古文書とスーパーコンピュータに関するシンポジウム

古文書解読とくずし字資料の利活用サービス「ふみのは」と「くずし字AI-OCR」

2024年3月15日

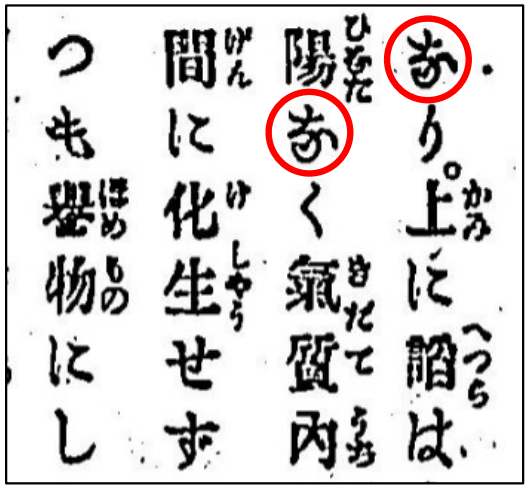
TOPPAN株式会社

TOPPAN 株式会社　について

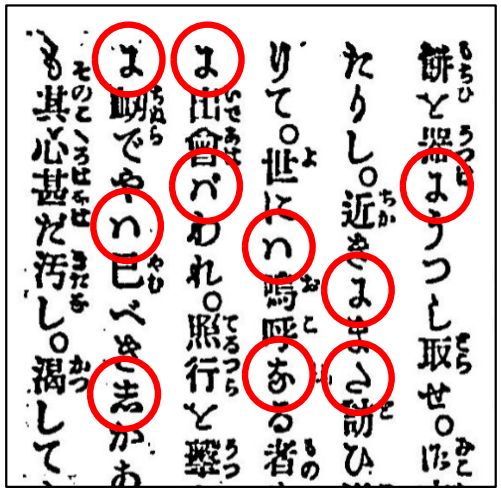
・創業：明治33年(1900年) ※
・2023.10社名変更
凸版印刷株式会社　⇒　TOPPAN株式会社
・売上高：約1兆4,600億円(連結)
・従業員因数：約53,000人(連結)

※凸版印刷創業の明治33年は、「小学校令」が施行され、平仮名が1音1字に制定された年です。

「変体仮名」が多用されている明治期の活字本



尾崎紅葉『紅鹿子』明治23年 dl.ndl.go.jp

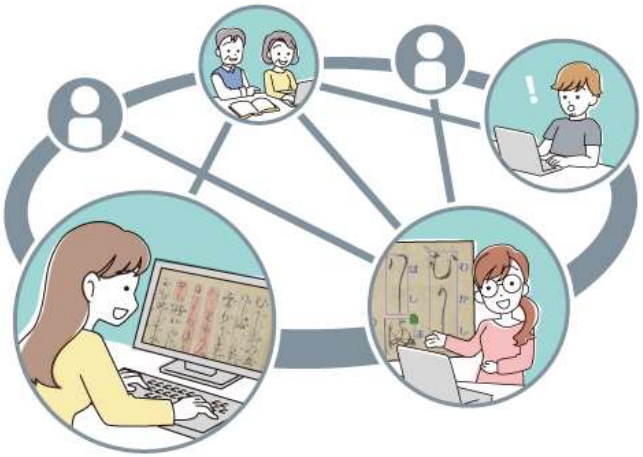


滝沢馬琴『三国一夜物語』明治20年 dl.ndl.go.jp

印刷博物館所蔵の「変体仮名の金属活字」

「ふみのは」サービス概要

TOPPAN は、現代人には難読になってしまった「くずし字」を、コンピュータの力を借りて読み解く試みとして、「くずし字OCR」技術の開発に取り組んできました。これまでの実証事業の結果をふまえ、2021年より**古文書解読とくずし字資料の利活用サービス「ふみのは®」**としてサービス提供しています。OCR＝光学文字認識技術



① 古文書解読サービス

最高水準の高精度くずし字AI-OCRを活用し効率的に解読した結果をお客様にお返しします。解読したテキストは、館内展示やインターネット公開に適した「ふみのはビューア」形式へ変換可能です。

② 古文書解読システム「ふみのはゼミ」（ASP提供）

高精度くずし字AI-OCRを搭載した古文書解読システムをPC/ブラウザで動作するソフトウェアとしてご提供します。複数人での同時作業をサポートし、授業・ワークショップ等での使用も可能です。

