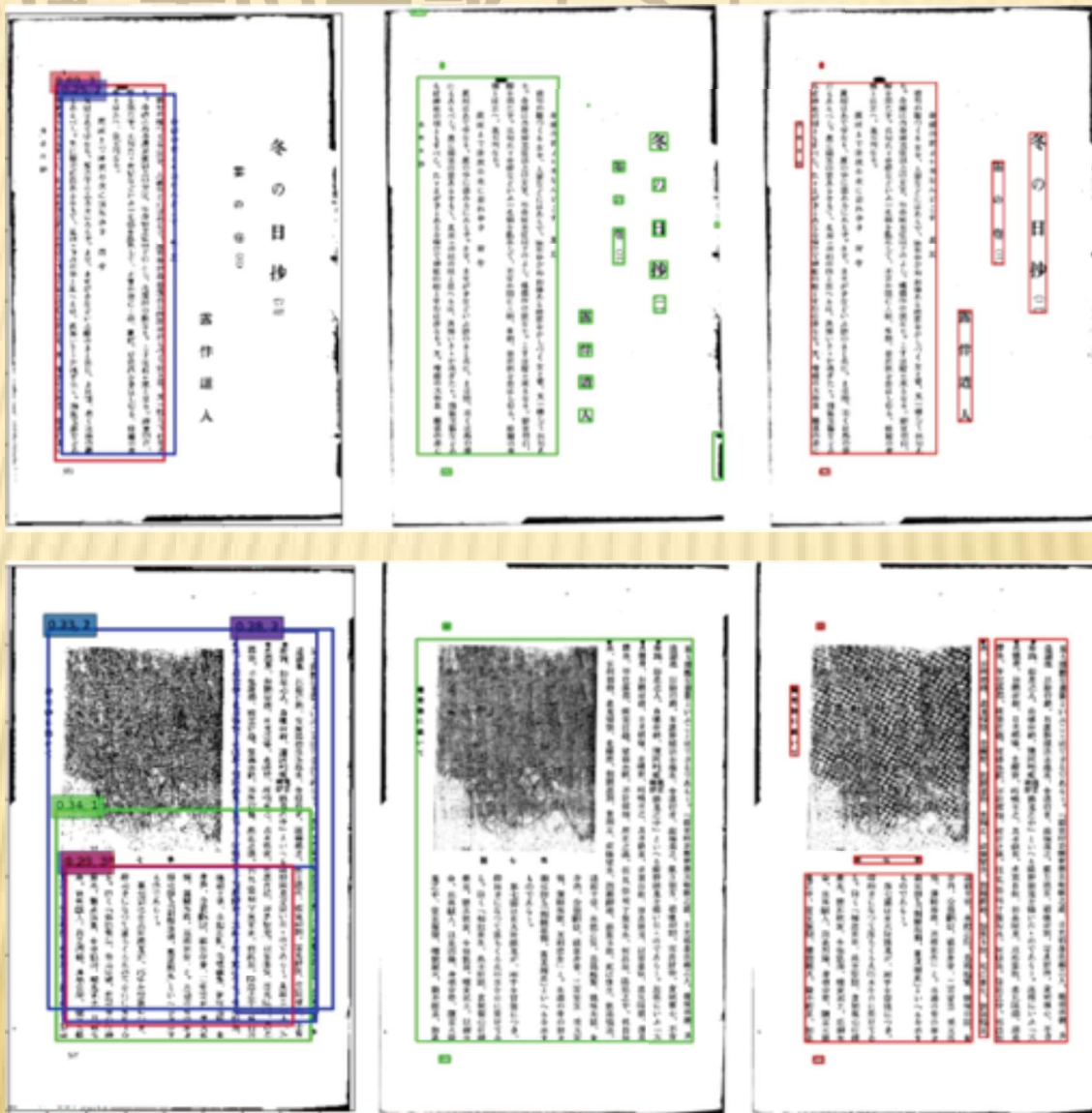
SSDによる認識結果

WinReaderPROによる認識結果

正解レイアウト



現状の結果の一部（2）



デジタル化資料の活用

**混沌とした知識を活用するためには
「知の構造化」が必要**

**「分野、組織、時勢を越えて、
知を(再)活用できるようにする」**

知(コンテンツ)を捉える

人が読む

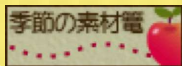
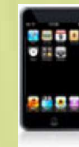
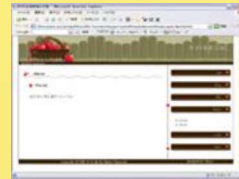
MIMAサーチ「思想」構造化

