論文

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――

植村　要*・山口真紀**・櫻井悟史***・鹿島萌子****

Ⅰ　はじめに

本稿では、書籍をテキストデータ化するには、イメージ・スキャナとOCRソフトを用いて行うより、DTPで組版された印刷用データを用いて行う方が低コストかつ正確にデータ化できることを実証的に示し、視覚障害などによって読みに困難が生じている者に対して出版社・印刷所は印刷用データから製作されたテキストデータを提供する必要があることを主張する。また、組版がDTPで行われるようになる以前に刊行された書籍は、今後もイメージ・スキャナとOCRソフトを用いてテキストデータ化することになるため、その作業に要するコストを実証的に明らかにする。

読みに困難を感じる人として最も早く可視化され対策をとられてきたのが視覚障害者であり、従来、点訳・音訳といった方法で対応されてきた。これらの作業は、点字図書館が中核となりながらも、そのほとんどが無償ボランティアによって担われてきた。そのための人材が十分ではないこともあって、刊行されている書籍の全てが点訳・音訳されているわけではなく、一般書に偏る傾向がある。また、新刊の刊行後、点訳・音訳が完成するまでに数ヶ月を要することも珍しくない。そのため、より専門的な書籍を、必要が生じた時点で速やかに読む必要がある学生、特に大学や大学院に在籍する視覚障害者にとっては、現在の情報保障環境はかならずしも必要を充足するものにはなっていない。

日本学生支援機構は、2008年5月1日現在で、国内の大学、大学院、通信制、短大、高等専門学校（以下、学校）1,218校を対象に、障害学生の修学支援に関する実態調査を実施した。回収率は100％で、以下の結果が得られた。障害学生が在籍する学校は、719校（59％）で、そのうち障害学生に対して何らかの支援を行っている学校は543校であった。障害学生は、調査を開始した2005年度以降最多の6,235人（0.2％）であり、そのうち3,440人が学校から何らかの支援を受けている（日本学生支援機構学生生活部特別支援課2009）。

大学で学ぶ障害をもつ学生に対する支援の必要性については、いまや誰も否定しないだろう。障害学生の修学支援については、多くのシンポジウムや研究会が開催され[1]、文献も多数出版されている[2]。そこで論じられている内容は、種別・程度によって異なる障害についての理解、従来不可視だった発達障害やディスレクシアを障害学生支援の対象にすることの必要性、国内外の各教育機関での実践報告、多機関とのネットワーク化の必要性、支援技術・コーディネートのノウハウの共有、障害をもつ学生と支援をする学生との交流、などである。つまり、障害学生支援としてなされる行為を適切かつ円滑に行うための方法が検討されているのである。しかしながらここには、不思議なほどに資金面の課題が登場しない。資金面での問題が一切生じていないためかというと、そうではないにもか

キーワード：視覚障害者の読書環境、DTP、OCR、テキストデータ化、コストの比較

＊　立命館大学大学院先端総合学術研究科　2006年度入学　公共領域
＊＊　立命館大学大学院先端総合学術研究科　2006年度入学　公共領域
＊＊＊　立命館大学大学院先端総合学術研究科　2007年度入学　公共領域
＊＊＊＊　立命館大学大学院先端総合学術研究科　2008年度入学　表象領域

