立命館人間科学研究，26，99–110，2013.

実践報告（Practical Research）

未校正書籍テキストデータの読書アクセシビリティ

――大学図書館における読書障害学生支援に向けて――

松　原　洋　子・植　村　　要

（立命館大学大学院先端総合学術研究科）

Immediate Access to Digital Text Files Converted from Printed Books for Students with Print Disabilities at University Libraries

MATSUBARA Yoko and UEMURA Kaname

（Graduate School of Core Ethics and Frontier Sciences, Ritsumeikan University）

In this study, we conducted an experiment to examine issues that arise when a university library provides unedited text data for users with disabilities, such as impaired vision, to sample books. The experiment compared book layouts, the method and device used to scan the book, and usage of image correction, in addition to measuring and comparing the time it took to prepare the text data. The quality of data created by each method was further evaluated visually as well as by using a screen reader. As a result, we found that all methods generated legible text data, although the quality varied. We also found that it was quicker to scan copied pages than the original book when creating data. Books with simple layouts can be quickly provided in a data format, without correcting the images. Images of books with complicated layouts that include figures should be corrected by removing the figures; however, in that case, information on the removed parts should be provided to the user. Furthermore, accurate text data should be prepared at least for the table of contents because the main text would include incorrectly recognized words.

Key Words : students with print disability, reading accessibility, university library, digital text files without proofreading

キーワード：読書障害をもつ学生，読書アクセシビリティ，大学図書館，未校正テキストデータ

1. はじめに

2010 年 1 月に改正著作権法が施行され，第 37 条第 3 項の規定により「視覚障害者その他視覚による表現の認識に障害のある者」を対象に，大学図書館においても書籍等の著作物を著作権者の許諾を経ずに利用者に必要な方式で複製し，提供できるようになった。特にテキストデータは，ソフトフェアで音声や点字に自動変換して読むことができ，視覚による読書に障害をもつ学生にとっては大変有効である。この著作権法改正を受けて，立命館大学図書館では，2010 年 7 月から所蔵資料のテキストデータ提供を試行的に開始した。提供の対象となる所蔵資料は，主に貸し出し可能とされている図書ならびに論

文雑誌である。開始当初は障害学生支援室と連携していたが，2011年1月からは図書館で本格的な運用を開始し，図書館の中にテキストデータ作成スペースと作業担当者を確保してテキストデータ作成作業を行い，CD-ROMにデータを格納して提供している（2012年11月1日現在）。

しかし，すでに指摘されているように，書籍等の印刷物の内容を正確にテキストデータ化するには多くの人手と時間を要する（立命館大学障害学生支援室，2010；植村・山口・櫻井・鹿島，2010）。印刷物をスキャンしてOCR化したテキストデータは，誤認識による文字化け等が発生するため校正が必要であるが，この作業にテキストデータ化に伴う労力の大半が費やされ時間も要する。立命館大学図書館では，当初，利用者からテキストデータ化を請求されたすべての資料について，丁寧に校正を行いデータの貸し出しをしてきた。しかし，条件によっては校正済みのテキストデータ貸し出しまでに時間がかかり，レポートや論文の執筆など勉学や研究の目的で利用する学生らのニーズに必ずしも応えられない場合もあった。

そこで立命館大学図書館では2012年4月から，より迅速に対応するため未校正のデータも希望があれば提供するようになった[1)]。校正されたテキストデータが利用できるほうが望ましいことは確かだが，人手や財源に限りがある状況においては，それを常に迅速に実現することは難しい。有限な資源を効率的に利用して，正確さと迅速さを求める利用者のニーズを充足するためには，精読に先立つ試し読みの段階で未校正データを活用することも考えられる。

勉学や研究上，精読すべき資料か，そうでないかを決めるには，複数の資料にあたって試し読みをする必要がある。視覚を利用して読書する者は，目次，文中のキーワード，文章の一部，索引等を点検して資料の概要を理解し，その資料を精読する必要があるか否かを判断する。

一方，視覚障害者は，印刷された資料のままでは精読以前に試し読みもできない。長時間かけて校正された資料を入手しても，精読を必要としない本であることが読んではじめてわかる場合もある。未校正テキストデータを試し読みに活用できれば，利用者にとっても図書館にとっても無駄な時間と労力をかけることなく，図書を有効に利用できるようになる。精読が必要と判断されれば，未校正テキストデータを校正した資料の貸し出しを改めて申請すればよい。

すでに米国では，ブックシェア（Bookshare）というNPOのオンライン図書館が，図書のデジタル複製データをエラーの数が少ない順に優（Excellent），良（Good），可（Fair）の3ランクに分け，未校正データも含めて視覚障害者等に提供してきた実績がある（国立国会図書館，2003）。また国立国会図書館でも，所蔵資料のデジタル化画像データの全文テキスト化実証実験の一環として，視覚障害者等を対象に，OCR認識率が70%，90%，93%，98%の書籍の自動音声読み上げによる理解度の調査を実施している。さらに「読上げにおいて正確に理解できるためには，98%以上のOCR認識率が必要である」としながらも，「OCR認識率が低い図書についても，書籍の正確な内容は不明でも，種類程度は判断できる可能性が高いことから，公開されることが望ましいという意見があげられた」と報告している（国立国会図書館，2011b）。OCRソフトの認識率に限界があるなかで，比較的人手をかけずに速やかに作成できる未校正テキスト