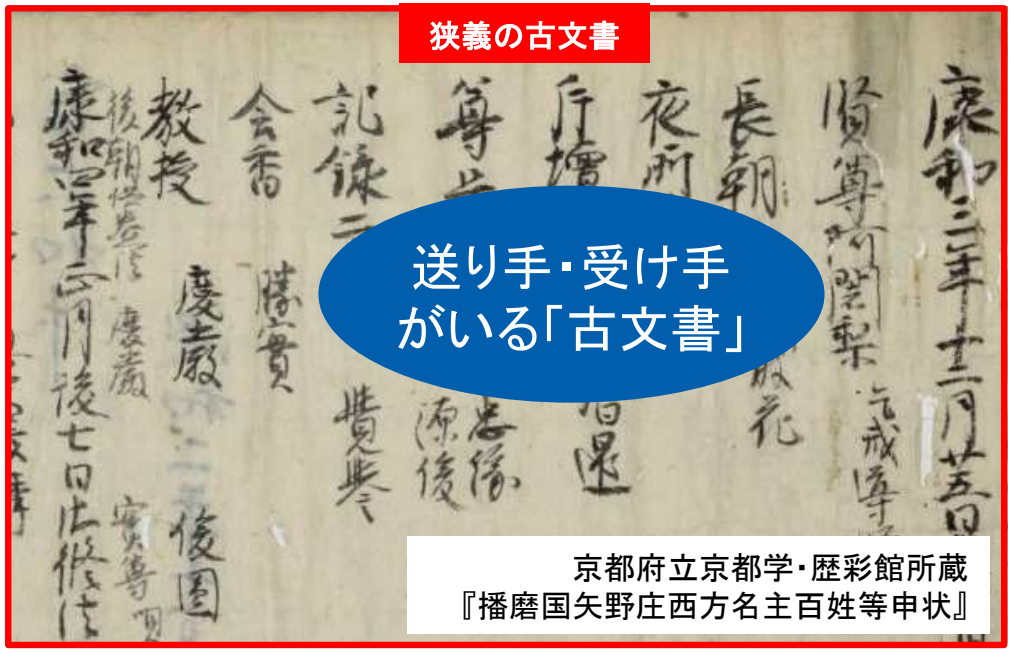
③ 古文書解読スマホアプリ「古文書カメラ」

ふみのはゼミで培ったくずし字AI-OCRを搭載。カメラと一体になることで、一般のお客様向けのサービス提供、資料館の整理業務でのご利用が可能になりました。

「ふみのは」のサービス対象

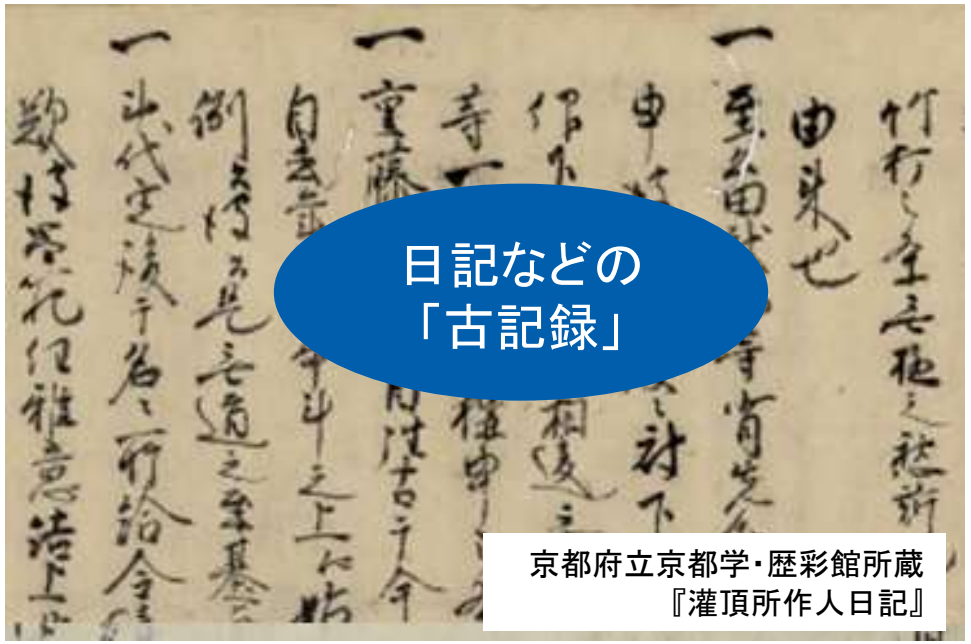
広義の古文書

狭義の古文書



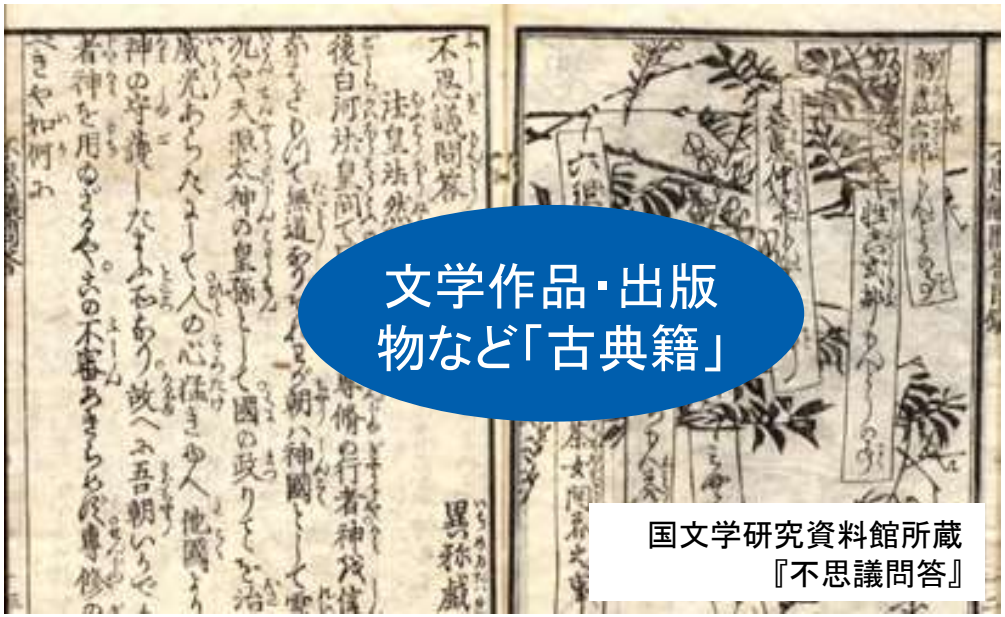
送り手・受け手がいる「古文書」

京都府立京都学・歴彩館所蔵
『播磨国矢野庄西方名主百姓等申状』



日記などの「古記録」

京都府立京都学・歴彩館所蔵
『灌頂所作人日記』



文学作品・出版物など「古典籍」

国文学研究資料館所蔵
『不思議問答』

浮世絵など絵画、美術品

東京大学総合図書館所蔵
『しんよし原大なまづゆらひ』

「くずし字AI-OCR」開発の経緯

創業（明治33年 ※1900年）以来培ってきた、
印刷会社としての文字・組版に対するノウハウ

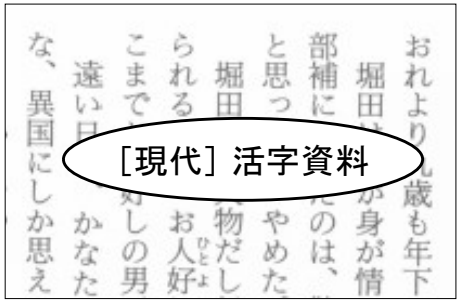
2013年～高精度全文テキスト化サービス

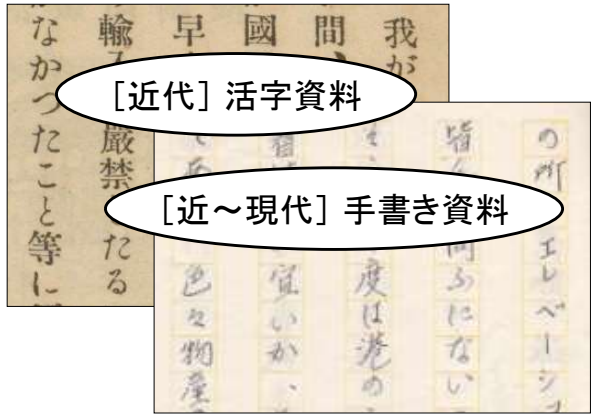
OCR活用による効率的な電子書籍制作

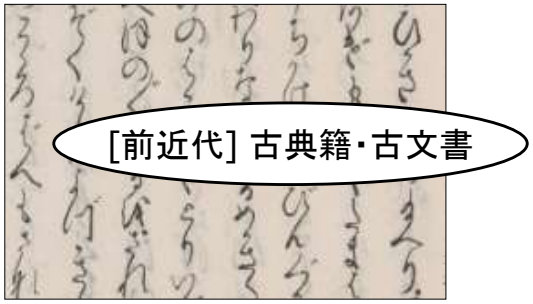
資料アーカイブ ＋ OCRによるテキスト化
※近代（明治～昭和初期）の活字・手書き資料

