

OCR を用いたデジタル画像
の全文テキスト化
実施結果報告書

平成 22 年度

国立国会図書館

目次

1. プロジェクトの概要	3
1.1. プロジェクトの背景	3
1.2. プロジェクトの実施方法	3
1.2.1. 対象資料	3
1.2.2. 作業実施者	3
1.2.3. OCRソフト	3
1.2.4. 作業期間	3
1.2.5. 作業実施方法	4
2. プロジェクトの実施結果	5
2.1. 文字認識率の集計	5
2.1.1. 文字認識率算出結果(サマリ)	5
2.1.2. 低認識精度(サマリ)	6
2.1.3. 低認識精度(画像)	7
2.2. 文字認識率のクロス集計	9
2.2.1. 文字認識算出結果(刊行年別)	9
2.2.2. 文字認識算出結果(NDC分類別)	10
2.3. 辞書更新	11
2.3.1. 辞書更新対象資料	11
2.3.2. 辞書更新対象文字	11
2.3.3. 辞書登録文字数	11
2.3.4. 辞書更新効果	12
3. 課題	14
3.1. プロジェクト実施工数	14
3.1.1. 本案件での課題	14
3.1.2. 今後の検討課題	14
3.2. 画像品質向上	15
3.2.1. 本案件での課題	15
3.2.2. 今後の検討課題	15
3.3. 縦横文書	15

3.3.1. 本案件での課題.....	15
3.3.2. 今後の検討課題.....	16
3.4. ノイズの除去.....	16
3.4.1. 今後の検討課題.....	16
付録.....	19
1. OCR作業に関する設定値.....	19
2. 文字認識率算出方法.....	20
2.1 作業対象.....	20
2.2 手順.....	20
2.3 算出上のルール.....	20
2.4 文字認識率の算出例.....	20
3. データ容量.....	22

1. プロジェクトの概要

1.1. プロジェクトの背景

デジタル・ネットワーク社会における出版物の利活用の推進及び平成 22 年 1 月に施行された改正著作権法による視覚障害者等へのサービス拡充に向けた取り組みの一環として、国立国会図書館では、平成 22 年度内に、全文テキスト化実証実験を実施することとなった。この実証実験では、和図書を中心とする当館所蔵資料のデジタル画像からテキストデータを作製し、①視覚障害者等向けの読上げサービス等に関する課題(アクセシビリティの向上)、②全文テキストデータの検索に関する課題(サーチャビリティの向上)の 2 つの技術的課題を検証する。

本案件で作製する全文テキストデータは、実証実験で構築するプロトタイプシステムに投入するものであり、主として全文テキストデータの検索に関する課題を検証するためのテストデータとして使用される。大量かつ幅広い年代の資料を対象に実験を行うことで、様々な文字認識率の全文テキストデータに対する検索性能等の検証を実施することが可能となる。

1.2. プロジェクトの実施方法

1.2.1. 対象資料

全 20,000 冊(2,614,703 コマ¹) 明治期～昭和戦後期刊行図書

表 1.1 作業対象資料

区分	対象冊数	対象コマ数	ファイル形式	解像度	階調
明治期刊行図書	5,000 冊	637,375 コマ	JPEG2000	400dpi	2 値
大正期刊行図書	10,000 冊	1,227,471 コマ	JPEG2000	350dpi	グレイ
昭和戦前期行図書	4,790 冊	702,715 コマ	JPEG2000	350dpi	グレイ
昭和戦後期刊行図書 ²	210 冊	47,142 コマ	JPEG2000/ JPEG	400dpi	カラー

1.2.2. 作業実施者

東芝ソリューション株式会社

(工程管理支援：アクセンチュア株式会社)

1.2.3. OCRソフト

ドキュメントリーダー Express Reader Pro 東芝ソリューション(株)製

1.2.4. 作業期間

平成 22 年 11 月～平成 23 年 1 月

¹ 1 コマは、見開き 2 ページを示す

² 昭和戦後期刊行図書には雑誌 113 冊を含む

1.2.5. 作業実施方法

- (1) デジタル画像全 20,000 冊を対象とし、OCRを用いて全文テキスト化作業³を実施する。
- (2) テキスト化に当たっては、校正作業⁴・辞書更新⁵を行い、その前後での認識率の向上の有無を評価した。なお、認識率の算出⁶は、各冊 1 コマを抽出して行った。

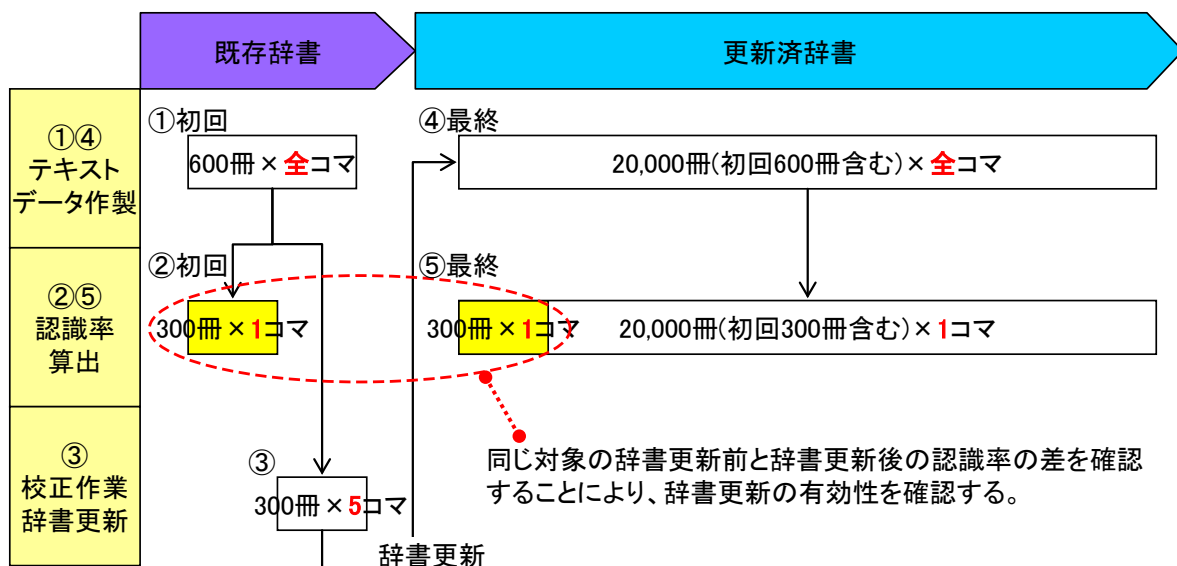


図 1.1 作業の概要

³ 全文テキスト化を実施時の OCR 設定項目は、「付録 1」参照

⁴ テキストと実際の画像を比較し、OCR 誤認識が発生している文字を修正

⁵ 校正作業を実施した文字を OCR 辞書に登録すること。登録対象方法は「2.3.2. 辞書更新対象文字」参照

⁶ 文字認識率(%) = (正しく認識された文字数 / 元画像上の文字数) × 100。認識率算出方法の詳細は「付録 2」参照

2. プロジェクトの実施結果

2.1. 文字認識率の集計

2.1.1. 文字認識率算出結果(サマリ)