1）立命館大学障害学生支援室では，視覚障害学生の授業支援として未校正データも含め，書籍のテキストデータ提供を行なってきた。しかし，2010年7月以降は著作権法改正を受けて，図書館蔵書の貸し出しに関しては，障害学生支援室ではなく図書館がテキストデータ提供の主体となっている。本研究では，2010年7月以降の図書館主体のテキストデータ提供サービスに焦点を絞っている。立命館大学の「障害学生サービス」については，下記を参照のこと。
www.ritsumei.ac.jp/acd/mr/lib/shogaiservice.html（2013年1月28日アクセス）。

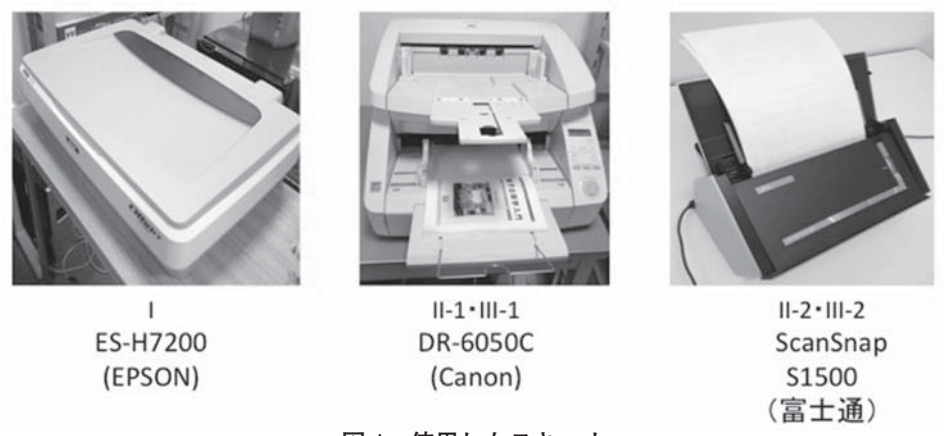

図１　使用したスキャナ

データの活用は，このように国立国会図書館の障害者サービスとしても注目されつつある。

そこで本研究では，未校正のテキストデータを試し読みに利用する場合，データ化作業においてどのような課題があるかを明らかにするために，実証実験を行った。実験では大学図書館で調達可能な装置を使用して，レイアウトが異なる２種類の書籍のテキストデータを作成した。そして，装置，書籍，データ作成方法の種類別に，作業に要した時間と未校正テキストデータの質を評価した。最後に，学習・研究のために大学図書館で提供する書籍として，試し読み用に未校正テキストデータを提供する際に，配慮すべき対応について考察した。

2. 方法

2－1　装置

本実験では図書館蔵書のデータ提供を前提としているため，本を裁断せずに原本をスキャンする方法（以下，I）とコピーをスキャンする方法（以下，II）を採用した。それぞれの方法でレイアウトの特徴が異なる２種類の書籍をテキストデータ化した。

Iでは，手作業で原本を見開きの状態にして直接スキャンした。スキャナは ES-H7200（EPSON），OCR ソフトは WinReader PRO ver.12（メディアドライブ）を使用した。

IIでは，原本を見開きの状態にしてコピー機で紙のコピーを作成し，自動給紙機能を使ってコピーをスキャンした。スキャナと OCR ソフトの組み合わせとしては，高価格の DR-6050C（Canon）+ WinReader PRO ver.12（II-1）と，比較的低価格で個人にも普及している ScanSnap S1500（富士通）+ 読取革命 ver.14（Panasonic）（II-2）の二通りで実験した[2)]。

比較対照として，本を裁断して自動給紙によりスキャンする方法（以下，III）でも同様の実験を行なった。スキャナと OCR ソフトは，III-1 は II-1 と，III-2 は II-2 と同じ仕様にした（図 1）。

2）装置の価格（税別）は，以下の通り。
I：スキャナ ES-H7200（85,524 円），OCR ソフト WinReader PRO（198,000 円，ただし ver.13 の価格），合計価格 283,524 円。
II-1：スキャナ DR-6050C（840,000 円），OCR ソフト WinReader PRO（198,000 円，ただし ver.13 の価格），合計価格 1,038,000 円。
II-2：スキャナ ScanSnap S1500（47,429 円），OCR ソフト読取革命（ver.14，希望小売価格税別で 12,190 円），合計価格 59,619 円。価格は以下で確認した（いずれも 2012 年 11 月 12 日アクセス）。
・ES-H7200　http://rcpt.kyoto-bauc.or.jp/guide/labo/index.html?mode=detail&product_id=3833
・DR-6050C　http://cweb.canon.jp/imageformula/lineup/dr-6050c/index.html
・WinReader PRO ver.13　http://mediadrive.jp/products/wrp/index4.html
・ScanSnap S1500　http://www.pfu.fujitsu.com/direct/scanner/detail_s1500a.html
・読取革命 ver.14　http://panasonic.co.jp/snc/pstc/products/yomikaku/shopping.html

使用したPCはすべてPC-MJ18XAZR8（NEC, OS:Windows 7）であった。また未校正テキストデータの自動音声による試し読みには，VDMW700-PC-Talker ver. 2.11（高知システム開発）を使用した。

2－2　使用した書籍

データ作成にあたっては，レイアウトの特徴が異なる以下の2種類の書籍を使用した。著作権者には本実験の目的を説明し，これらの書籍のテキストデータを作成し研究に利用することについて許諾を得た。

1. 中村正『家族のゆくえ――新しい家族社会学』（以下，書籍A）

【書誌情報】人文書院，1998年，四六判（18.8×13.4×2.4 cm），ハードカバー。

【構成とレイアウトの特徴】

・目次：章と節番号がシンプルな飾り数字。節の下位項目がスラッシュで区切られ列記されている。計4頁。

・「はじめに」，第1～3章，「おわりに」：縦書き一段組で，引用文献が本文中〔　〕内に横書きで，英文が4ヶ所，和文が35ヶ所挿入されている。図表や注はない。計212頁。