2016年：実証実験開始～2021年：サービス開始

くずし字OCR







本日お話しする内容

- くずし字とAI-OCR
- 人力 vs OCR

  製造業からみた「OCR技術」
- 「ふみのは」サービスについて

くずし字とAI-OCR

文字の歴史とくずし字による「情報の断絶」

（原始）	旧石器時代	– 紀元前14000年頃
	縄文時代	前14000年頃 – 前4世紀
	弥生時代	前4世紀 – 後3世紀中頃
古代	古墳時代	3世紀中頃 – 7世紀頃
	奈良時代	710年 – 794年
	平安時代	794年 – 1185年
中世	鎌倉時代	1185年 – 1336年
	室町時代	1336年 – 1573年
	安土桃山時代	1573年 – 1603年
近世	江戸時代	1603年 – 1868年
近代	明治時代	1868年 – 1912年
	大正時代	1912年 – 1926年
現代	昭和時代	1926年 – 1989年
	平成時代	1989年 –

楷書・行書・草書は、前漢～唐代に確立。

「書聖」王義之（303-361）は東晋の政治家

日本へ輸入

草書ベースに「かな」が発明される

文字の形が安定していて、大量の文書が残っている

＝OCR向き

近代の手書き文字は最難関

くずし字の時代（1000年）

？

？

1,000年分の膨大な史資料に直接アクセスできない状況

断絶

旧字旧仮名遣い

断絶

新字新仮名遣い

活字の時代（150年）

400年前の本（バルセロナ）

カソリック教会と歴代教皇を中心として、世界各地で生じた出来事を編年体で綴った書物。100年以上かけて全6巻構成で刊行された歴史書の第3部に当たるもの。秀吉によるバテレン追放令や、天正遣欧使節に関する記事などが随所に掲載されており、貴重な作品。

TERCERA PA
DE LA HISTORIA PO
FICAL Y CATOLIC

Vida de Gregorio XIII. Pontifice

CAPITVLO PRIMERO. C

Eleccion de Gregorio XIII. Ratifica la liga que su antecessor auia hecho con el Rey Catolico, y Venecianos. Da priesa para que salga el armada, y por salir tarde no haze el efeto que se desseaua.

VE Gregorio Decimoter cio natural de Bolonia, ciu dad bien conocida en Ita. lia, que este principio me cõuiene lo sea de mi histo- ria: porque si bien el Doctor Gonçalo de Illescas en el segundo tomo de la suya (que yo quiero D proseguir) dexò começada la vida de Gregorio, no me he podido escusar de escriuirla desde su principio: tanto, porque lo q̃ della

lio este año de quinien
Quiẽ dize, que le ocasi
nas diſſenſiones dome
no le ſucedia muy prot
q̃ auia tomado de la le
ua los premios mereci
tras, coſa q̃ de ordinar
Vniuerſidades, adonde
del mundo) nunca falta
quien mejor lo merece
coſa marauillosa la diu
trató, y todos con gran
dencia. En el Pontifica
Concilio de Trento: y
ma, fue teniente del A
en aquella sazon lo er
despues Cardenal de S