全 20,000 冊のテキスト化をスケジュール通りに完了した。またその品質確認として、認識率を算出したところ、明治期 87.7%、大正期 88.2%、昭和戦前期 92.7%、昭和戦後期 96.6%となった。認識率 70%以下の資料については、次項でその要因を検討した。

表 2.1 文字認識率算出結果(4期)

区分	対象冊数	ファイル形式	解像度	階調	認識率
明治期刊行図書	5,000 冊	JPEG2000	400dpi	2 値	87.7%
大正期刊行図書	10,000 冊	JPEG2000	350dpi	グレイ	88.2%
昭和戦前期刊行図書	4,790 冊	JPEG2000	350dpi	グレイ	92.7%
昭和戦後期刊行図書	210 冊	JPEG2000/ JPEG	400dpi	カラー	96.6%

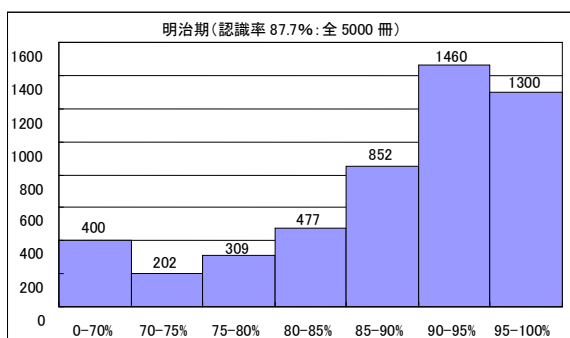


図 2.1-1 文字認識率算出結果(明治期)

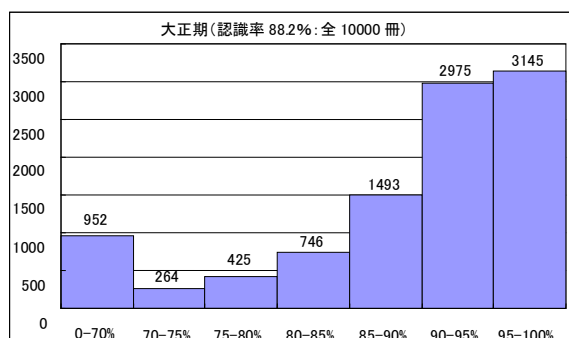


図 2.1-2 文字認識率算出結果(大正期)

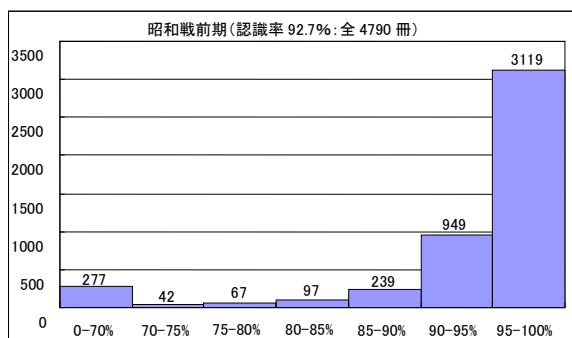


図 2.1-3 文字認識率算出結果(昭和戦前期)

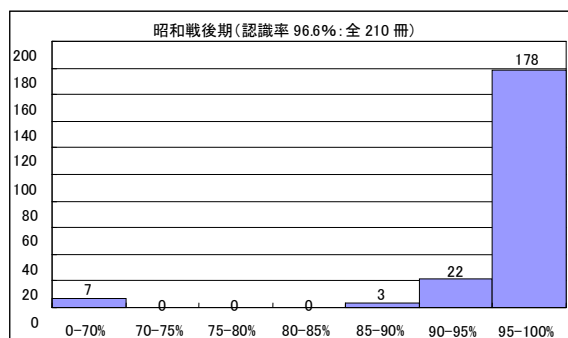


図 2.1-4 文字認識率算出結果(昭和戦後期)

2.1.2. 低認識精度(サマリ)

全 20,000 冊のうち、認識率が 70%以下であった資料 1,636 冊についてその原因を検討した。

本案件では、和装本は対象外としたが、本来対象外とすべき手書き資料 385 冊(23.5%)、図表のみの資料 219 冊(13.4%)が含まれていた(図 2.2-1、2 参照)。また、汚れ、文字の薄さやつぶれを含んだ画像品質の低い資料 500 冊(30.5%)が認識率に大きな影響を与えた(図 2.2-3、4、5 参照)。さらに、OCR の設定においてルビは認識対象外としているが、文字とルビの間隔が近い場合には、ルビを含めた文字を 1 文字として認識してしまうため、225 冊(13.8%)は認識率が低かった(図 2.2-6 参照)。

表 2.2 低認識精度内訳(4 期)

誤認識理由	明治期	大正期	昭和戦前期	昭和戦後期	計
手書き	87(21.8%)	96(10.1%)	195(70.4%)	7(100.0%)	385(23.5%)
汚れ	205(51.3%)	35(3.7%)	1(0.4%)		241(14.7%)
ルビの影響	11(2.8%)	168(17.6%)	46(16.6%)		225(13.8%)
図表	12(3.0%)	203(21.3%)	4(1.4%)		219(13.4%)
文字が薄い	11(2.8%)	143(15.0%)	13(4.7%)		167(10.2%)
文字つぶれ	30(7.5%)	58(6.1%)	4(1.4%)		92(5.6%)
カナ誤認識	5(1.3%)	59(6.2%)	1(0.4%)		65(4.0%)
レイアウト誤り	7(1.8%)	38(4.0%)	1(0.4%)		46(2.8%)
裏写り		37(3.9%)	6(2.2%)		43(2.6%)
文字が小さい	4(1.0%)	31(3.3%)	1(0.4%)		36(2.2%)
注釈線・点	9(2.3%)	17(1.8%)	4(1.4%)		30(1.8%)
画像傾き	1(0.3%)	26(2.7%)			27(1.7%)
行またがり	8(2.0%)	15(1.6%)			23(1.4%)
押印の影響	1(0.3%)	15(1.6%)	1(0.4%)		17(1.0%)
漢文	9(2.3%)	3(0.3%)			12(0.7%)
数式		7(0.7%)			7(0.4%)
英文		1(0.1%)			1(0.1%)
計	400(100%)	952(100%)	277(100%)	7(100%)	1636(100%)

2.1.3. 低認識精度(画像)