・「コラム」：2章と3章の間（200～224頁）に挿入されている。縦書き二段組で，見開きに2つのコラムが上下段に分けて掲載されており，一つのコラムがそれぞれ2頁に渡っている。上下段の間に四角の囲みで横書きのタイトルが挿入されており，上段のコラムのタイトルが前頁，下段のコラムのタイトルが次頁に掲載されるやや変則的なレイアウトになっている（図2）。計25頁。

・「マンガ書評」：58頁に「マンガ書評」という雑誌記事が画像として再掲されている。縦書き四段組の頁レイアウトで文字はかなり小さく，写真やイラストを含む。計1頁。

・参考文献：横書き1段組で参考文献の書誌情報が列挙されている。計4頁。

・その他:扉，標題紙，各章表紙と奥付。計6頁。

以上，スキャンした印刷部分は合計252頁。さらに白ページが2頁あった。見開きでは128面（コピーで128枚）となった。

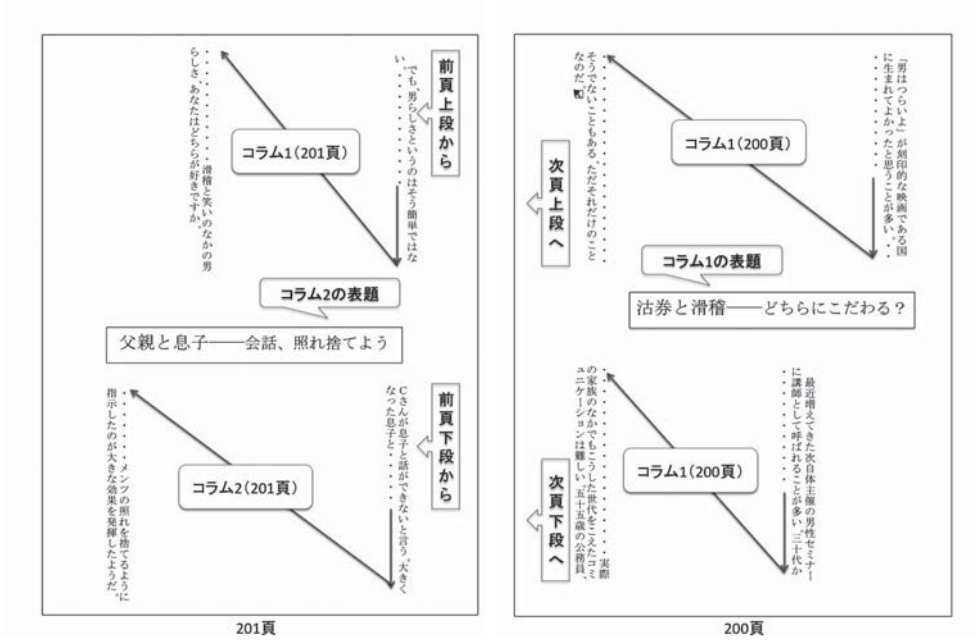

図2　書籍A「コラム」の段組

2. 湯浅俊彦『電子出版学入門――出版メディアのデジタル化と紙の本のゆくえ　改訂2版』（以下，書籍B）

【書誌情報】出版メディアパル, 2011年, A5判（21×14.8×1.4 cm），ソフトカバー。

【構成とレイアウトの特徴】

目次：横書き一段組で，章・節のタイトルと頁番号の間に長い「…」がある。計2頁。

「まえがき」，第1～7章：横書き一段組で，各節が見開き2頁または1頁で完結している。節番号とタイトルに飾り枠がついている。図表も豊富であり，書籍Aと比較して視覚的効果を生かすレイアウトになっている。脚注が各頁の最下部の罫線の下に記載されている。計81頁。

囲み記事：節の終わりに四角枠で囲んだ横書き一段組，1頁の記事が13箇所挿入されている。計13頁。

年表：巻末に「資料編」として横書きの表形式で「年」「月」「事項」が3列で表示されている。計19頁。

索引：横書き3段組で文字が小さい。計2頁。

章の表紙：各章および年表の冒頭に1頁ずつ挿入されている。章番号または「資料編」というタイトルに飾り枠が付き，続いて章および年表のタイトルが表記され, その下に「この章の概要」が四角枠で囲んだ横書き一段組のタイトルが縦書き1行で表示されている。計8頁。

その他：扉と奥付。計2頁。

以上, スキャンした印刷部分は127頁。白ページはなかった。見開きでは64面（コピーで64枚）となった。

2－3　実験手順と評価

I～IIIの全てについて，スキャンとOCR化に要した時間をそれぞれ計測した。さらにIIではコピー作業時間を，IIIでは本の裁断時間を計測して加算した。コピー（等倍で書籍AはB5, 書籍BはA4）および裁断は書籍A，書籍Bそれぞれ1回のみの作業とし, その時間を計測した。

また「画像補正あり」と「画像補正なし」の2つのパターンでデータを作成した。「画像補正あり」では，スキャンで画像を読み込んだ後OCR化に先立ち画像を補正した。補正では以下の部分を除外して文字部分を選択した（図3）。