apple

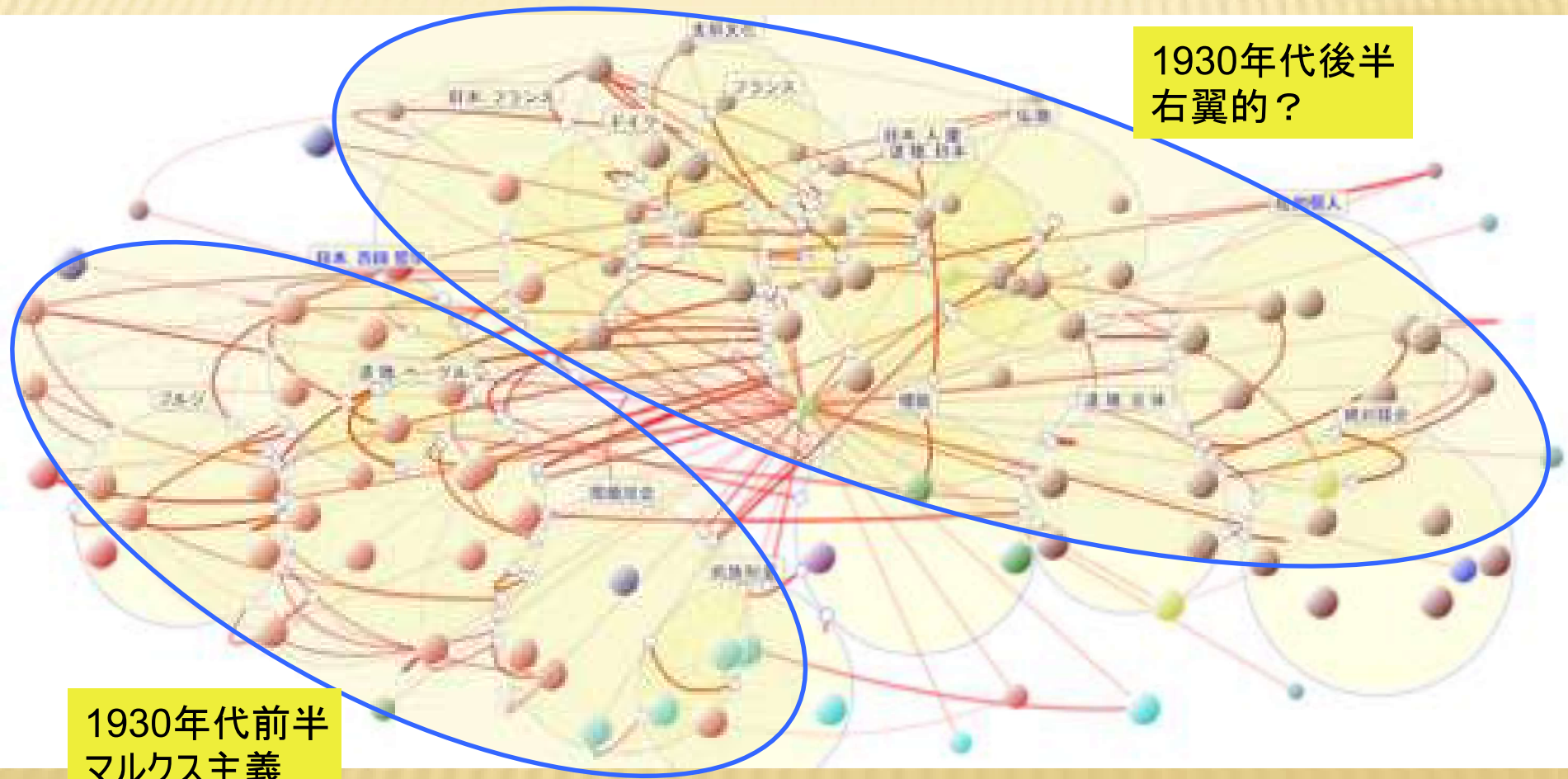
検索

果物

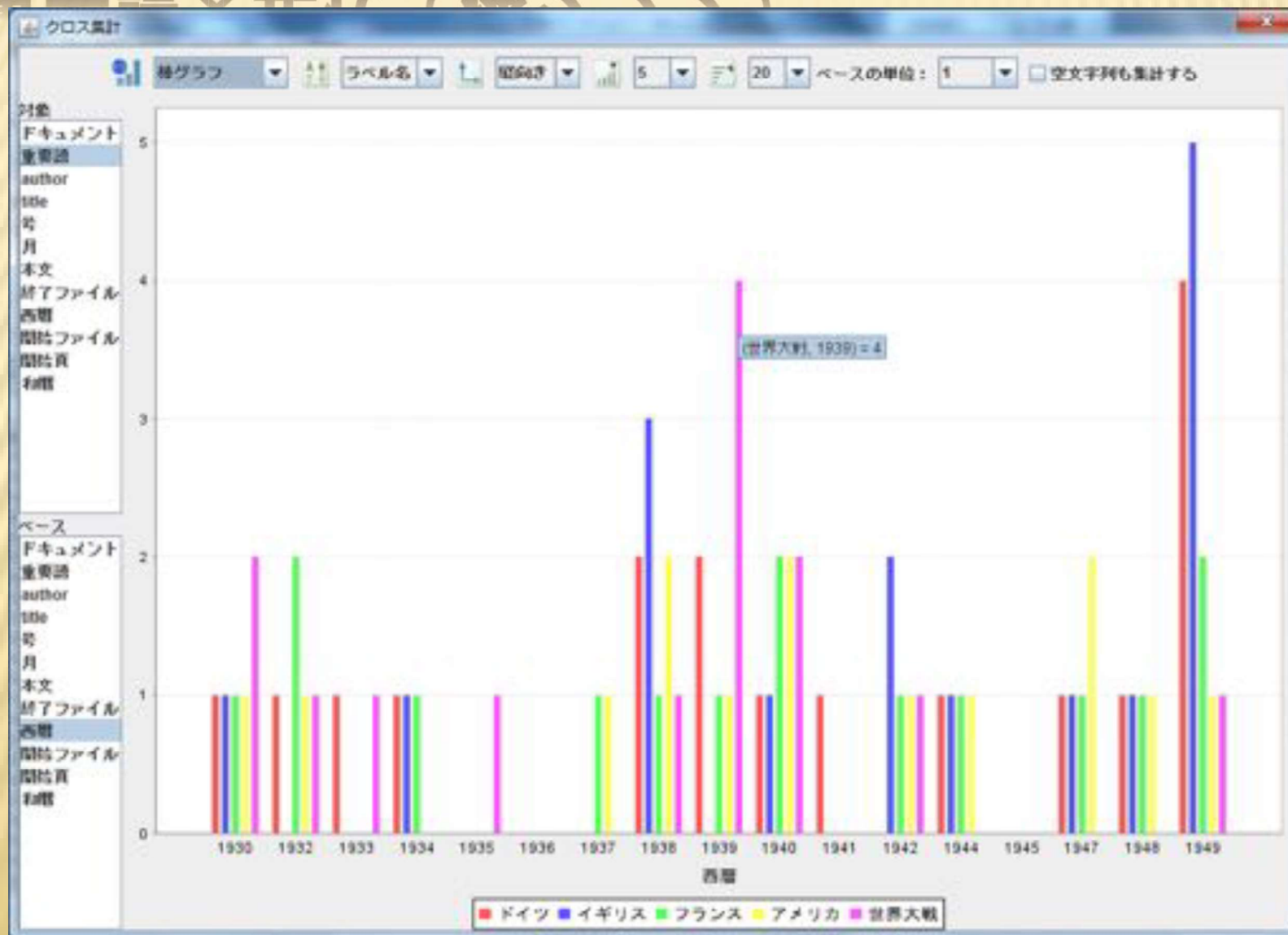
IT



MIMAサーチ可視化分析 「大和民族」に関する論述



属性毎のクロス集計 重要語×年代（棒グラフ）



属性毎のクロス集計 重要語×著者（円グラフ）



言葉の表現の集計

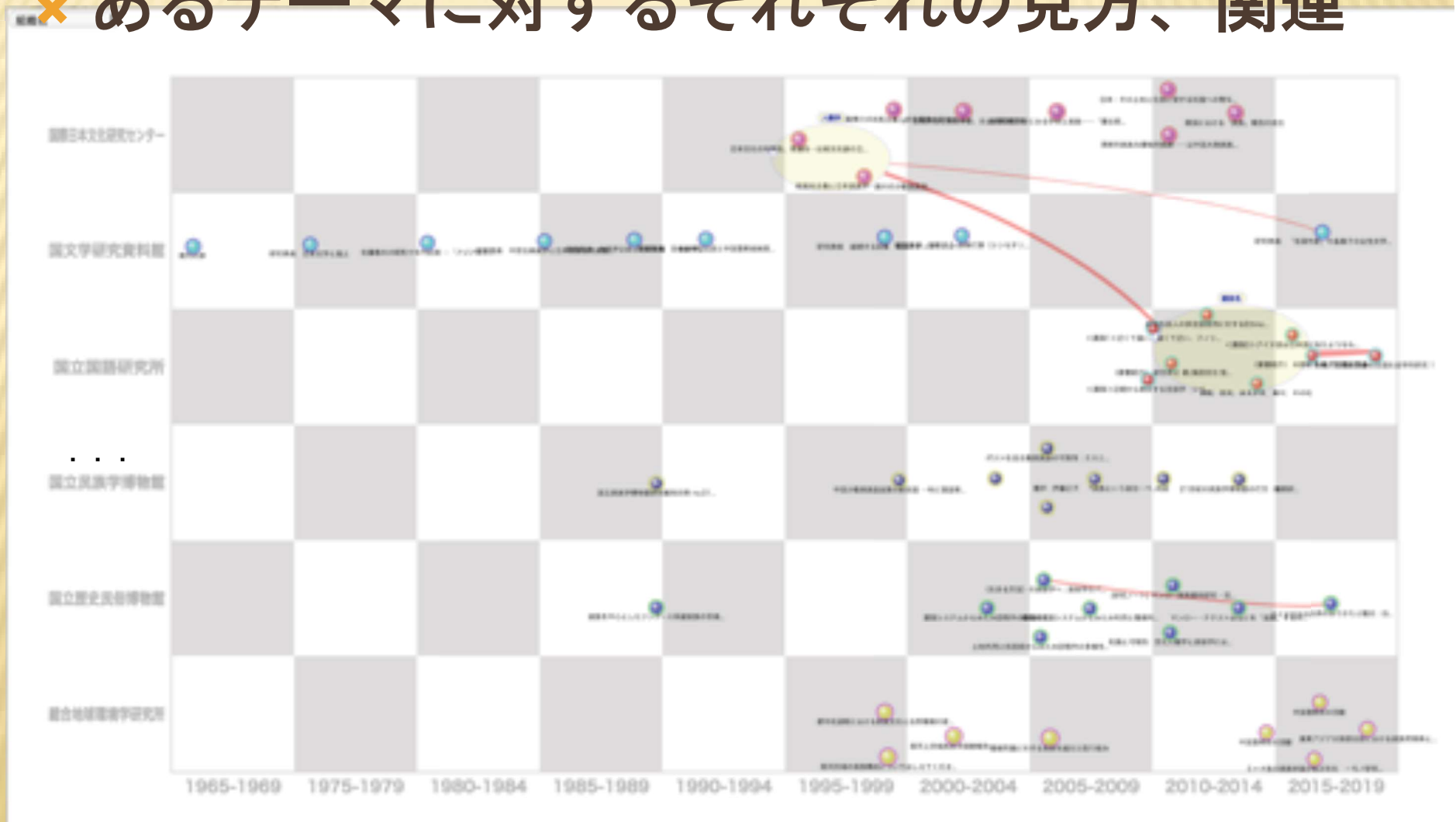
× 著者ごとによく使う表現を集計

著者: (著者リスト) キーワード:
 スコア順 頻度順 著者別順

著者	係り元	係り先	頻度	スコア	用例
三木清	構想力	論理	121	772.609	→用例
	論理	有る	89	186.431	→用例
	論理	論理	50	167.034	→用例
	形	論理	20	113.841	→用例
	論理	無い	42	109.242	→用例
	論理	言う	30	91.590	→用例
	経験	論理	16	91.073	→用例
	直観的悟性	論理	13	83.008	→用例
	論理	就く	21	78.669	→用例
	理性	論理	11	62.613	→用例
	論理	もの	19	59.415	→用例
	論理	考える	15	58.504	→用例
	論理	する	21	52.794	→用例
	論理	従う	13	51.835	→用例
	悟性	論理	8	42.293	→用例
	論理	因る	12	38.996	→用例
	想像	論理	6	38.311	→用例
	論理	居る	14	37.403	→用例