認識率が低い値となった画像例は以下の通り。

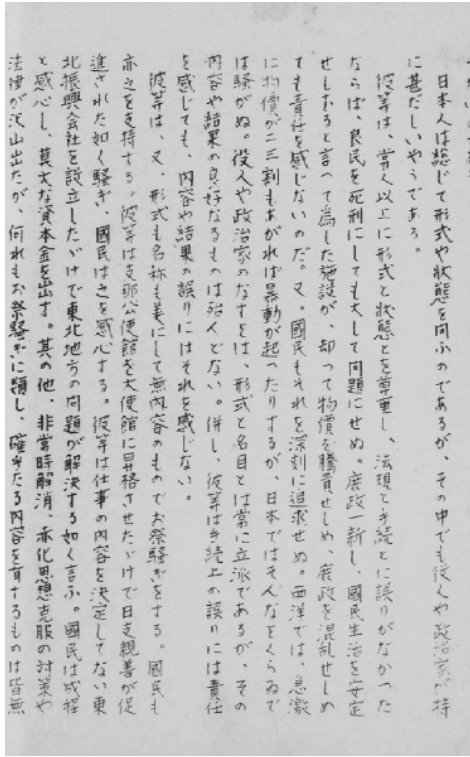


図 2.2-1 手書き(認識率 9.9%)

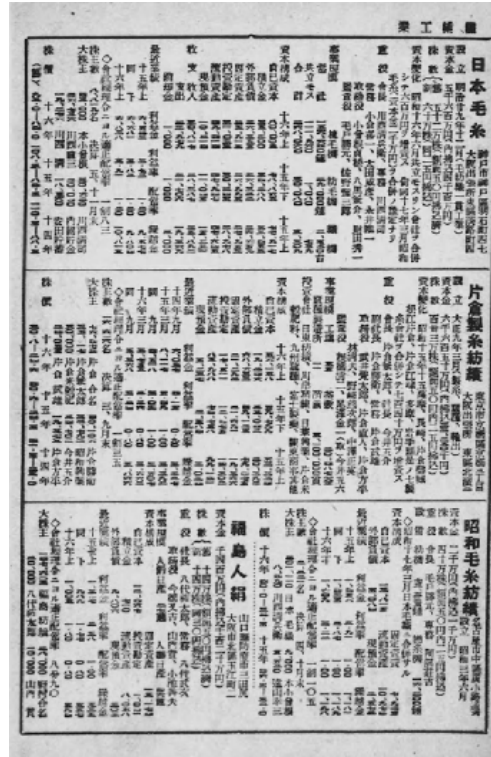


図 2.2-2 図表(認識率 60.5%)

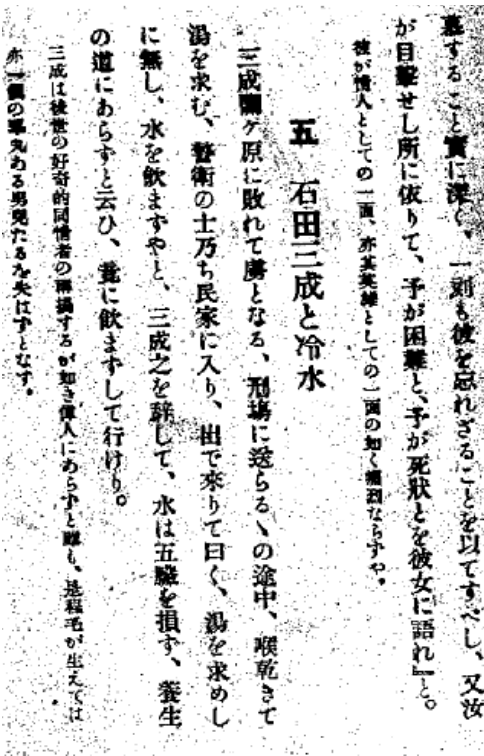


図 2.2-3 汚れ(認識率 64.2%)

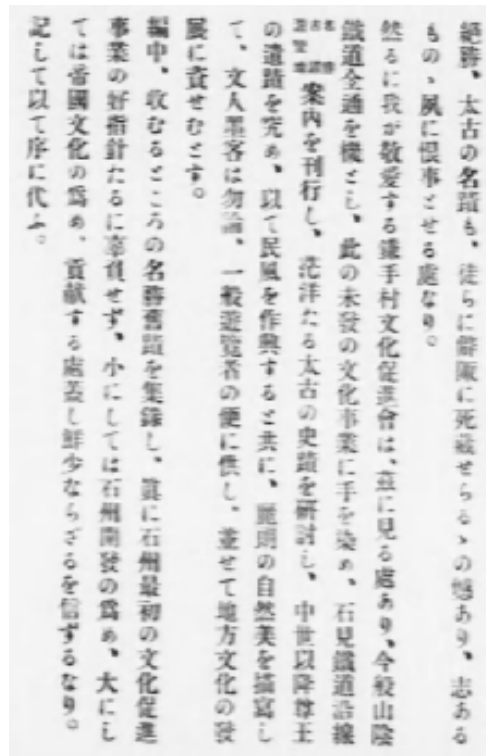


図 2.2-4 文字が薄い(認識率 40.7%)

2.2. 文字認識率のクロス集計

2.2.1. 文字認識算出結果(刊行年別)

全 20,000 冊における刊行年毎の認識率を算出した。

1800 年代は年代が古くなるほど認識率が低く、1950 年以降は横ばいという結果になった。

1940 年以前は認識率のばらつきが大きいという結果になった。

表 2.3 文字認識率算出結果(刊行年別)

刊行年	対象冊数 ⁷	認識率 ⁸
1860 年代	1 冊	62.0%(-27.7%)
1870 年代	90 冊	72.2%(-17.5%)
1880 年代	557 冊	80.0%(-9.7%)
1890 年代	1,445 冊	86.6%(-3.1%)
1900 年代	2,047 冊	89.6%(-0.1%)
1910 年代	6,137 冊	88.5%(-1.2%)
1920 年代	5,409 冊	89.2%(-0.5%)
1930 年代	2,828 冊	92.1%(+2.4%)
1940 年代	1,347 冊	94.0%(+4.3%)
1950 年代	5 冊	97.0%(+7.3%)
1960 年代	28 冊	97.7%(+8.0%)
1970 年代	25 冊	97.0%(+7.3%)
1980 年代	21 冊	96.8%(+7.1%)
1990 年代	5 冊	97.4%(+7.7%)
計	19,945 冊	89.7%

⁷ 刊行年不明(42 冊)、2000 年以降(13 冊)については、算出対象外。2000 年以降は主に手書き資料であるため

⁸ 括弧内は平均との差を示す

2.2.2. 文字認識算出結果(NDC分類別)

全 20,000 冊における NDC 分類毎の認識率を算出した。

特に認識率が高いカテゴリはなかったが、0. 総記、4. 自然科学、8. 言語については、平均と比較すると認識率は 3%以上低いという結果になった。

表 2.4 文字認識率算出結果(NDC 分類別)

NDC 分類	対象冊数 ⁹	認識率 ¹⁰
0. 総記	132 冊	85.9%(-3.7%)
1. 哲学	2,379 冊	88.9%(-0.6%)
2. 歴史	3,714 冊	87.6%(-2.0%)
3. 社会科学	8,648 冊	91.1%(+1.5%)
4. 自然科学	891 冊	83.8%(-5.8%)
5. 技術	524 冊	88.9%(-0.7%)
6. 産業	1,750 冊	89.1%(-0.5%)
7. 芸術	332 冊	86.8%(-2.8%)
8. 言語	215 冊	86.2%(-3.4%)
9. 文学	1,286 冊	90.1%(+0.5%)
計	19,871 冊	89.6%