書籍A:ノンブル（頁番号), 柱（ヘッダーやフッターの章タイトルなど），イラスト。

書籍B：ノンブル，柱，章タイトルの飾り枠，図表（ただしキャプションと出典の文字情報は選択）。

テキストデータ作成の経験がある作業者1と作業者2が, データ作成作業と時間の計測を行った。1名が作業している間，もう1名が作業時間をストップウォッチで計測した。書籍A，書籍BともにI～IIIのスキャンとOCR化について, 作業者1が「画像補正あり」, 作業者2が「画像補正なし」を担当した。

OCRソフトで作成したテキストデータについて，誤認識の程度や質，また誤認識がなくとも問題があるケースを視認によって第一著者が確認するとともに，スクリーンリーダーで音声に

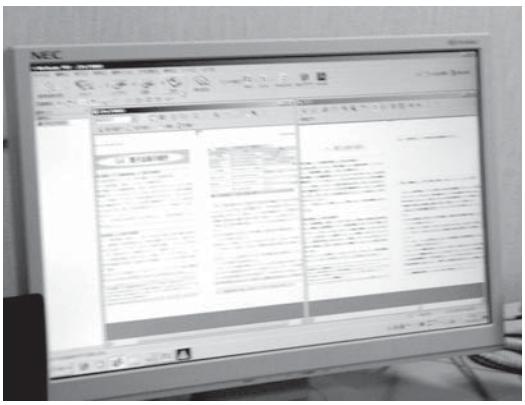

図3　画像補正

ディスプレイ左側に調整中画面が表示されている。OCRソフトで読み取る部分を選択すると，ディスプレイ右側に読み取られたテキストが表示される。

よる試し読みの良好さを第二著者（視覚障害者）が確認し，判定した。

3. 結果

書籍 A，書籍 B それぞれについて，テキストデータ作成時間と未校正テキストデータの品質を評価した結果を，表 1 に示した。データ作成時間については，コピー，裁断，スキャン，OCR 化（「画像補正あり」の場合は，画像補正作業時間を含む）の各作業工程の時間と，その合計を示した。

品質については，スクリーンリーダーによる試し読みを前提として，以下の3段階に分類した。

表 1　テキストデータ作成時間と品質

【書籍 A】

スキャン方法（枚数）	スキャナ・OCR種類	画像補正	コピー or 裁断	スキャン	OCR化	合計	品質
Ⅰ 原本（128）	Ⅰ	あり		63分59秒	40分39秒	104分38秒	優
		なし		51分09秒	15分21秒	66分30秒	優
Ⅱ コピー（128）	Ⅱ-1	あり	32分25秒	09分59秒	19分51秒	62分15秒	良
		なし	32分25秒	09分48秒	07分48秒	50分01秒	良
	Ⅱ-2	あり	32分25秒	11分58秒	35分28秒	79分51秒	良
		なし	32分25秒	11分36秒	18分50秒	62分51秒	良
Ⅲ 裁断本（252）	Ⅲ-1	あり	29秒	10分10秒	31分19秒	41分58秒	優
		なし	29秒	10分06秒	09分40秒	20分15秒	優
	Ⅲ-2	あり	29秒	07分56秒	39分45秒	48分10秒	良
		なし	29秒	09分52秒	16分28秒	26分49秒	良

【書籍 B】

スキャン方法（枚数）	スキャナ・OCR種類	画像補正	コピー or 裁断	スキャン	OCR化	合計	品質
Ⅰ 原本（64）	Ⅰ	あり		33分26秒	31分38秒	65分04秒	可
		なし		26分29秒	10分20秒	36分49秒	可
Ⅱ コピー（64）	Ⅱ-1	あり	13分43秒	05分41秒	24分13秒	43分37秒	良
		なし	13分43秒	05分44秒	09分43秒	29分10秒	可
	Ⅱ-2	あり	13分43秒	07分11秒	27分29秒	48分23秒	可
		なし	13分43秒	07分13秒	14分17秒	35分13秒	可
Ⅲ 裁断本（127）	Ⅲ-1	あり	25秒	05分44秒	28分48秒	34分57秒	良
		なし	25秒	05分39秒	07分17秒	13分21秒	可
	Ⅲ-2	あり	25秒	05分58秒	30分36秒	36分59秒	良
		なし	25秒	05分57秒	13分24秒	19分46秒	可

優：　特別なレイアウト部分以外は文字の誤認識が非常に少なく，十分理解可能である。
良：　本文の誤認識が「優」よりやや多いが，無視できる程度である。
可：　本文の誤認識が「良」よりやや多く，文章の欠落や順序の間違いなど，内容の理解を大きく損なうエラーが部分的に発生している。しかし誤認識がない部分については，理解可能である。
不可：本文の文字の誤認識に加えて，文章の欠落や順序の間違いなど，内容の理解を大きく損なうエラーが発生しており，ほとんど意味不明である。

ただし3－2で述べるように，レイアウトの特徴が原因となり，「優」「良」であっても文章の内容の理解を損なう問題が発生している部分を含む場合がある。

3－1　データ作成時間

【合計作業時間】

データ作成時間の合計は，書籍A，書籍Bいずれの場合も，「画像補正あり」「画像補正なし」ともに，時間が長い順にI＞II-2＞II-1＞III-2＞III-1となった。

【装置による違い】

コピーの作業工程がないにもかかわらず，Iの合計作業時間がIIよりも上回ったのは，スキャン時間の差が大きい。IとIIでは，スキャン枚数は同じであるが，Iでは手作業で原本を直接スキャンしたため，自動給紙でスキャンしたIIよりも約5～6倍の作成時間を要した。書籍Aのコピー作業は1枚あたり15.2秒であったが，Iのスキャン作業は「画像補正あり」で1枚あたり30秒であった[3)]。1枚スキャンする度に本をスキャナに押し付けて30秒待つ作業は負担が大きく，作業者に疲労と集中力の低下がみられた。