『教皇とカソリックの歴史：第3部』1609年刊

200年前の本（日本）

江戸時代後期、「草双紙」と呼ばれた一般的な絵入り小説の形式。

本の大きさは現代の単行本コミックスと同等。平仮名が多く、丸みを帯びた形で書かれている。

『東海道中膝栗毛』の作者十返舎一九（じっぺんしゃいっく）作の『小夜衣(さよごろも)物語』文政4 (1821)刊行。

たった200年前の娯楽小説が
現代日本人には読めない！

『小夜衣物語 6巻』　1821刊　国立国会図書館蔵

「製造業」からのご提案

人材不足

高齢化

大量に存在する未翻刻古文書とデジタル化の遅れ

属人化

手作業

非効率

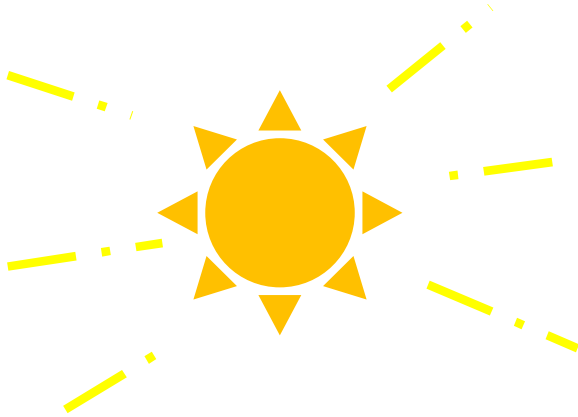
TOPPAN

作業工程の分解と再定義

専門家/非専門家/自動処理の分業

OCR関連技術活用

電子化や文字のノウハウ



効率化

古文書のデジタル化と利活用による知の創出

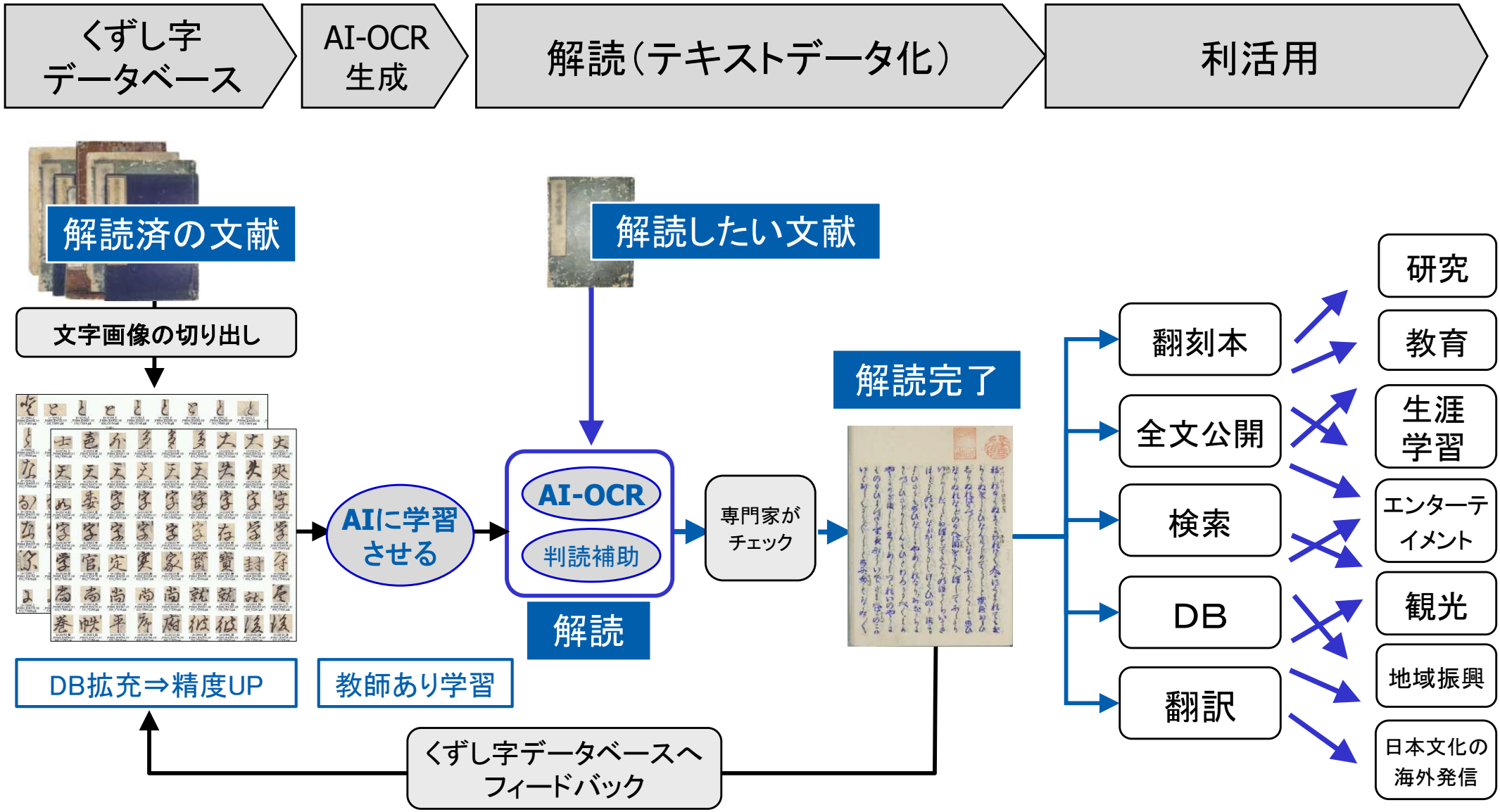
後進育成

文化継承

地域創生

くずし字AI-OCRの仕組み

AI（人工知能）技術の進化により、従来は不可能だった、高精度に「くずし字」を識別する文字認識エンジン（AI-OCR）の開発が可能になりました。



くずし字データベースのイメージ 「な」

現状、近世（江戸時代）の類型的なくずし字は版本・手書きともにかなりの精度で認識できるようになっています。

ただし、「書き癖の強い古文書も80%以上の精度で読める」ためにはまだまだ不十分のため、現在急ピッチで拡充中です。

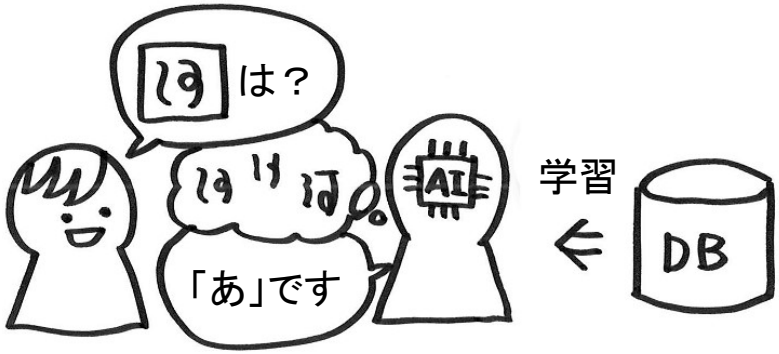
※国文学研究資料館所蔵「絵入源氏物語」（承応3年-1654年）

（現在の）くずし字AI-OCRにできること

膨大な字形データから形のよく似た文字を提示

膨大な字形データベース（教師データ）に含まれる「形のよく似た文字」を高速に検索して候補を出してくれる仕組みです。

文字の切れ目（連綿体の区切り位置）と文字コードの推定を同時に行うことも可能です。



AIが読めなかった文字は人が教えることが可能

教師データに含まれずＡＩが判読できなかった字形をAIに再学習させることで、認識精度を段階的に改善していくことが可能です。

（現在の）くずし字AI-OCRへの過剰な期待…

AIだから当然
できるでしょ？

- あらゆる筆跡のくずし字が読めるようになった。
- 人間が間違いを修正するとリアルタイムでAIが学習し、どんどん賢くなる。
- 学習データを用意しなくても、AIが勝手に学習してくれる。
- 人類未解読の文字もAIが読み解いてくれる。（人間が読めない文字も読める）