⁹ NDC 分類不明(129 冊)については、算出対象外

¹⁰ 括弧内は平均との差を示す

2.3. 辞書更新

2.3.1. 辞書更新対象資料

300冊(×5コマ)に対して辞書登録を行った。

表 2.5 辞書更新作業対象資料(3期)

区分	対象冊数	ファイル形式	解像度	階調
明治期刊行図書	75冊	JPEG2000	400dpi	2値
大正期刊行図書	150冊	JPEG2000	350dpi	グレイ
昭和戦前期刊行図書	75冊	JPEG2000	350dpi	グレイ

2.3.2. 辞書更新対象文字

辞書登録数を増やせば単純に認識率が高くなるというわけではなく、対応する文字が多くなるため、かえって認識を誤るケースが存在する(「2.3.4. 辞書更新効果」参照)。そのため、次の文字は登録対象外とした。

- ① 句読点：、。、
- ② カッコ：()「」など
- ③ 長音：ー
- ④ 記号
- ⑤ 文字切出し誤り
- ⑥ 文字潰れ、擦れ
- ⑦ 判読不明文字、特異なフォント
- ⑧ 手書き

2.3.3. 辞書登録文字数

辞書登録文字数は41,729件(1コマ当たり27.8件)となった。

表 2.6 辞書登録文字数

区分	辞書登録数
漢字(旧字体含む)	17,406件
ひらがな	11,620件
カタカナ	12,683件
数字	20件

2.3.4. 辞書更新効果

辞書更新による効果を算出したところ、明治期+6.3%、大正期+3.1%、昭和戦前期+2.7%上昇し、大部分の資料について効果があることが判明した。ただし、3冊は効果がなく、29冊は認識率が下がる結果になった。

認識率低下の原因は、辞書登録により、かえって対応すべき文字が多くなり、誤認識が生じたと推測される。認識率が低下した資料を確認したが、特別な理由は発見できなかった。

表 2.7 辞書更新効果算出結果(3期)

区分	対象冊数 ¹¹	認識率 効果有	辞書更新前 認識率	辞書更新後 認識率
明治期刊行図書	67冊	65冊(97.0%)	84.5%	90.8% (+6.3%)
大正期刊行図書	130冊	108冊(83.1%)	87.9%	91.0% (+3.1%)
昭和戦前期刊行図書	66冊	58冊(87.9%)	93.0%	95.7% (+2.7%)

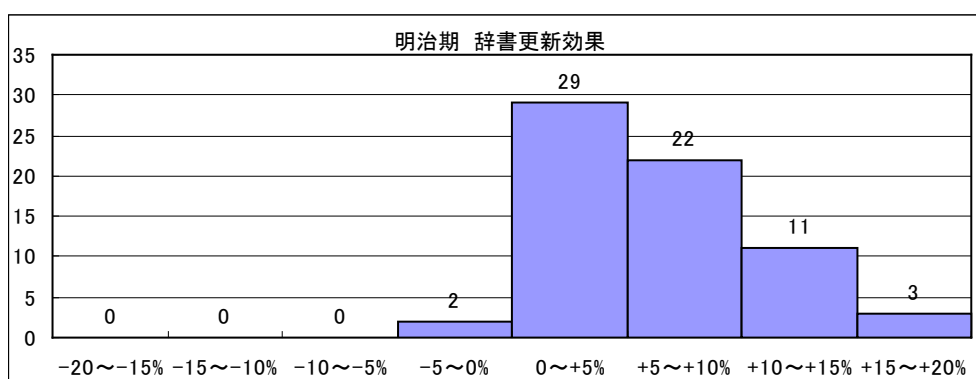


図 2.3-1 辞書更新効果(明治期)

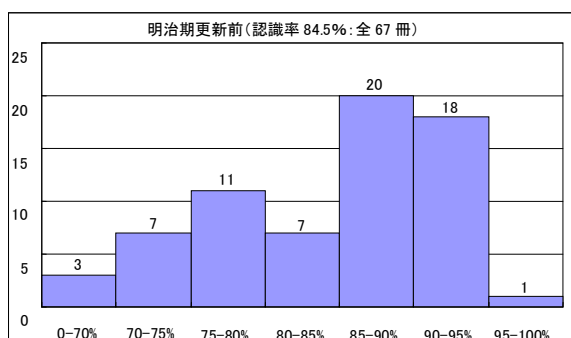


図 2.3-2 辞書更新前文字認識率算出結果(明治期)

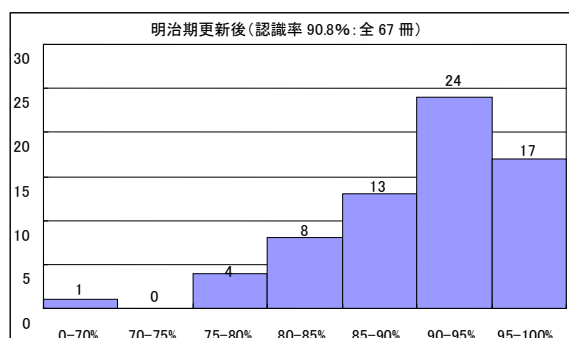


図 2.3-3 辞書更新後文字認識率算出結果(明治期)

¹¹ 算出に当たり 300冊中 37冊は対象外

- ①手書き資料：10件(明治期1件、大正期1件、昭和戦前期8件)
- ②認識率差20%以上：20件(明治期5件、大正期14件、昭和戦前期1件)
→レイアウト漏れ等辞書更新以外による要因が大きいため
- ③認識率60%以下：7件(明治期2件、大正期5件、昭和戦前期0件)
→画像が薄い、図表等辞書更新による効果が現れにくいため

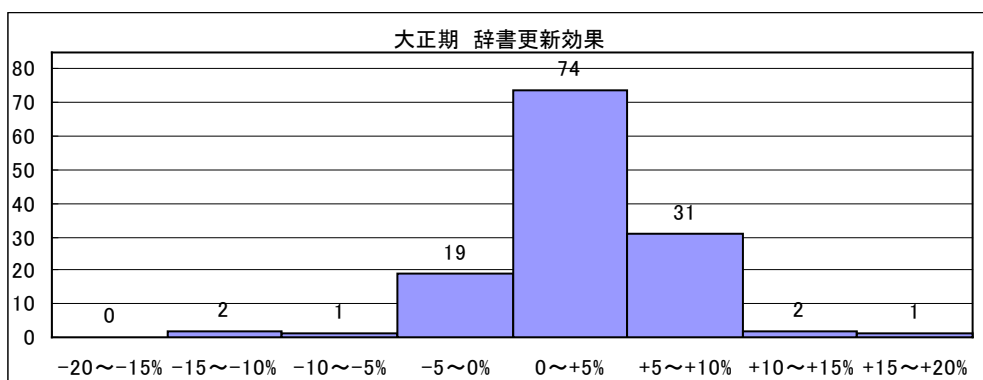


図 2.4-1 辞書更新効果(大正期)