自動給紙でコピーをスキャンしてOCR化したIIについては，スキャン時間，OCR化時間ともにすべてII-2＞II-1となり，比較的高価格な装置で作業時間が短縮された。特に書籍Aでは，「画像補正なし」でのOCR化時間についてII-2（18分50秒）がII-1（7分48秒）の2.4倍となり，II-1の装置の優位性が顕著であった。

【レイアウトの影響】

書籍のレイアウトの違いがOCR化時間に影響した。本文では1頁あたり，書籍Aが18行（44字／行），書籍Bは31行（35字／行）であった。さらに書籍Bでは，飾り枠や図表が多数使われており，スキャン1枚あたりの情報量は書籍Aよりも多かった。コピーの影の影響を排除するため，III（裁断本）の「画像補正なし」で比較すると，III-1について書籍Aは2.3秒／枚，書籍Bは3.4秒／枚であった。またIII-2については1枚あたり，書籍Aは3.9秒／枚，書籍Bは6.3秒／枚であった。このようにIIIの「画像補正なし」では，OCR化時間について書籍Bが書籍Aの1.5倍（III-1）または1.6倍（III-2）となった。

3－2　品質

【全体的傾向】

書籍A，書籍Bともに「不可」はなかった。
書籍A：全般に良好であった。縦書き一段組の本文については，I～IIIの全てでスクリーンリーダーの読み上げに支障がなかった。特に原本を直接スキャンし画像の補正を行ったI「画像補正あり」の品質が，最も良好で精読が可能であった。一方，比較的低価格の装置を使用したII-2，III-2では「ン」を「ソ」とする誤認識が頻発するなど（たとえば人名の「ゴフマン」は26ヶ所中正しく認識されたのは1ヶ所のみで，24ヶ所が「ゴフマソ」，1ヶ所が「ゴフ-r ソ」），品質はやや落ちたが，試し読みとしての品質を大きく損なうことはなかった。「画像補正なし」では，ノンブ

3）Iのスキャン作業は1枚あたり「画像補正あり」（作業者1が担当）で30秒，「画像補正なし」（作業者2が担当）で24秒であった。スキャンには画像補正作業の有無は影響しないため，6秒の差は作業者の個人差による。

ルや柱を削除していないため，頁番号や欄外に記載された章のタイトル等が文章の途中で挿入された。定型的な文言が頻繁に挿入されるので聞いていて煩わしいが，逆に定型的なので文章を理解する流れが中断されても，文意を理解できる。したがって，慣れれば試し読みは可能である。

表1の品質評価の「優」「良」の差は，「マンガ書評」という画像で再掲された雑誌記事の認識精度の差である。なお，「コラム」ではレイアウトが原因で，I～IIIの全てにおいて問題が生じた（後述）。「コラム」は本文ではないため，試し読みにおける影響は相対的に低いと判断し，品質評価の対象としなかった。

書籍B：全般に書籍Aよりも誤認識が多く，相対的に品質が低下した。主な理由としては，後述するように2頁見開きでスキャンしたI，IIで本文の文意の理解を大きく損なう誤認識が発生したこと，目次のレイアウトのためスクリーンリーダーによる音読が大変聞きづらくなったことがある。

I～IIIの「画像補正なし」では，文章の途中で頁番号や章タイトルが挿入されるのに加え，表の数字の羅列や図の誤認識（文字化け）が発生したため，理解が困難になった。キーワードを拾って内容を推測するという方法であれば，試し読みは可能である。

一方，裁断本（III）の「画像補正あり」では，見開きのスキャンやコピーに伴う問題がなく，III-1，III-2ともにノンブル，柱，図表を削除したため，品質は良好であった。

【レイアウトの特性と誤認識】

①　目次

書籍A：章や節の番号が飾り文字であるため，本文にはほとんど誤認識のない「画像補正あり」のIでも，番号の脱落が多発した。

書籍B：章の下に節が順序良く配列されているが，OCR化では目次の正確な再現が困難であった。

例えば最も品質が良好であったIII-1「画像補正あり」でも，章と節の番号のみが先に列挙され，かつイタリックで表示された節番号が文字化けする（「*1.1*」→「z/」，「*1.2*」→「L2」など）という誤認識が生じた。また，章や節のタイトルの後ろに「…」が続き，右端に頁番号が表記さ