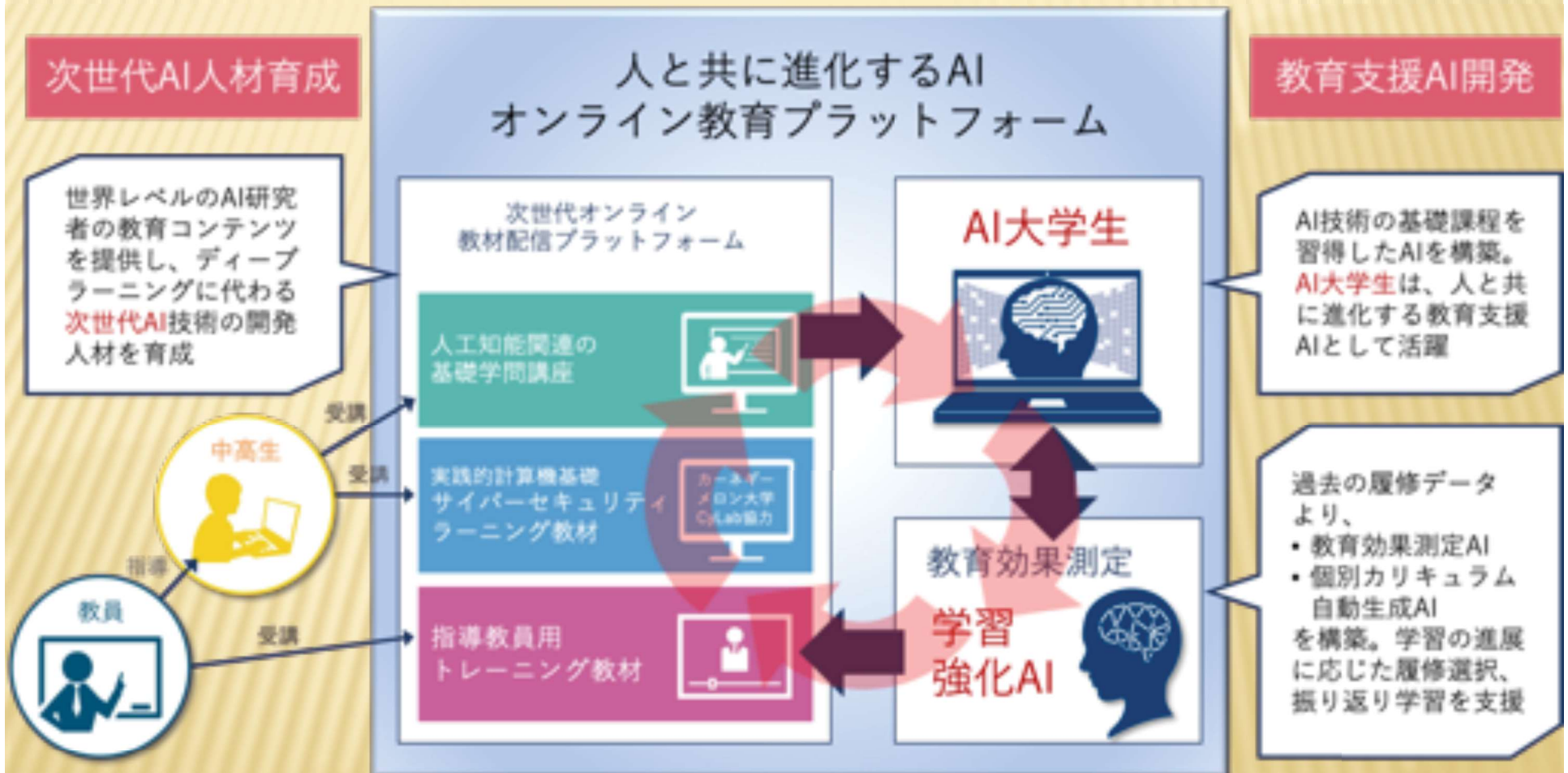
横断的な知見

✕ あるテーマに対するそれぞれの見方、関連



AIが読む

人と共に進化するAI オンライン教育プラットフォーム



活用のまとめ

■ AIが読む

- 質問応答
- 未来予測

人材育成

■ 人が読む

- 資料体へのアクセス・公開
 - 教育支援、教材づくり
 - 海外の研究者の支援
 - 近代の言語資源の価値化

ご清聴ありがとうございました

mima.hideki.8y@kyoto-u.ac.jp

MIMAサーチ

検索