かわらず、である[3]。

このような視覚障害者など読みに困難を生じている者の読書環境が改善される可能性が出てきた。2009 年 6 月 12 日に改正著作権法が成立し、2010 年 1 月 1 日から施行されることとなったのだ。今回の改正は、情報通信技術の進展に対応したものであり、次の 3 本柱から構成されている。①インターネット等を活用した著作物利用の円滑化を図るための措置。②違法な著作物の流通抑止のための措置。③障害者の情報利用の機会の確保のための措置。従来、視覚および聴覚を中心に障害者の著作物の利用、つまりは障害者が情報を入手することを可能にするための媒体の変更は、第 37 条を中心に規定されていた。今回も③に関連する改正は、第 37 条を中心に行われた。それは、次の 3 点を主な内容としている。①障害の種類を限定せず，視覚や聴覚による表現の認識に障害のある者を対象とすること。②デジタル録音図書の作成，映画や放送番組の字幕の付与，手話翻訳など，障害者が必要とする幅広い方式での複製等を可能とすること。③障害者福祉に関する事業を行う者（政令で規定する予定）であれば，それらの作成を可能とすること。この改正によって、従来は情報の入手に困難を生じている者が視覚障害者と聴覚障害者に限定されていたのだが、その範囲が広げられたこと、また変更が認められていた媒体は点字・録音・字幕に限定されていたものが、その幅が広げられたことなど、障害者をめぐる情報保障の環境は大きく改善される見通しが出てきた[4]。これによって、既に読みに困難を生じている人に対する有効な媒体として提案･開発されてきた書籍のテキストデータ、およびマルチメディア・デイジーの製作と頒布が認められることになった。

今回の改正著作権法では、国立国会図書館がその保存を目的に原資料のデジタル化を行うための改正も行われた。これに先立つ 2009 年 5 月には、国立国会図書館の資料を大規模にデジタル化するための補正予算約 127 億円（前年比約 100 倍）を計上し、成立している。同館の全蔵書 917 万冊のうち、この補正予算で約 92 万冊（同館の国内図書の 4 分の 1 近く）のデジタル化が終わる計算であり、今年度中に国内図書の 1968 年刊行分までの約 77 万 3,000 冊をデジタル化する計画である。これは、米グーグル社による書籍の全文ネット検索サービスに対抗するものだという（「国会図書館、デジタル化予算前年比 100 倍計上」朝日新聞 2009 年 5 月 16 日）。この著作権法改正、および補正予算成立を受けて、国立国会図書館は 2009 年 9 月 17 日、国立国会図書館新館講堂において出版社を対象に「出版社を対象とする国立国会図書館の資料デジタル化に関する説明会」を開催した。予算成立の直後、河村宏は、このデジタル化が書籍をスキャンして pdf 形式にするのみで終わりそうであることに対して、発達・学習障害などで「読める」教科書のない子どもたちに対する教科書整備を優先すべきであり、デジタル化はデイジーのような適切な形式で行うことを望む旨、投書している（「（声）デジタル化は教科書優先で」朝日新聞 2009 年 5 月 31 日）。ここで注意しなければならないことは、国立国会図書館が行うデジタル化がテキストデータ化を意味するのではなく、スキャンをして pdf 形式にするのみで終わるものだということ、画像ベースの pdf データは、視覚障害者の読書環境の改善に何ら資するものではないことである。デジタル化には視覚障害者など読みに困難を感じている人の読書環境を改善する可能性があるにもかかわらず、そのような目的は含まれていないのである。

書籍のテキストデータは、今後の視覚障害者の情報保障環境を整備する上で欠かすことのできない媒体である。テキストデータは、スクリーンリーダーをインストールしたパソコンを用いることで、視覚障害者にもそのままの形式で読むことができる。テキストデータは、さらに別の媒体の製作に利用もされる。自動点訳ソフトを用いる今日の一般的な点訳の工程では、その作業の前段階としてテキストデータを作成しなければならない。また、書籍のテキストデータを XML 形式に加工し、音声データや画像データと同期させた複合的情報提供システムであるマルチメディア・デイジーにも活用される。近年、マルチメディア・デイジーは、視覚障害者のみならずディスレクシアなど、読みに困難を感じる多様な人に対して、情報の入手を容易にする媒体の一つとして注目されている。書籍をテキストデータ化することに対しては、データの複製・改ざんが容易であること、外部への流出の可能性が払拭できない、といった危惧が出版社から示されている（植村 2008）。マルチメディア・デイジーは、この危惧に対する有効な技術的対策をとっている。