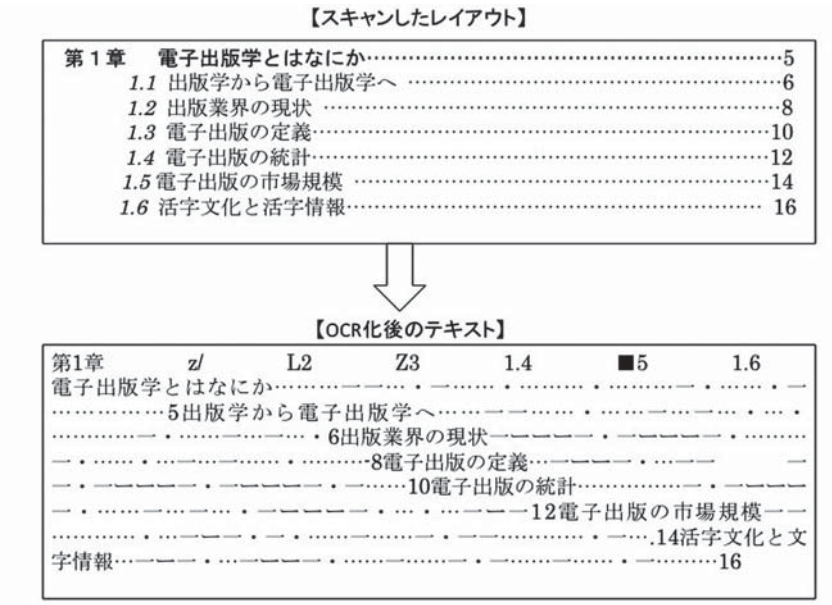

図4　書籍B目次の誤認識

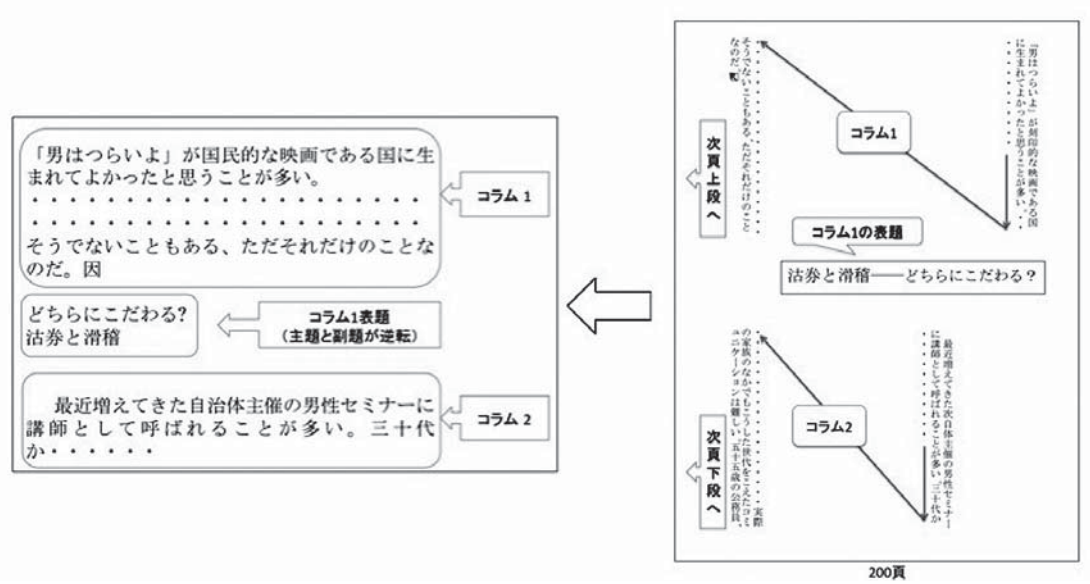

図 5　書籍 A「コラム」の誤認識

れている。そのため点の集合が一部漢数字の「一」と誤認識された（図 3）。学習機能によって，目次の 2 頁目から「…」はすべて削除されたが，目次の冒頭では「テンテンテンイチイチテンテンイチテンテン」というように逐一読み上げられるため，試し読みにしても聞き続けるのが苦痛であった[4)]。なお，OCR ソフトが異なる III-2 では「…」が「一」に誤認識されることはなかった。

②　画像と段組

書籍 A：四段組の雑誌記事画像が再掲された「マンガ書評」の認識精度に差が出た。I（原本）と III（裁断本）は良好で文意を十分理解できたが，コピーをスキャンした II は誤認識が多く理解不能であった。

縦書き二段組の「コラム」は，上下段にわかれたコラムを，それぞれ右頁から左頁に通して読むレイアウトになっていた（図 2）。OCR では頁ごとに文字認識されたため，異なる 2 つのコラムが半分ずつつなげられ，その間にコラムの見出しが入った（図 5）。そのため「コラム 1 前頁→コラム A 表題→コラム 2 前頁→コラム 1 次頁→コラム B 表題→コラム 2 次頁」の順に配列され，異なる内容の文章がパッチワーク状につながれる結果となった。

書籍 B：I「画像補正あり」では，本文に関して，見開きの左右で同じ位置にある文章を，頁を越えて一行として認識してしまい，全く文意が通らなくなっている部分があった。6-7 頁の「1.1 出版学から電子出版学へ」を例にとると，6-1 → 6-2 → 脚注→ 7-1 → 7-2 → 7-3 の順に文章を読み進むが，例えば I「画像補正あり」では，6-1 → 7-1 → 7-2 → 6-2 →脚注→ 7-3 の順に誤って認識され文章が配置された（図 6）。

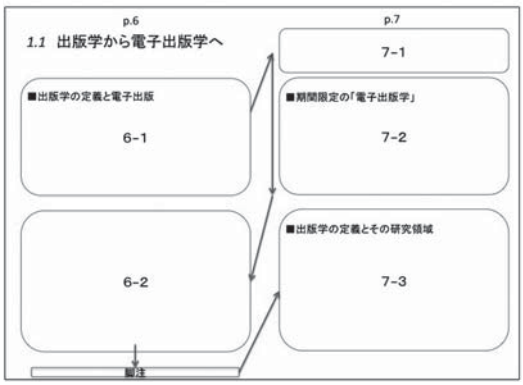

図 6　書籍 B 見開き文章配列の誤認識

4）DTalker for Mac Ver3.0 では，「…」は読みあげられずスキップされた。

また，見開きの前後を逆に認識する（II-2「画像補正なし」），ページの脱落（I「画像補正なし」，II-2「画像補正あり」）が生じている部分もあった。

巻末の罫の入った年表や三段組の索引はレイアウト通りに秩序だってテキスト化されず理解困難であった。また索引は文字が小さく，誤認識が起こりやすくなった。

③　引用と注

書籍 A：本文は縦書きだが本文中に挿入されている引用文献がすべて横書きであるため，その部分はすべて文字化けを起こした（図 7）。引用文献の文字化け発生頻度は，平均して本文 5.4 頁あたり 1 ヶ所であり（和文献が 35，欧文献 4 で合計 39，本文 212 頁），試し読みには大きな支障はなかった。なお，OCR ソフトに「読取革命」を使用した II-2，III-2 では，英文のみ正しく認識した。