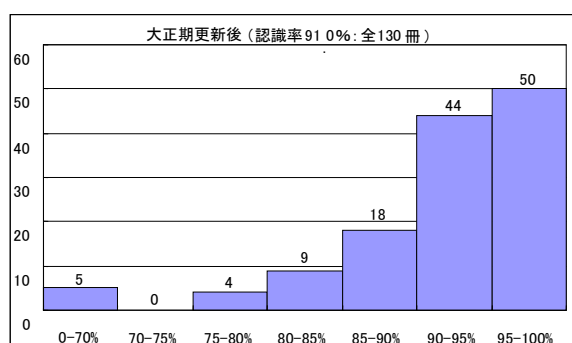
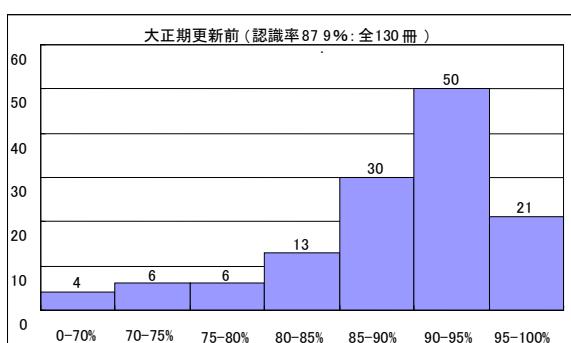


図 2.4-2 辞書更新前文字認識率算出結果(大正期)

図 2.4-3 辞書更新後文字認識率算出結果(大正期)

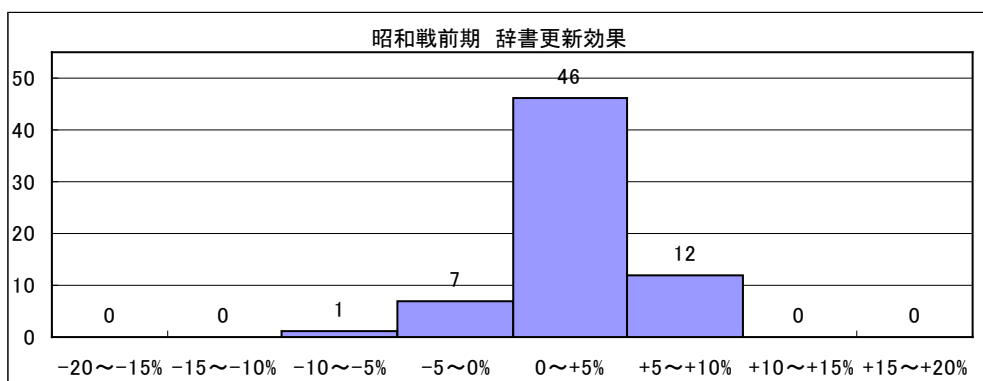


図 2.5-1 辞書更新効果(昭和戦前期)

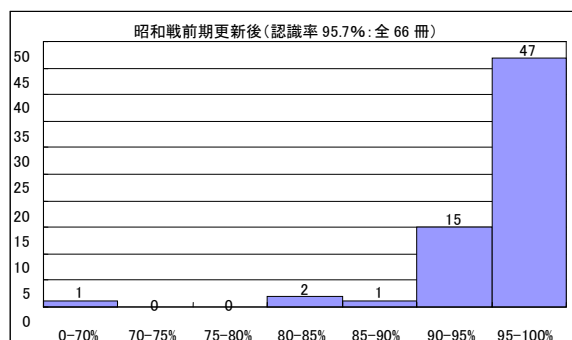
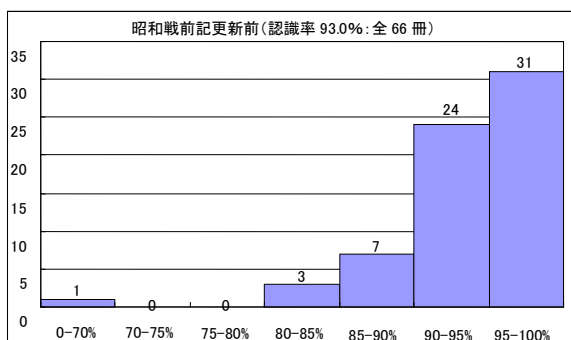


図 2.5-2 辞書更新前文字認識率算出結果(昭和戦前期)

図 2.5-3 辞書更新後文字認識率算出結果(昭和戦前期)

3. 課題

3.1. プロジェクト実施工数

3.1.1. 本案件での課題

表 3.1 の (2) (5) 認識率算出、(3) 校正作業は、人手作業であるため、それぞれ 1 コマ当たり約 14 分、約 69 分の時間を要した。

(3) 校正作業は、下記のステップで同作業経験者が 1 人 1 コマを担当して実施した。旧字は、旧字及び新字のどちらでテキスト化されていても問題ないこととしたが、正しく新字でテキスト化されているかの確認は有識者が行う必要があり、旧字新字の切り分け作業に時間を要した。校正作業時間の内訳は次のとおり。

- ① 約 35 分：画像とテキストを比較し、校正作業実施
- ② 約 30 分：誤読増加に繋がる文字パターンの削除作業、文字パターンの正解チェック(校正の再チェック)
- ③ 約 4 分：文字パターン単位でのノイズ除去作業

表 3.1 プロジェクト実施工数表

工程	作業数量		工数※		作業 日数	稼働 時間	人員	PC	夜間 自動認識
(1)全文テキストデータの作製	600冊	90,000コマ	33.0分/書誌	0.22分/コマ	1日	3時/人日	2人	33台	10時間
(2)認識率算出(初回)	300冊	300コマ	14.0分/書誌	14.0分/コマ	7日	5時/人日	2人	2台	-
(3)校正作業	300冊	1,500コマ	346分/書誌	69.1分/コマ	9日	6時/人日	32人	32台	-
(3)OCR辞書の更新	300冊	1,500コマ	22.0分/書誌	4.40分/コマ	11日	5時/人日	2人	2台	-
(4)更新済みOCR辞書を用いた全文テキストデータの作製	20,000冊	3,000,000コマ	32.8分/書誌	0.22分/コマ	21日	3時/人日	2人	26台	20時間
(5)認識率算出(最終)	20,000冊	20,000コマ	13.2分/書誌	13.2分/コマ	23日	6時/人日	32人	32台	-

※工数：(1) (4) PC 処理時間、(2) (3) (5) 人員作業時間

3.1.2. 今後の検討課題

(2) (5) 認識率算出には 1 コマ当たり約 14 分を要するため、大量の OCR 化における実施は難しい。そのため、ある適度システムの予測が可能かの検証が必要である。

本案件で使用した Express Reader Pro では、登録されている辞書の中からテキスト化対象文字に最も似た文字を抽出する際に、独自のロジックで点数化を行い、その点数が最も高いものを表示するという方法を取っている。標準機能では、その点数を表示させることはできないが、内部的にログを取得しており、抽出は可能である。そのため、その点数の平均、バラつきと、人手で実施して算出した正しい認識率との相関関係が見つければ、ある適度妥当な認識率を自動的に算出可能になると考えられる（「図 3.1 相関関係比較イメージ図」参照）。