現状の「教師あり学習」ベースのAI－OCRに対しては、過剰な期待。

ただし、昨今の大規模言語モデル（LLM）に代表される、急激なAI関連技術の進歩により、状況は変わる可能性が高い。

おそらく、課題はコスト

人力 vs OCR

精度95%の壁 （対象：活字または現代の手書き文字）

経験則

高精度テキスト（99.5%以上）の作成を目的とするならば、

- 認識精度90%程度のOCRでは、最初から手入力の方が低コスト
- OCRを導入して効果が上がるには、安定して認識精度95%以上が必要

精度90%

文字を目視判読＋手入力
（100文字／1分）

＜ コスト

OCR＋目視間違い探し＋再入力
（100文字／1分以上）

最初から全文を手入力
（慣れるとゲーム感覚で楽しい）

＜ コスト

OCR誤認識文字の修正作業
（イライラ、辛い…）

タッチタイピング練習ソフト

辛い作業は、報酬を高めに設定しないと作業者の調達ができないから

「精度95%の法則」は、くずし字にも適用可能か？

経験則

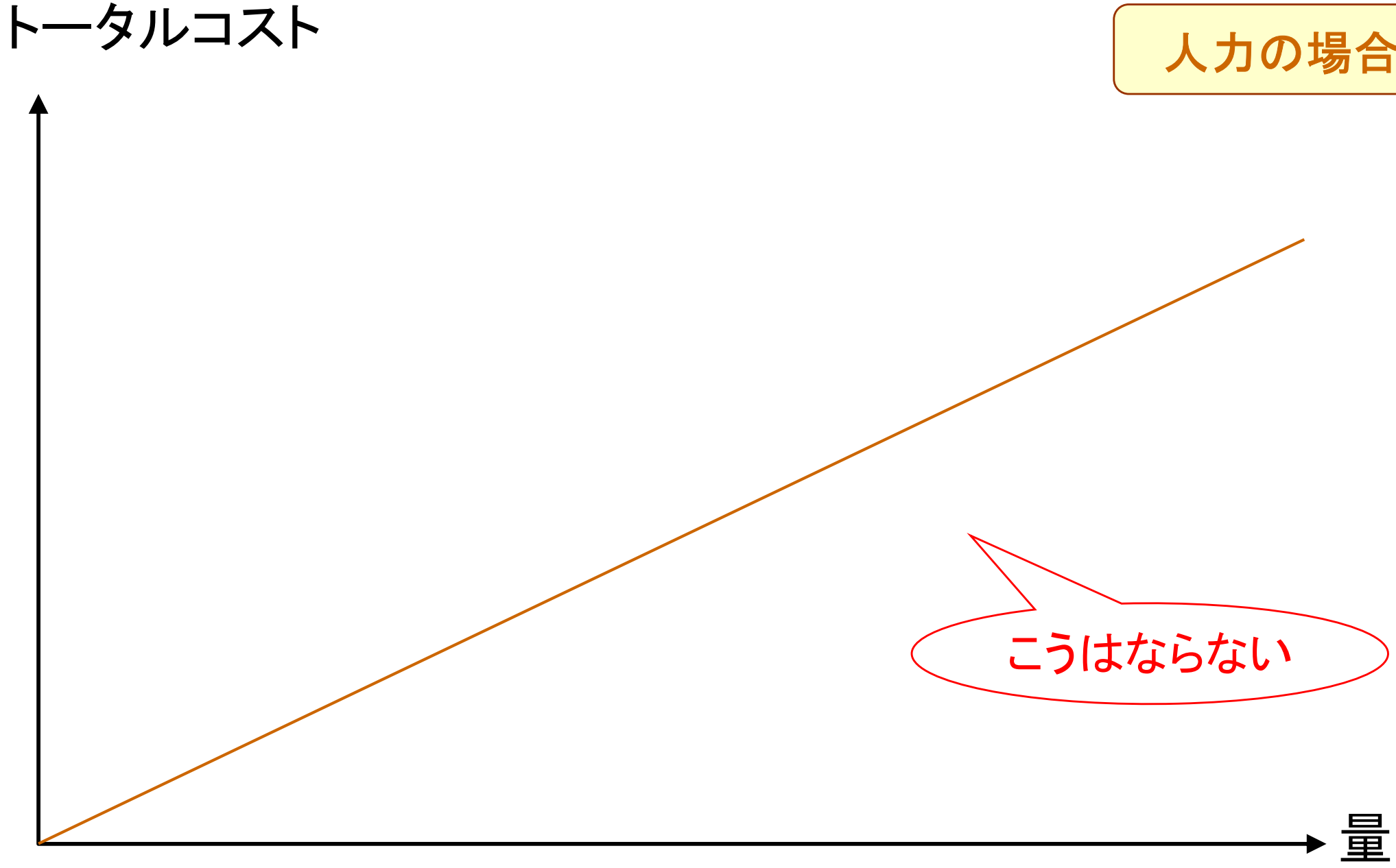
高精度テキスト（99.5%以上）の作成を目的とするならば、

- 認識精度90%程度のOCRでは、最初から手入力の方が低コスト
- OCRを導入して効果が上がるには、認識精度95%以上が必要

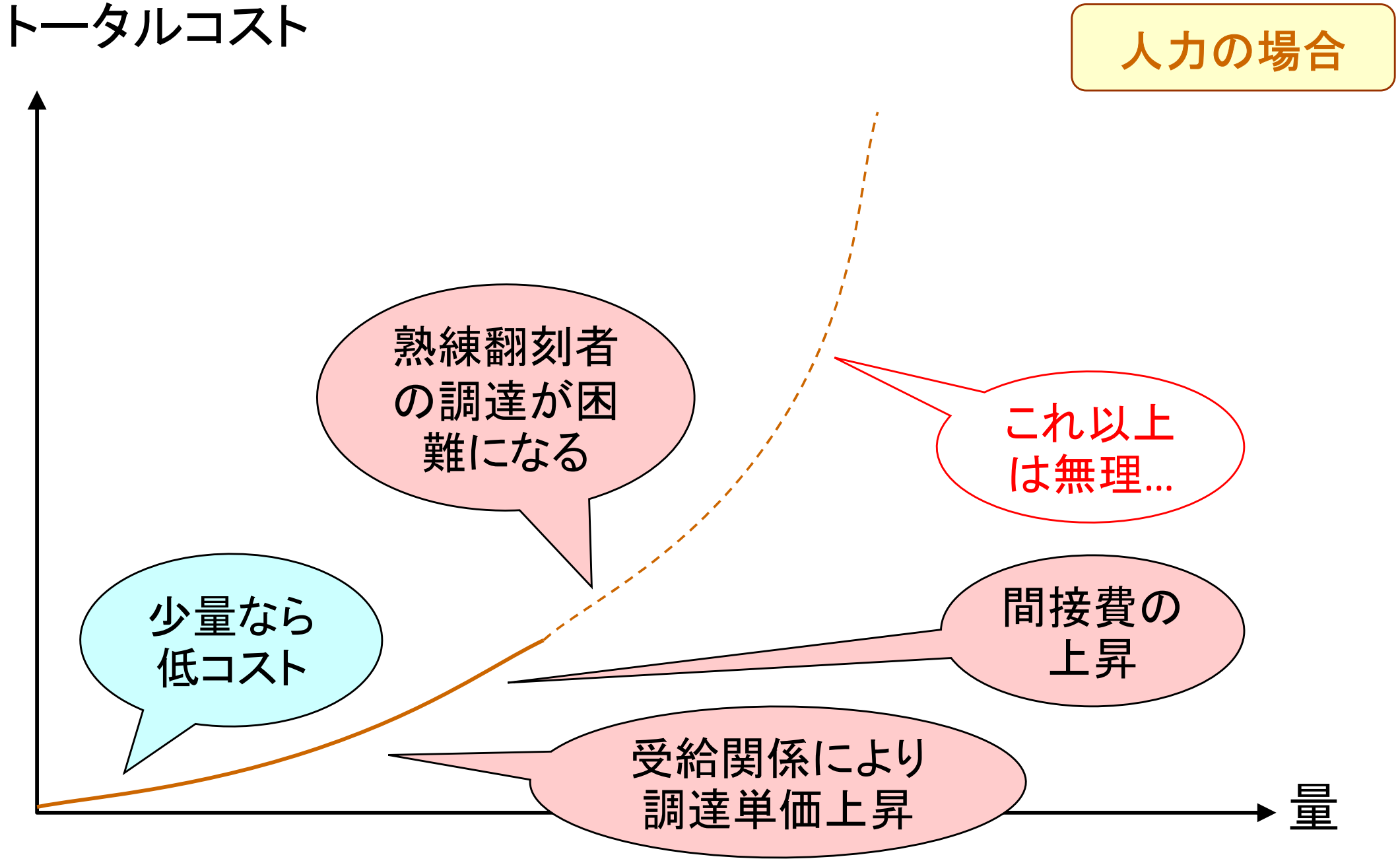
くずし字でも同じことが言えるか？

- くずし字で100文字／1分の入力は困難では？
- 読める人材の大量調達は困難では？
- OCR認識結果の修正はそこまで苦行か？
- AIに「下読み」をさせることで効率化するのでは？
- 候補文字が出ることで初～中級者も参画できるのでは？