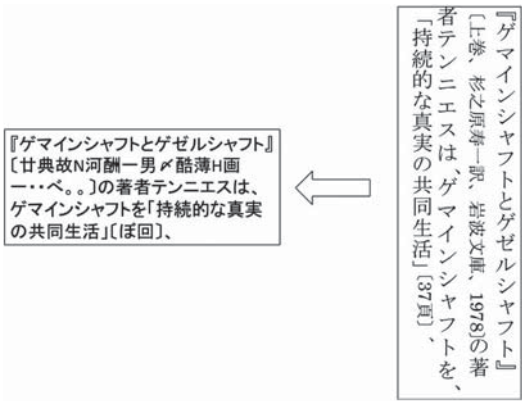

図 7　書籍 A 引用文献の誤認識

書籍 B：脚注の文自体は概ね正確に認識された。文章の途中で脚注が挿入されるため，スクリーンリーダーで聞くと文意の理解にやや支障をきたすが，試し読みは可能である。

④　飾り枠

書籍 B では，本文の節のタイトルが飾り枠の中に表示されている。グレーの四角い枠の中央がぼかしで白抜きになり，そこに節の番号とタイトルが表示され，視覚的には印象的なデザインになっている。しかし，OCR 化ではあらかじめ四角の枠をはずしても，ぼかしが入っているせいか，正しく認識されなかったり，文字全体が脱落したりする場合が多かった。

⑤　コピーの影

II ではコピーの影が品質に影響した。見開き 2 頁のコピーの左端に影が入っており，これが読み取り開始部に当たるため，頁切り替えの最初に数行～ 20 行程度にわたって誤認識された文字が飛び飛びに入っている場合が多かった（図 8）。また，頁始めの文章が脱落したりしている場合もあった。

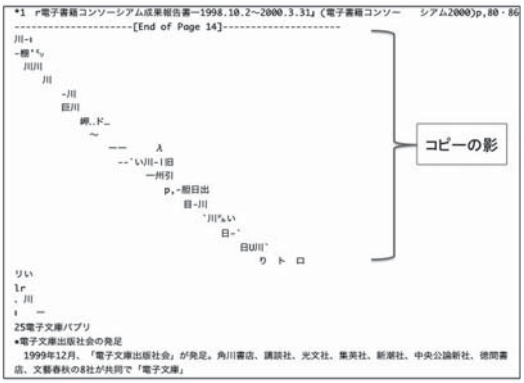

図 8　書籍 B コピーの影の文字化け

4. 考察

4－1　作成時間短縮と「画像補正なし」での提供の可否

本実験の結果，全ての装置，書籍の種類，データ作成方法において，品質には「不可」はなかった。「可」では文意がとれなくなる部分や，誤認識による意味不明の文が挿入されるため，スクリーンリーダーの読み上げ音声を全て聞くことは苦痛である。しかし，適当に飛ばし読みをし，どのような語句が使用されているかを知ることは可能である。

図書館蔵書のデータ提供を前提とする I と II で，最も短時間でデータを作成できたのは，書籍

A，書籍Bともにコピーのスキャンで装置が比較的高価格のII-1「画像補正なし」であった。比較的低価格の装置を使用したII-2「画像補正なし」では，II-1よりも書籍Aで7分50秒，書籍Bで6分3秒，作成時間が長くなった。しかし，自動給紙で自動変換であるため，労力は要さない。また品質もII-1よりやや劣るものの，書籍Aでは「良」，書籍Bでも「可」であり，今回対象とした書籍に関してはII-2の装置も実用的であった。

一方，Iの原本のスキャンでは，手作業でのスキャンであるため労力と時間を要し，現状では実用性に乏しい。しかし，原本を自動めくりし高速かつ高い精度でスキャンする装置が開発され，2013年度にも図書館や博物館向けにサービスを開始するとの報道もある[5)]。将来はコピーの段階を挟まずに，効率的に原本からダイレクトにテキストデータを作成する装置を，大学図書館で活用することが望まれる。

作成時間については「画像補正なし」が勝るが，品質については「画像補正あり」の方が優れている。「画像補正あり」ではノンブル，柱，図表などの文章の理解を妨げる部分が除外されるので，格段に読みやすくなる。「画像補正あり」と「画像補正なし」の作成時間の差は以下のとおりである。

書籍A　II-1：12分3秒／II-2：16分38秒

書籍B　II-1：14分30秒／II-2：13分12秒

画像補正は手作業であるため，この時間の差は労力の差でもある。書籍Aの本文のようにレイアウトが単純で「画像補正なし」でも認識精度が高い部分については，補正せずに提供してもよいだろう。一方書籍Bについては，飾り枠や図表が多いため画像補正が必要である。