書籍のテキストデータを製作するには、二つの方法がある。一つは、DTP によって製作された印刷用データを活用するものである。近年、書籍の製作は、DTP（Desktop Publishing）によって組版されるようになったことで、印刷用データからテキストデータを作成することが比較的容易に可能になった。一部の書籍には奥付に「テキストデータ引換券」が添付され、これを読者が出版社に送ることによって出版社から読者にテキストデータが提供され

るようになった。また、「テキストデータ引換券」が添付されていない書籍であっても、直接連絡をすることでテキストデータを提供する出版社も複数存在する。その一方で、書籍のテキストデータを一切読者に提供しない出版社も複数存在する。植村は、そのような対応がなされる背景に、法的要素、技術的要素、コスト要素、出版社内のルールが関わっていること、そして、コスト要素には次の2種が含まれていることを調査に基づいて示した（植村2008）。①初版刊行時の著作権者と出版社との出版契約が印刷による出版のみを想定したものになっているため、読者へのテキストデータの提供に際して、改めて著作権者に連絡をとって許諾を得る必要があること。②印刷用データをtxt形式でエクスポートした際に生じる文字化けなどを修正する必要があること。今回の著作権法改正後も、コスト要素および技術的要素、これに対する出版社内のルールは問題となる。ただし、コスト要素①は、既刊書については発生するが、新刊書については、初版刊行時の出版契約に読者へのテキストデータ提供条件を明記すれば解消されるものである。であるなら、出版社が問題にするコスト要素②、および技術的要素にかかわって生じるコストがどれほどのものなのかを、実証的に明らかにする必要がある。本稿は、ここに注目する。

もう一つの方法は、イメージ・スキャナとOCRソフトを用いて作成する方法である。これは上記した技術的要素から生じるものである。ただし、今日新規に出版される書籍の全てがDTPで組版されているわけではない。またDTPの技術が開発される以前に出版された書籍には、いうまでもなく印刷用データはない。これらの書籍は、イメージ・スキャナとOCRソフトを用いて、次節で詳述する工程でテキストデータ化することになる。

出版社が営利企業である以上、コストを問題にするのは、当然ともいえる。しかし国立国会図書館がデジタル化を行うためには巨額の予算が投じられるのに対して、視覚障害者などの読書環境の改善には、法改正のみで、何らの予算的措置もない。僅かな無償ボランティアに依存する現状を、容認どころか、むしろ積極的に推奨しているようでさえある。本稿では、上記のコスト要素②に関わって、視覚障害者などを取り巻く今日の貧困な読書環境が、出版社のどれほどのコストと引き換えにもたらされているものであるかを明確にする。加えて、出版社が拒否したコストを視覚障害者などがこうむるとき、それがどれほどのコストに膨張しているかを明確にする。これらの作業は、上記の技術的要素によって印刷用データの存在しない書籍のテキストデータ化に要するコストを明確にすることにもつながる。

以下のⅡでは、印刷用データ、およびイメージ・スキャナとOCRソフトを用いてテキストデータ化する作業の工程を、具体的な手順に即して詳述する。ⅢではⅡに示した工程に基づき、条件の違いによって要する時間の違いを実験する。Ⅳでは、Ⅲの実験結果を考察する。

Ⅱ　テキストデータ化の工程

テキストデータ化の作業は、大きく二つの工程からなる。すなわち、未校正データを作成する作業と、校正作業である。さらに未校正データを作成する方法には、Ⅱ－Ⅰに記すDTPで組版された印刷用データからエクスポートする方法と、Ⅱ－Ⅱに記すイメージ・スキャナとOCRソフトを用いる方法の二種がある。どちらで作成された未校正データであっても、校正作業は、Ⅱ－Ⅲに記す方法で行う。本章では、この作業工程を、具体的な手順に即して詳述する。