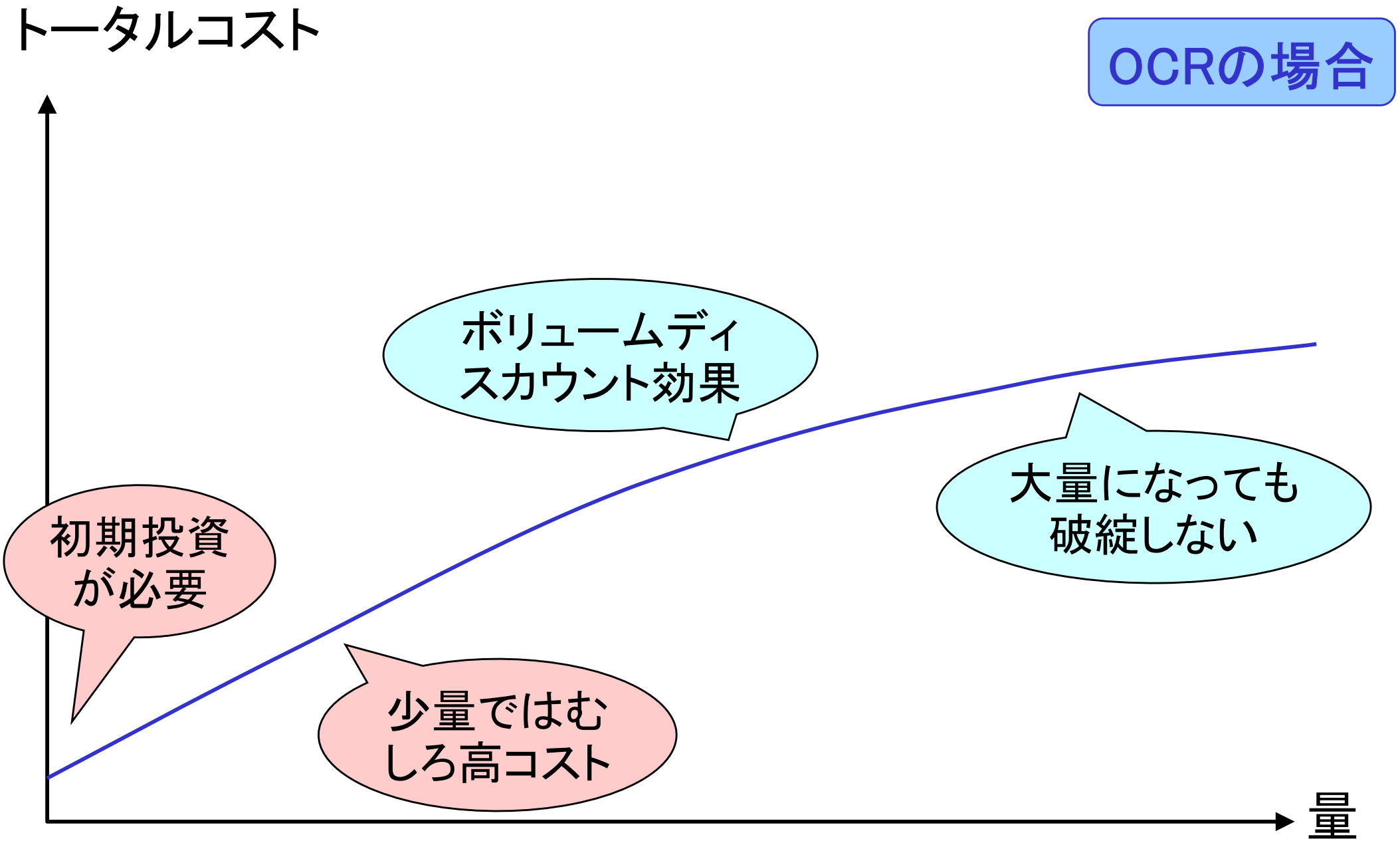
コスト構造比較（人力）



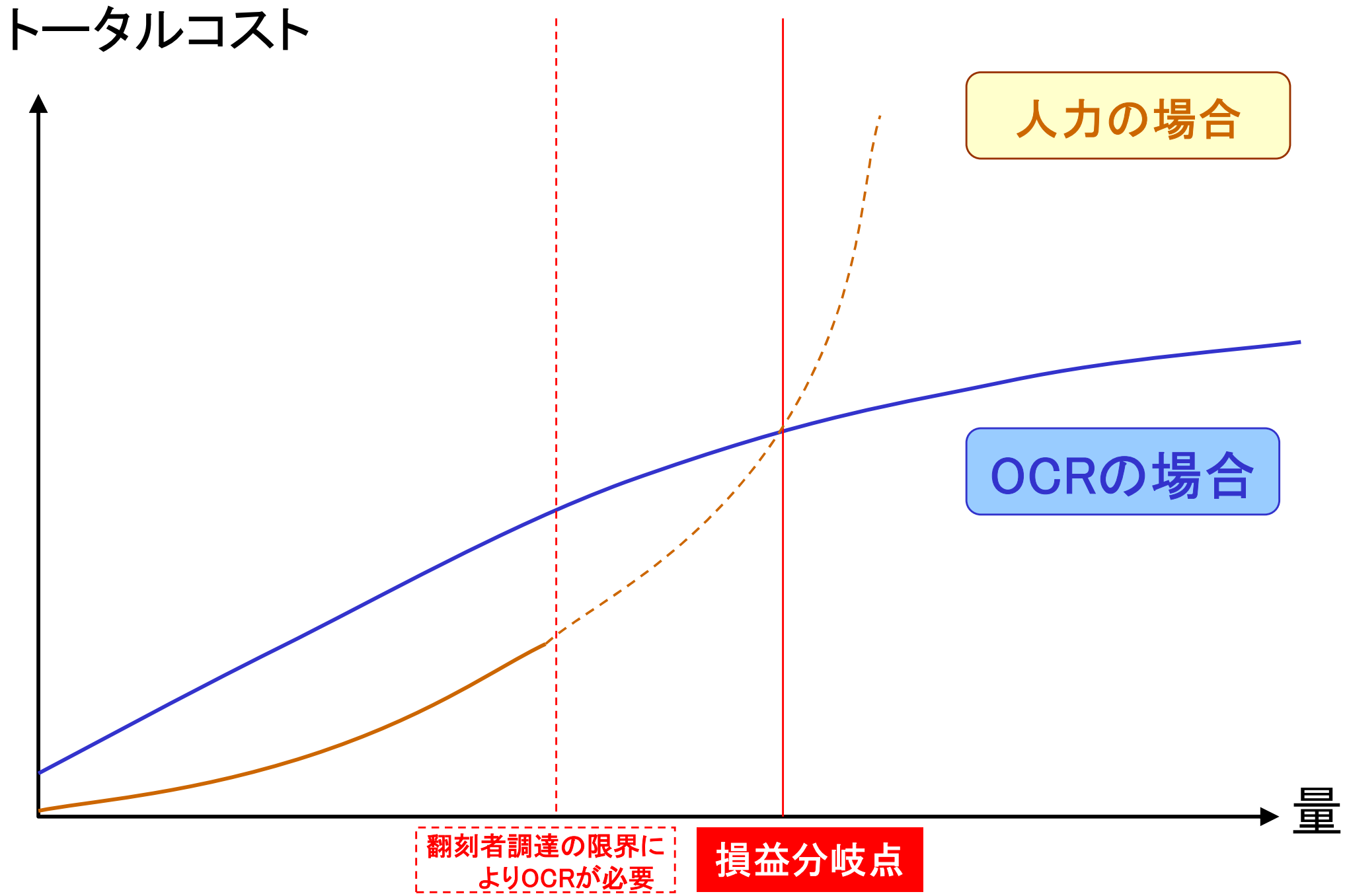
コスト構造比較（人力）



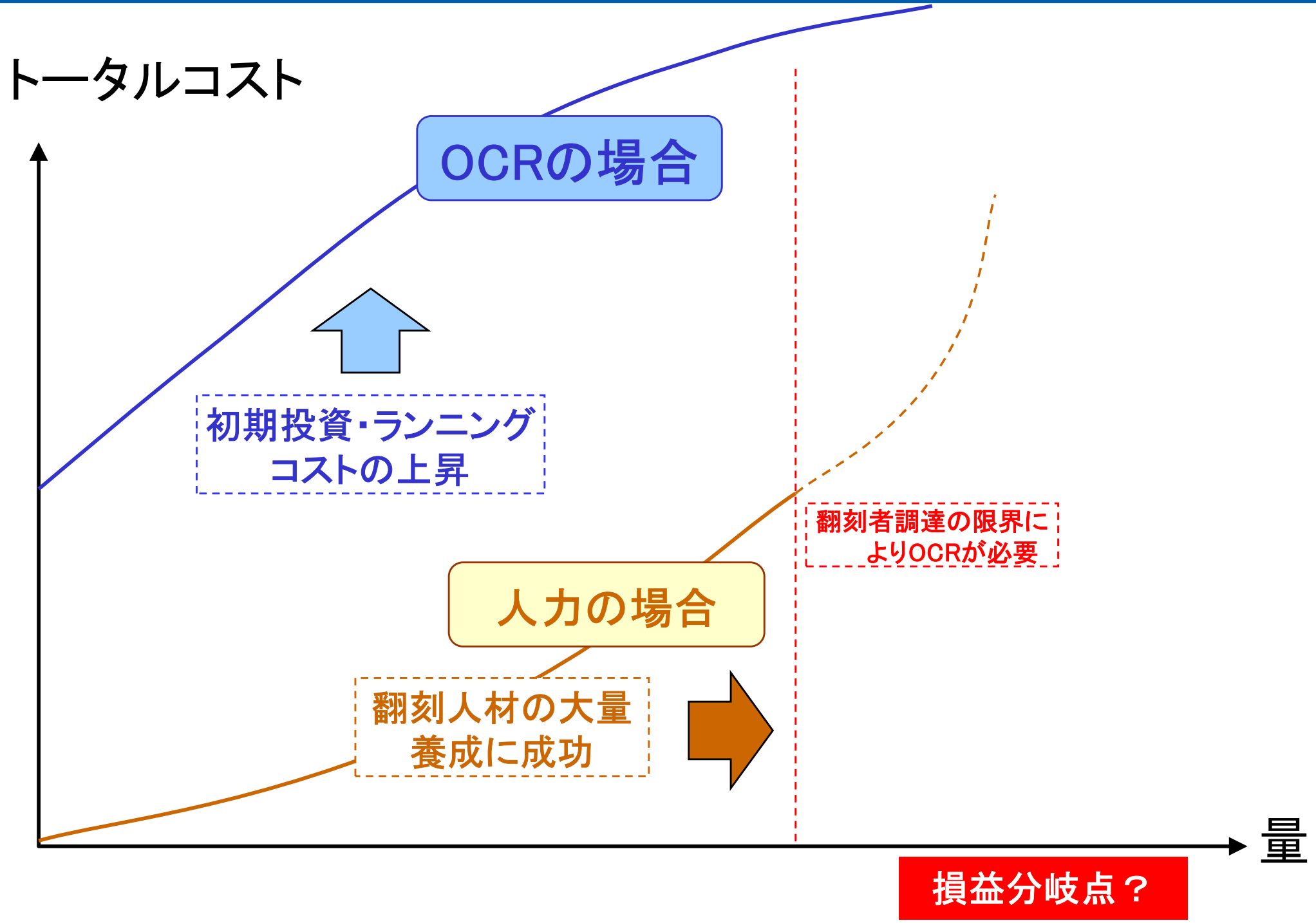
コスト構造比較（OCRの場合）



コスト構造比較



コスト構造比較（OCRにとって不都合なシナリオ）



くずし字AI-OCRの精度

Information

TOPPAN

2023 年 12 月 12 日
TOPPAN 株式会社

TOPPAN、古文書解読スマホアプリ「古文書カメラ」のアップデートを実施

AI-OCR の改良と画像編集機能の追加により解読精度を大幅に向上。
2023 年 12 月 16 日、17 日開催のお城 EXPO 2023 などに出展

TOPPAN ホールディングスのグループ会社である TOPPAN 株式会社（本社：東京都文京区、代表取締役社長：齊藤 昌典、以下 TOPPAN）は、スマートフォンで撮影したくずし字資料を高精度の AI-OCR 技術によりその場で手軽に解読できるくずし字解読アプリ「古文書カメラ®」を、2023 年 6 月より iOS 版、2023 年 10 月より Android 版で提供しています。

2023.12.12のニュースリリースより抜粋

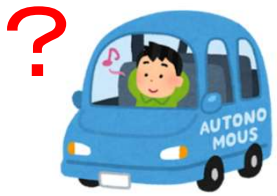
対象資料	6月版	12月版（今回）	精度改善
サンプルA（版本）	約85%	約90%	**+5pt**
サンプルB（古文書　読みやすいもの）	約55%	約84%	**+29pt**
サンプルC（古文書　読みにくいもの）	約46%	約69%	**+23pt**