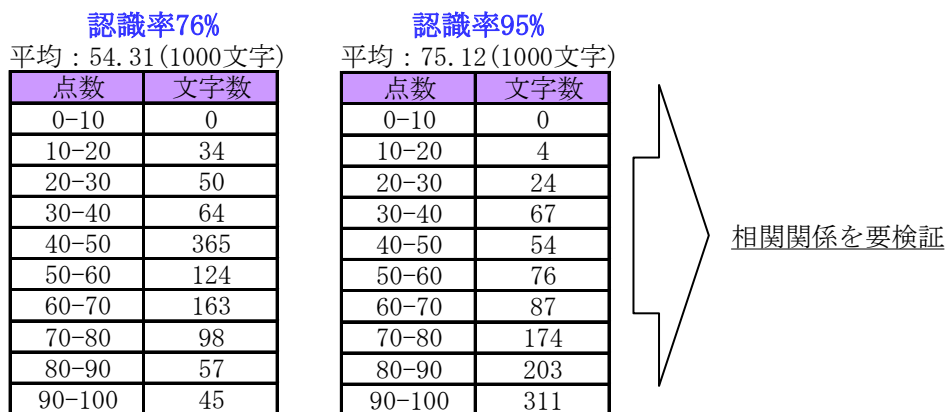


図 3.1 相関関係比較イメージ図

(3)校正作業を人手により実施する場合は、1 コマ当たり約 69 分を要するが、認識率が低い資料ほど校正作業の工数が増大する。画像をもとに 1 からタイピングする方が早いことも考えられる。

3.2. 画像品質向上

3.2.1. 本案件の課題

本案件の元画像は、主に明治期、大正期という古い資料を対象としたため、文字が極端に薄い又は下地が濃い画像、クリームがかかった白地の画像が多く、OCR 処理に誤認識が多く発生した(「2.1.2. 低認識精度(サマリ)」参照)。

3.2.2. 今後の検討課題

文字が極端に薄い又は濃い画像、クリームがかかった白地の画像については、OCR 処理前に 2 値化することで、文字と白地を鮮明に区分可能になる。それにより OCR 認識率の向上が見込まれるため、検証が必要である。

3.3. 縦横文書

3.3.1. 本案件での課題

本案件では、縦書き横書きの判断をフィールド¹²毎に自動的に実施したが、誤認識した箇所が発生した。縦横の誤認識が発生すると、そのフィールド内の認識率は 0%になってしまうため、影響は大きい。

縦横の判断は、基本的には、縦横の文字の揃い具合や間隔により行っているが、形態素解析機能¹³により意味が通るかどうかによっても判定している。

元画像が傾いていると文字間隔を正確に認識できず、誤認識が発生する。また、OCR に搭載されている形態素解析は現代文専用であるため、古い資料では、通常ひらがなで表現する部分がカタカナで表現されている、助詞の使い方が現代文と異なる、単語自体が現代と異なるなどの理由により形態素解析が正常に機能せず、誤認識が発生する。

¹² OCR 読取時に、元画像上で文字の固まり毎で分離するレイアウト解析が自動実行されるが、それによって分離された部分のこと

¹³ 形態素区切り機能と品詞付与機能から成り、日本語として成立しているかを読取る機能のこと

3.3.2. 今後の検討課題

誤認識の原因が画像の傾きである部分は、左右ページ別々にスキュー補正¹⁴を行うことで、改善に努めた。ただし、その効果と誤認識件数は未算出であるため、今後検討が必要である。

形態素解析が正常に機能しない部分は、時代に則した文法を加味した解析機能を採用することで対応可能と考えられる。しかし、作業実施者によると、技術的には可能であるが、汎用性やニーズが少なく、商品化は難しいとのことであった。

3.4. ノイズの除去

3.4.1. 今後の検討課題

本案件における文字認識率は、「正しく認識された文字数÷元画像上の文字数」で算出しており、ノイズ（文字ではない部分を文字として認識してしまったもの）については考慮していない。そのため、限りなく100%に近い文字認識率であっても、ノイズが多ければ、読上げソフトや検索システムでは使えない可能性がある。対応策としては、事前の画像処理等によるノイズ除去が考えられる。

全体では1コマ当たり平均93文字のノイズが検出され、1910年以前(明治期)の資料は、ノイズが多いという結果になった。

認識率が高いものについては、比較的ノイズも少ない傾向が見られ、認識率97.5%以上の資料のノイズは平均47文字で、他と比較して少ない結果になった。

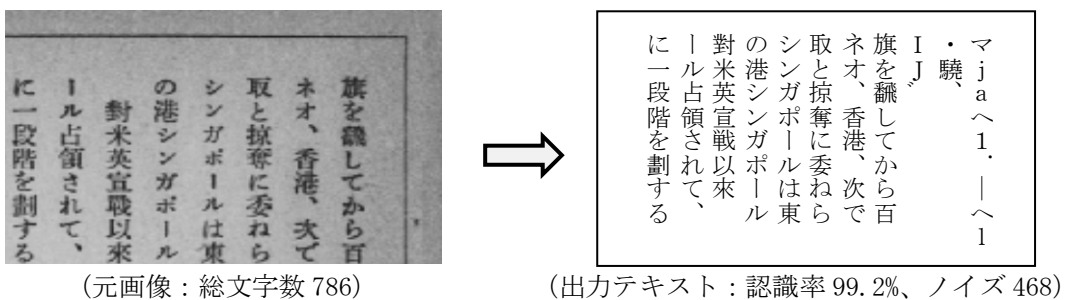


図 3.2 ノイズの例

¹⁴ 画像の傾きを補正する機能のこと

表 3.2 ノイズ数算出結果(刊行年別)

刊行年	対象冊数 ¹⁵	ノイズ ¹⁶	ノイズ(標準偏差)
1860年代	1冊	42文字(-51)	0
1870年代	89冊	106文字(+13)	93.7
1880年代	552冊	141文字(+48)	145.0
1890年代	1,426冊	123文字(+30)	150.6
1900年代	2,010冊	131文字(+38)	168.8
1910年代	6,002冊	99文字(+6)	226.6
1920年代	5,195冊	81文字(-12)	233.7
1930年代	2,755冊	59文字(-34)	147.7
1940年代	1,326冊	74文字(-19)	154.3
1950年代	5冊	27文字(-66)	20.0
1960年代	28冊	49文字(-44)	63.7
1970年代	25冊	32文字(-61)	33.9
1980年代	21冊	42文字(-51)	34.3
1990年代	5冊	58文字(-35)	55.0
計	19,440冊	93文字	202.2

表 3.3 ノイズ数算出結果(認識率別)