Ⅱ－Ⅰ　印刷用データを用いる手順

今日、広く用いられているDTPソフトに、InDesign（アドビ）とQuarkXPress（Quark）がある。これらのソフトは、作成した印刷用データをPDFやXML、txtなどの形式でエクスポートする機能を備えている。したがって、紙媒体の書籍を印刷するために製作された印刷用データからテキストデータを作成するには、特別の作業を要するものではなく所定の操作で可能である。あるいは、印刷用データを全文選択し、テキストエディタやワープロソフトにコピー＆ペーストすることでも可能である。しかし印刷用データにはフォントや段組などの指定、外字やルビ、図表や写真なども入っている。そのため、単純にtxt形式にエクスポートしただけでは、文字化けしたり、段組部分が入れ替わったりなどする。

Ⅱ－Ⅱ　スキャン作業の手順

本稿では、スキャン作業を、イメージ・スキャナで文書を読み取り、それを OCR ソフトでレイアウト・認識させるまでの一連の工程と定義する。スキャン作業は以下の手順で進められる。

1. 書籍を電動裁断機で裁断して、ばらばらの紙の束にする。図書館で借りたものなど裁断できないものの場合は、コピーをとる。
2. 文書をスキャナにセットし、OCR ソフトで文書の読み取りを行う。この時点では、読み取った文書は画像形式で認識される。
3. 必要に応じて文書の反転を行う。文書をスキャナにセットするときの向きによって、ディスプレイ上に表示された文書の方向が、上下・左右に転倒している場合がある。このような転倒は、読み取った文書を一括して文字認識する際に、文字化けして認識される。そこで、これを正対する向きに回転させる。
4.「見開き自動補正」を行う。特に見開きのコピーにおいてよく生じることとして、複写された文書がコピー紙に対して斜めになっていることがある。横書きの文書の場合、OCR ソフトは左の点から同じ高さの右の点まで、定規で線を引くようにまっすぐ読み取っていく。そのため、斜めになっているまま文字認識を行うと、文書内の行と行とを横断して認識してしまい、ほとんどの文字が文字化けしてしまうことになる。そこで、OCR ソフトの「見開き自動補正」機能を使い、可能な限り画面と平行にしてから文字認識を行うことで文字化けを防ぐことができる。ただし、コピー紙に文書が平行に複写されている場合、この工程は必要ない。
5.「自動レイアウト」機能によって、認識範囲の確定を行う。OCR ソフトは、黒いものを文字として認識しようとするため、見開きでコピーした際に生じる中央の黒い影や、文書の汚れ、埃なども文字として認識しようとする。これは、影や汚れを無理やり文字として認識するのであるから、当然のこととして文字化けに繋がる。そこで OCR ソフトの「手動レイアウト」機能によって、文字認識の対象にする範囲と、対象から除外する範囲を選択し確定する。これによって、より正確な認識を可能にする。
6. 同じく「自動レイアウト」機能によって、認識順の調整を行う。文書内にウインドウや段組がある場合、その塊ごとに認識の順序が決められる。ごく稀にではあるが、この順序が読みの順序と異なっている場合があるので、調整する。
7.「領域属性」の一つの「改行コード挿入指定」機能によって、文字認識の際に改行コードを挿入する位置を指定する。ここで「毎行改行」を選択すると 1 行ごとに、「自然改行」を選択すると段落部分にのみ改行コードが挿入される。しかし、誤認識などによって、正確な位置に改行コードが挿入されない場合もある。後述するように校正作業において、改行コードは段落部分にのみ挿入し、それ以外は削除するのであるから、「自然改行」に設定する方が作業量は削減できる。しかし、「毎行改行」の方がリズムをもって校正できる、とする校正者もいる。そこで、「毎行改行」と「自然改行」は、その文書の校正をする担当者の希望に応じて選択する。
8. 文字認識を行ってテキストデータにする [5]。

Ⅱ－Ⅲ　校正作業の手順

以下、立命館大学障害学生支援室（2009）によるガイドブックに即して記す。校正作業の基本は、原本に忠実に行うことである。原本に明らかな誤植と判断される部分があっても、修正することなく、OCR の段階で生じた文字の誤認識や文字化けのみを校正する。校正作業は、次の A、B の作業を同時に行うものである。