（Web翻刻システム）

「ふみのはゼミ」システムのご紹介

凸版のAI-OCR付き翻刻システム（ふみのはゼミ）の開発コンセプト

人とテクノロジーとのちょうどよい距離感の設計＝潜在能力のブースト



AIによる自動運転

ではなく

電動アシスト自転車

ロボット

ではなく

パワードスーツ

（1）人間の作業をアシストしてくれるようなAI

- 人間にもAIにも簡単に読める文字は予め入力しておいてほしい
- 専門家でも判断に迷ったときに、AIが判断する候補文字があると参考になる
- 初学者が学習を始めるハードルを下げ、分野のすそ野を広げる

（2）AIは間違うという前提で、

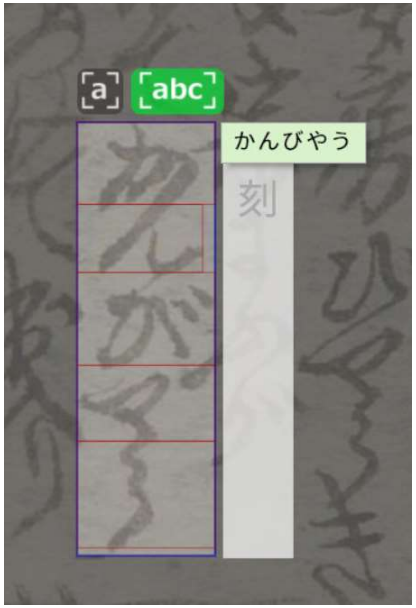
- 間違った結果を人間が簡単に修正できる仕組みが必要
- 正しい結果を簡単にAIにフィードバックすることができる仕組みが必要

「ふみのはゼミ」システムの特長

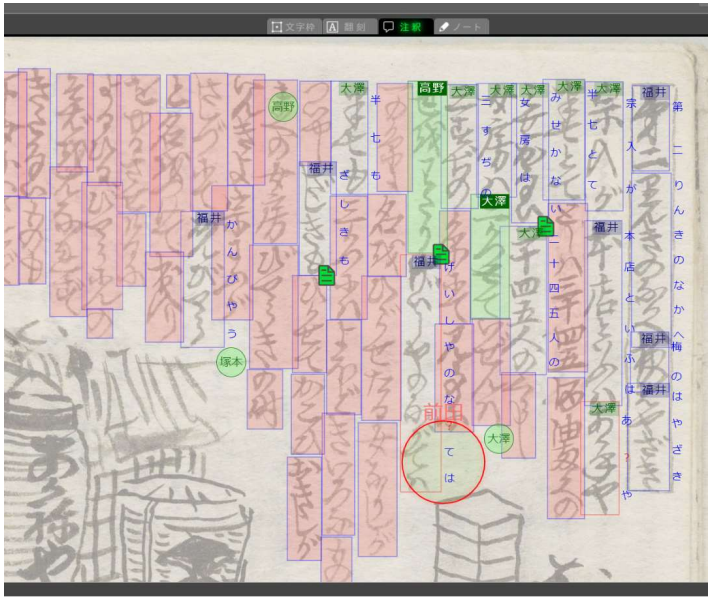
特長

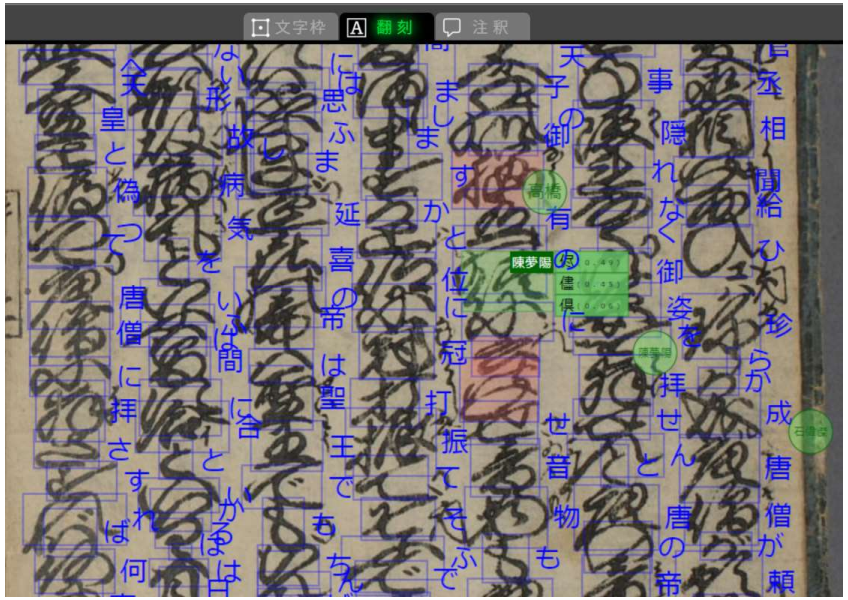
- 最新のくずし字AI-OCRが人の解読力をサポート
- オンラインでの複数人共同作業
- 教育機関やワークショップでの利用を想定した採点機能※等

※別途正解データが必要



AI-OCRによる文字認識





ふみのはゼミを使用した複数人での同時翻刻の様子

印刷・製造業のノウハウを結集した効率的な解読作業を実現！

スマートフォン用古文書解読アプリ「古文書カメラ」

「ふみのは」サービスで培った高精度AI-OCRエンジンを搭載。
その場で手軽に解読できる古文書解読アプリです。

撮影
読みたい古文書を撮影して取り込むだけ

解読
古文書 AI-OCR搭載でくずし字を読み解く

編集
解読結果はアプリ内で編集・書き出しが可能

ふみのは
2024.6 iOS版配信開始
2024.10 Android版配信開始

2つの解読方法 （全自動、部分解読）

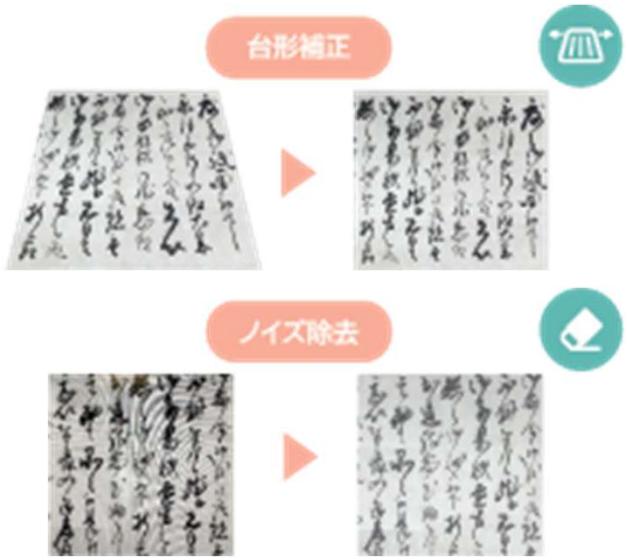

「全ての文字を解読」ボタンをタップ


解読後の文字枠をタップして修正もできます


読みたい文字をタップ


AI の候補一覧から選択できます

画像編集機能



ご清聴ありがとうございました