認識率	対象冊数 ¹⁷	ノイズ ¹⁸	ノイズ(標準偏差)
0-10%	137冊	125文字(+32)	359.7
10-20%	73冊	230文字(+137)	231.8
20-30%	102冊	325文字(+232)	547.7
30-40%	188冊	191文字(+98)	298.7
40-50%	177冊	190文字(+97)	390.5
50-60%	279冊	195文字(+102)	380.6
60-70%	524冊	230文字(+137)	385.2
70-80%	1,217冊	184文字(+91)	282.7
80-90%	3,782冊	120文字(+27)	193.0
90-92.5%	2,166冊	87文字(-6)	184.8
92.5-95%	3,155冊	71文字(-22)	136.0
95-97.5%	4,401冊	57文字(-36)	141.0
97.5-100%	3,293冊	47文字(-46)	150.5
計	19,494冊	93文字	202.2

¹⁵ 刊行年不明(42冊)、2000年以降(13冊)、テキストの文字数が実際の画像文字数より少ない資料(506冊)については算出対象外。2000年以降は主に手書き資料であるため

¹⁶ 括弧内は平均との差を示す

¹⁷ テキストの文字数が実際の画像文字数より少ない資料(506冊)については算出対象外

¹⁸ 括弧内は平均との差を示す

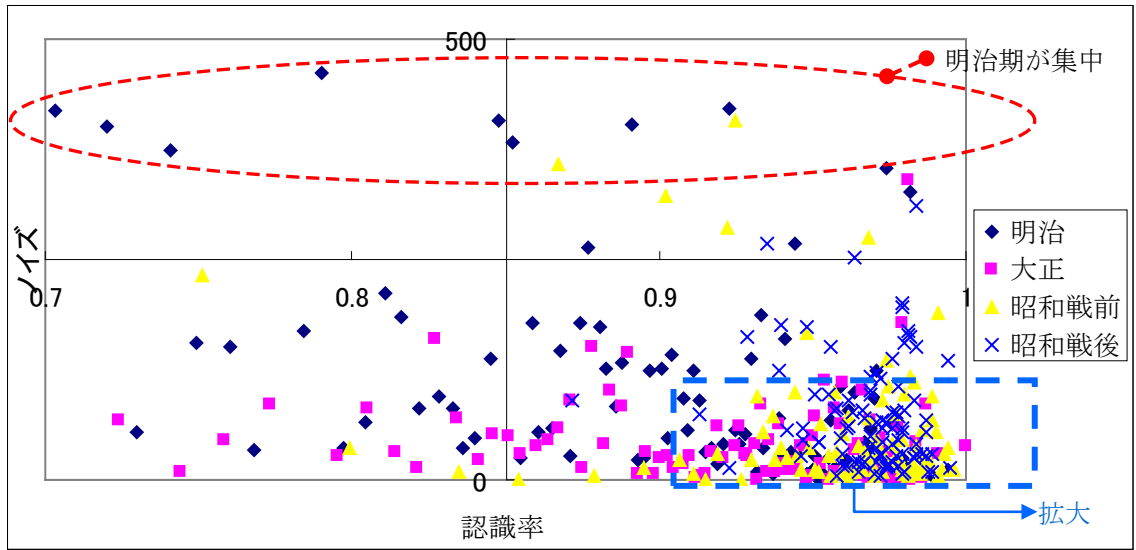


図 3.3 認識率・ノイズ相関関係(全体)

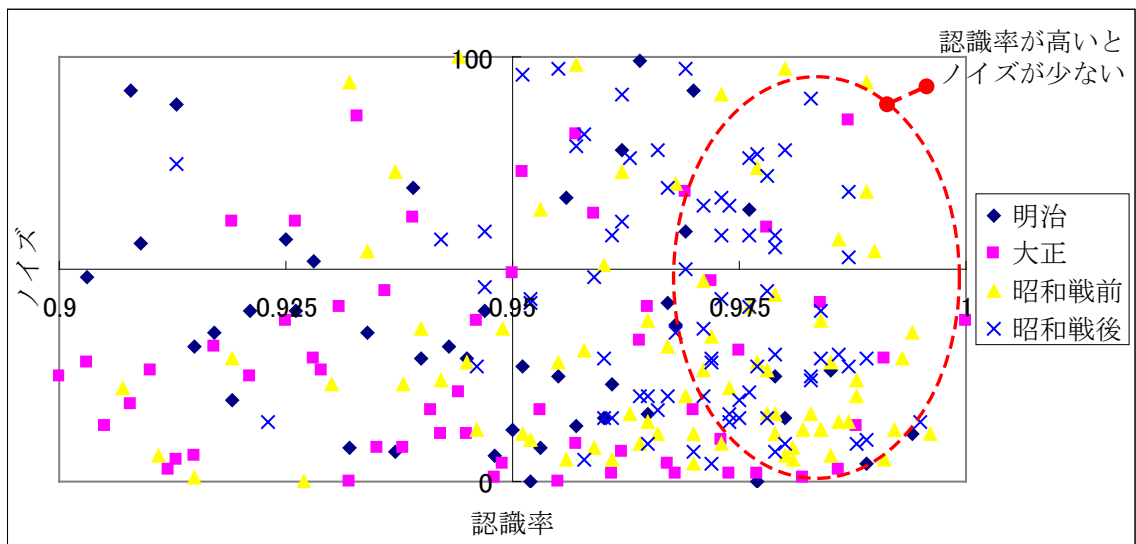


図 3.4 認識率・ノイズ相関関係(拡大)