A. OCR の段階で生じた文字の誤認識や、文字化けの修正

OCR の段階で生じる文字の誤認識や文字化けは、比較的ひらがなやカタカナには少なく、漢字や英数字に多く生じる。アルファベットと数字の誤認識、漢字と英数字の誤認識、輪郭の似た漢字の誤認識、1 文字の漢字のへんとつくりを 2 文字に誤認識するなどである。具体的に、以下に例示する。その他、文字化けも含まれている。これらを一つ一つ原本と突き合わせて確認し、修正していくのである。

誤認識の例示
・「0」（数字）⇔「O」（アルファベット大文字）⇔「o」（アルファベット小文字）⇔「○」（丸印記号）
・「1」（数字）⇔「I」（アルファベット大文字）⇔「i」（アルファベット小文字）
・「､」「｡」（句読点）⇔「,」「.」
・「一」（漢数字）⇔「－」（マイナス）⇔「―」（ダッシュ）⇔「ー」（長音）
・「者」と「老」、「日」と「目」のように輪郭が似ている漢字、「国」や「園」のように部首が同じ漢字も頻繁に文字化けする

B. レイアウトの修正

・段落の改行とスペースの挿入

OCR で認識をした段階では、原本の 1 行ごとに改行コードが入る。時折、文章が続いている箇所に改行コードが挿入されていることもある。この改行コードを、段落の変わり目を除いて、全て削除していく。つまり、段落の変わる部分にのみ改行コードを入れるのである。

これには文字置換機能を用い、一括して改行コードを削除する方法もある。この場合は置換後に段落の変わり目を探し、改めて改行コードを挿入することになる。

そして新たな段落の冒頭にはスペースを一つ挿入する。

・ページ番号の挿入

ページが変わるところでは、改行コードを挿入し、ページ番号を挿入して、再度改行コードを挿入する。これで、そのページの始まりの部分に、ページ番号が表記されることになる。ページ番号は、「p. ○○」と半角英数字で記入する。文章の途中でページが変わっている場合でも、原本通りに改行してページ番号を挿入する。

・脚注の処理方法

文中にある脚注の番号は、「(注 1)」や「(*1)」などと記す。これに対応する脚注の文章には、文献によってページ脚注、章末脚注、巻末脚注などの様式があるが、テキストデータ化に際しては、ファイルの最後にまとめて記すことが望ましい。

・図表などの省略と明記

図表やイラスト、写真などは、キャプションのみ記し、基本的に割愛する。図表によっては文章化できるものもあるので、その場合は文章化する。割愛する場合は、「(校正者注：図○○省略)」など、割愛したことを明記する。その際、上のように括弧内の文字列が原本にある文字列ではなく校正者が書き加えたものであることを明確にしておく。

・ルビの校正

本文文字列の上、あるいは左右に、文字サイズを小さくして付されたルビは、多くが文字化けする。ルビは、当該単語の後ろに括弧に入れて挿入する。

・傍点、強調の校正

傍点や太字、斜体などで語句が強調されている文字列は、その文章末に、「校正者注：「○○」傍点)」と、傍点などが付された文字列を示す。

・目次の修正

目次は、独特なレイアウトが工夫されている。これは、OCR の段階でレイアウトそのものが崩れてしまう。たとえば、項目とページ番号を 3 点リーダーでつなげて表記していることがある。こうした場合、ページ番号の数字が

項目から切り離され、全く違う場所に置かれてしまうことがほとんどであり、OCR で読み込んだものを修正するよりも、校正者自身が最初から打ち込む方が早いこともある。

OCR ソフトで文字認識したデータを校正する際、校正者はテキストエディタやワープロソフトを用いて PC の画面上で修正を行う。この作業は、各校正者によって工夫がなされている。筆者らが行っている方法を、以下に 5 例示す。