4－2　試し読みに必要な配慮

①　正確な目次の提供

利用者の学習・研究のために，大学図書館が提供する試し読み用の未校正テキストデータは，その本を精読する必要があるかを点検するための暫定的なデータである。誤認識を含む試し読みのデータを読む利用者は，その本の分野に関する予備知識がない場合もある。例えば，書籍AのII-2，III-2で重要な引用文献の著者名である「ゴフマン」が，ほとんど「ゴフマソ」となっていた。それを誤認識だと知らなければ，「ゴフマン」への言及とわからず，精読に進むか否かの判断を誤る可能性がある。しかし，書籍Aの目次には「ゴフマン社会学」という記載がある。目次が正確であることを利用者が知っていれば，本文中の「ゴフマソ」は「ゴフマン」であるとの確信を持てる。内容の理解において重要な手がかりとなる目次が正確であることは，利用者の判断において重要である。書籍Aであれば，画像補正をする時間を目次の作成に使い，テキストデータを「画像補正なし」で正確な目次と共に提供すればよい。目次の作成においては，Webcat Plusの目次データベース等を活用し，手打ちで作成したほうがスキャンしたデータを校正するよりも能率的である。

②　画像補正での除外情報の提供

図表のほか，目次，索引，囲み記事などレイアウトが複雑で，多数の誤認識の発生が予測される場合は，それらをOCRの認識対象から除外する画像補正を行う。また，段組が多い部分は，認識順序を指定して文字認識するか，それが困難な場合は除外する。利用者には，目次に即してどの部分を除外したのかの情報を提供する。

③　鮮明な状態で1頁ごとにスキャン

裁断した本に近い状態の鮮明なコピーを作成しようとすると，原本を強くコピー台に押し付

5）大日本印刷と東京大学が，書籍を自動でめくりながら1分間に250頁をきれいによみとる装置を開発したと報道された。「大日本印刷と東大，自動でパラパラめくって，書籍ばらさず電子化，高速で読み取り」『日本経済新聞』2012年11月16日（夕刊）。この記事に関連した大日本印刷のニュースリリースとして，「大日本印刷　東京大学　世界最速レベルのブックスキャナーを開発」（2012年11月19日）http://www.dnp.co.jp/news/10061081_2482.html〈2012年11月23日確認〉。

けたりすることになり，本を破損する恐れがある。しかし，コピーの文字の歪みや影は誤認識の原因となるため，コピーはできるだけ丁寧に行う必要がある。また，横書きの本を見開きでコピーすると左右の頁がハの字に傾いてしまう場合があり，誤認識が増加する恐れがある。さらに，3－2の②で述べたように，横書きの見開きでは文章の認識順序を間違う場合がある（図5）。したがって見開きのコピーを中央で裁断するか，1頁ごとにコピーをとったものをスキャンすることが望ましい[6]。なお，OCR処理前に画像の歪みや影を補正することができれば，誤認識の回避に役立つ。

5. 終わりに

本研究でも明らかになったように，視覚的には理解しやすいレイアウトであっても，OCR化の際には誤認識が起こりやすい。また書籍Aのように，未校正テキストデータでも精読に足るほどの精度を確保できる場合もあれば，レイアウトに伴う誤認識などにより，異なる文脈の文章が混在して全く文意がつかめなくなることもある。

すでにAmazon.comが，ブックリーダーのKindleで電子書籍の自動音声読上げを実現しているように，日本においても視覚障害者が利用しやすい電子書籍が今後流通する可能性がある。しかし，アクセシビリティに配慮した新刊が刊行されるようになっても，知のアーカイヴである図書館の膨大な資料を印刷物や画像データからテキストデータに変換する必要は残る。とりわけ学術的成果を生み出し次代に継承する社会的責務をもつ大学の図書館においては，視覚障害等で印刷物を利用できない学生や研究者にも，蔵書を十分に活用できる環境が求められる。そのうえで試し読み用の未校正テキストデータを適切に作成し，利用者に提供することは，今後も一層重要になるだろう。

6）書籍の状態によっては，鮮明なコピーが難しい場合もある。コピーの影で誤認識が起こる場合，図8のようにあるまとまりとして視認できることもある。このような場合は，その部分だけ削除して利用者に提供する方法もある。

謝辞

著書の利用を許可してくださった中村正氏，湯浅俊彦氏に感謝致します。また実験データの作成にあたり，立命館大学大学院先端総合学術研究科院生の安田智博さん，平田剛志さんの協力を得ました。

本研究は2011年度立命館大学人間科学研究所研究所重点プログラム「読書障害学生支援における大学図書館の課題」，および立命館グローバル・イノベーション研究機構研究プログラム「電子書籍普及に伴う読書アクセシビリティの総合的研究」の支援を受けました。

引用文献

国立国会図書館（2003）デジタル環境下における視覚障害者等図書館サービスの海外動向．Current Awareness Portal. http://current.ndl.go.jp/files/report/no1/lis_rr_01.pdf（2012年7月10日）

国立国会図書館（2011a）全文テキスト化実証実験に係る調査及び評価支援等作業実証実験報告書．国立国会図書館．http://www.ndl.go.jp/jp/aboutus/digitization_fulltextreport.html（2012年7月10日）

国立国会図書館（2011b）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．国立国会図書館．http://www.ndl.go.jp/jp/aboutus/digitization/ocrzenbun.pdf（2012年7月10日）

立命館大学障害学生支援室（2010）テキスト校正ガイドブック．青木慎太朗（編）「視覚障害学生支援技法増補改訂版，立命館大学生存学研究センター報告」，*12*，174-202.

植村要・山口真紀・櫻井悟史・鹿島萌子　（2010）　書籍のテキストデータ化にかかるコストについての実証的研究―視覚障害者の読書環境の改善に向けて．Core Ethics, *6*, 37-49.

（2012. 7. 18　受稿）（2013. 1. 7　受理）