付録

1. OCR作業に関する設定値

本案件で使用した設定値を示す。

表 付.1 OCR 設定値

No.	項目	選択肢/本案件◎	備考
1	エンコード方式	◎UTF-8 UTF-16 UTF-16(逆ビット)	英文の場合、iso8859-1と互換性が高く、通信用の制御コードを使用しておらず扱いやすいため、UTF-8が一般的に使用される。
2	BOM	有り ◎無し	エンコード方式を区別するための記号で、有っても無くても、Windowsソフトで開けることが多い。有るとUnixの多くのソフトで文字化けが発生。
3	改行コード	CR(Mac) ◎LF(Unix) CR+LF(Windows)	CR+LFが最も普及している。
4	改行コード出力位置	出力しない ブロック単位 ◎原文の全行	全行に出力した方が、テキストエディタで開いた際、原文に最も近い結果になる。
5	スキュー補正	◎有り 無し	スキュー補正とは、OCR技術により原稿の傾きを補正する技術で、認識率の向上に役立つ。ただし、デメリットとして、喉元部分の文字切れ、また角に白い余白が生じる。
6	画像出力形式	◎CCITTFaxDecode DCTDecode ◎JPXDecode	CCITTFaxDecodeはCCITTG4形式と同様で、Tif形式、モノクロ2値に対応。 DCTDecodeはJPEG形式に対応。 JPXDecodeはJPEG2000形式に対応。
7	文字コード	◎UTF JIS シフトJIS EUC	HTMLでは4つの形式のいずれも使われる。Windows では従来はシフトJISが主流であったが現在ではUTFも利用できる。UTFは国際化の面で有利であるため、利用が広がっている。
8	ルビ	◎出力しない 出力する	基本認識処理に実装。
9	出力画像品質	1～100 ⇒◎設定値(50)	画像容量を押さえるため、標準設定値(90)から(50)に変更。
10	出力ファイル形式	◎透明テキストつきPDF ◎プレーンテキスト HTML XML CSV RTF	仕様書の通り。
11	ノイズ(ごみ)除去	◎有り 無し	基本認識処理(自動処理)に実装。
12	対応文字言語	◎日本語 英語	日本語文章処理で対応。(認識文字としては、アルファベット認識が可能) 今回インストールしたモジュール構成では他言語の認識モジュールは含まれていないが、他にもドイツ語・フランス語・簡単字中国語、繁体字中国語など19ヶ国語の認識モジュールも用意している。
13	改ページコード	有り ◎無し	テキストファイル出力時に改ページコードを出力すると、ページの切れ目を保存することができるが、今回は出力していない。
14	出力単位	◎ページ単位 1冊単位	最終的なPDFファイルは、1ページPDFファイルとなるように設定。
15	行頭空白	◎出力しない 1文字 複数文字	行頭に空白文字を出力すると、テキストエディタなどのソフトウェアで閲覧する際に原文に近いレイアウトで見ることができ、一方でコピー・貼り付け・読み上げなどの再利用の際に妨げとなる。
16	文書内の空白	◎出力しない 1文字 複数文字	文書内に空白文字を出力すると、テキストエディタなどのソフトウェアで閲覧する際に原文に近いレイアウトで見ることができ、一方でコピー・貼り付け・読み上げなどの再利用の際に妨げとなる。
17	認識対象文字種	JIS第1水準(約4000字) ◎JIS第1,第2水準(約7000字)	第2水準の文字を不用意に認識対象として加えると、通常の文書であれば出現頻度が低いので認識精度が落ちる。しかし、本案件では旧字体文書が主な認識対象であるため、出現頻度が極めて高く精度向上に役立つ。
18	認識モード	高速モード ◎高精度モード	第2水準の文字群には極めて類似した文字が含まれているため、認識対象とするためには今回使用したOCRソフトでは、やや低速な処理ながら計算量の多い、高精度モードを利用する必要がある。
19	行判定	◎自動判定 横書き 縦書き	横書き、縦書きを自動判定するモードを使用。
20	全角・半角	全角・半角を区別 ◎全て全角 英数記号のみ半角	区別して出力した場合、全角・半角は原稿の状態から自動判別され、原文レイアウトに近い形となる効果もあるが、半角カナが出力されシステムによっては処理できない場合もある。全てを全角で出力した場合、日本語文書ではバランス良く表示されるが、英文の再利用性が劣る。英数記号のみ半角で出力した場合、日本語文書は表示上のバランスが悪くなる場合があるが、英文の再利用性が高まる。

※ドキュメントリーダー Express Reader Pro 東芝ソリューション(株)製

2. 文字認識率算出方法

文字認識率の算出は、以下のとおり行った。

2.1 作業対象

各分冊の本文部分の先頭から1コマを対象とする。ただし、挿絵、図(表)、計算式、落書き、画像の品質が悪い、漢文(レ点、返り点付)、手書きが含まれるコマは、文字認識率算出の対象外とし、次のコマを対象とする。

2.2 手順

- (1)「元画像上の文字数」を数える。
- (2)OCR 処理により作製したテキストデータと元画像を比較し、「正しく認識された文字数」を数える。
- (3)上記(1) (2)で算出された数字をもとに文字認識率(※)を計算する。

$$\text{※文字認識率(\%)} = (\text{正しく認識された文字数} / \text{元画像上の文字数}) \times 100$$

2.3 算出上のルール

- (1)以下を数える。
 - ① 元画像の本文部分に含まれる文字並びに句読点及び括弧等の記号
- (2)以下は数えない。
 - ① 本文ではなく、図、挿絵、表等に含まれる文字
 - ② 計算式
 - ③ 漢文のレ点、返り点
 - ④ ノンブルや、ページ肩部分にある書名・章名など、本文以外の部分にある文字
 - ⑤ ルビ
 - ⑥ 元画像上のゴミ等、「文字でないもの」を間違っ文字として認識している場合
- (3)その他
 - ① レイアウト、階層、改行等、構造上の間違いがあっても、元画像上の文字自体が正しくテキスト化されていれば、正しく認識された文字として数える。
 - ② 元画像上の旧字がテキストデータ上では新字として認識されている場合は、正しく認識されたものとして数える。
 - ③ 文字自体は正しく認識されているが文字の順番が元画像と異なる場合は、元画像の順番のとおり認識されている部分のみを、正しく認識された文字として数える。
 - ④ 最終的な認識率については、小数点第2位を四捨五入し、小数点第1位までを算出する。

2.4 文字認識率の算出例

下記【テキストデータ】のうち、灰色の網かけにした文字が「正しく認識された文字」である。

この場合、

- (1)元画像上の文字数=237 文字
- (2)正しく認識された文字数=197 文字

であるため、文字認識率は、 $(197/237) \times 100 = 83.1(\%)$ となる。

【元画像】

動物物の凡らゆる遺跡である。
岩石中に保存された生物の遺跡が果して古生物なるや否やを決すべき目安は其の岩石の成立した時代で、遺跡が石に化して居るや否や、又は現世界に産する種類と同一なるや否やではない。勿論古生物が岩石中に埋没した後に大抵多少の變化を受けて、鑛物になつて居ることは事實である。しかし、稀有の事とは云へ、前世界産の象や犀の屍で其の儘西伯利亞の凍結地盤中から出た例もあり、又同じ前世界の産である琥珀中に、植物、昆蟲、蜘蛛等の、些の變化を受けずに保存してゐる例もある。

【テキストデータ】

鵜棲槍の凡らゆる遺跡である。
倉石中に保存された佳物の遺跡が果して古生物なるや否やを決すべき目安は真の岩石の成立した時代で、遺跡が石に化して居るや否や女は現世界に慶する樺鯛と同一なるや否やではない。勿論古生物が岩石中に埋没した後に大概多少の美化を受けて、義物になって居ることは奉賛である。しかし、稀有の事とは云へ、前世界塵の象や亀の屍で其の儘西伯利亞の確繡地盤中から出た例もあり女同じ前世界の塵である境鵬中に痛勒良牟蜘蛛等の、些の美化を受けや1こ保存してゐる例
▲○あるも

図 付.1 文字認識率の算出例

3. データ容量

本案件で定めている納品物は、透明テキストつきPDF¹⁹及びプレーンテキストファイルの2点で、1コマ当たり1ファイルとして作成した。刊行時期ごとの容量は以下のとおり。

表 付.2 データ容量

区分	対象冊数	PDF (計)	TEXT (計)	PDF (1冊当たり)	TEXT (1冊当たり)
明治期刊行図書	5,000冊	70.4GB	1.6GB	14.1MB	0.3MB
大正期刊行図書	10,000冊	150.5GB	2.9GB	15.0MB	0.3MB
昭和戦前期刊行図書	4,790冊	221.4GB	2.2GB	46.2MB	0.5MB
昭和戦後期刊行図書	210冊	8.9GB	0.3GB	42.4MB	1.4MB

¹⁹ PDFファイルの一種で、紙資料をスキャンしたデジタル化画像と、OCRソフトで作製したデータとを1つのファイルとしてまとめたファイル形式