校正者 A：OCR ソフトで文字認識した文章をディスプレイに表示し、通読する。このとき、矢印キーでカーソルをスライドさせながら読み進めていく。そして誤認識と思われる文字列や、意味の通じない文章が現れたところで、逐一原本にあたって必要な修正を施す。また、2 ページを一区切りに、傍点等で強調された文字列を原本で確認し、校正していく。

校正者 B：校正者 A と同じく、ディスプレイ上で、カーソルをスライドさせながら読み進め、誤認識などを逐一修正していく。またブックスタンド等を用いて、原本をディスプレイに隣り合わせて据え置く。こうすることで、ディスプレイと原本を同時に視野に入れることができ、加えて、両手を PC のキーボードに構え続けることが可能になる。これによって、誤認識を発見してから修正を施す過程が速やかになり、また傍点等の強調された文字列の校正を同時に行うことが容易になる。一通りの校正を終えてから見直しを 2・3 度行い、校正漏れの確認をする。

校正者 C：校正者 A と同じく、ディスプレイ上でカーソルをスライドさせながら読み進め、誤認識などを逐一修正していく。本節 B に例示した誤認識や文字化けしやすい文字列は、意識的に注意して読み進める。また読み進める過程で気づいた、そのデータにおいて誤認識されている頻度の高い文字列のメモをとっておく。校正が一通り終了したところで、「Ctrl」＋「F」キーで、メモした文字列を検索し、校正漏れの確認をする。肉眼では見落としやすい半角の「.」の検索は、ページ番号の挿入漏れの確認にもなる。ここまでを日本語チェック機能のあるワープロソフトで行い、続いて全文をメモ帳にコピー＆ペーストする。メモ帳の「右端を折り返す」機能を解除して、改行コードの消し忘れを確認する。また、ワープロソフトで校正する過程で、txt 形式では表示できない文字が使用されていた場合、メモ帳では「？」と表示されるため、検索して修正する。

校正者 D：最初に、メモ帳の「右端を折り返す」機能を解除して、段落部分の改行コードの修正を行い、同時にページ番号を全て入力する。これによって、ページ番号の挿入漏れを防ぐことができ、また、英数字を確実に半角に統一できる。また、次の誤認識の修正作業を行う際、誤認識の発見に注意力を集中しやすくなる。続いて、全文を日本語チェック機能のあるワープロソフトにコピー＆ペーストし、通読して誤認識などを逐一修正していく。文字化けの多いページは、そのページのみ再度スキャンと OCR ソフトによる文字認識をしなおすこともある。一通りの校正を終えてから任意のページを見直し、校正漏れがあった場合は、漏れの傾向から他のページの確認を繰り返す。

校正者 E：OCR ソフトで文字認識した段階の未修正のデータを、一度印刷する。そのさい、原本が縦書きであれば縦書きに、横書きであれば横書きにというように、行の向きを原本に合わせる。これを原本と突き合わせて誤認識された文字列を見つけ、ペンなどでひとつひとつ修正を書き込んでいく。大幅に文字化けしている箇所は、チェックを入れるに留める。その後に、PC のディスプレイ上での修正を行う。ここでは原本を用いず、書き込みを入れた印刷物を用いる。この方法には、ディスプレイを長時間凝視することからくる眼精疲労の軽減、原本と未修正データの印刷物を突き合わせることによる修正の見落としの削減、PC のディスプレイ上で修正するさいに修正の見落としを発見できる、といった利点が考えられる。一方で欠点としては校正作業に余計に時間を要している可能性、紙やインクなどのコストなどが考えられる。

Ⅲ　実験方法と結果

Ⅲ－Ⅰ　実験方法

本章では、前章に記したテキストデータ化の作業を実際に行い、データ化する文書の質的差異、およびデータ化方法の差異が、テキストデータの製作に要する時間に与える影響を実験した。実験に用いた文書、データ、機器は、以下のとおりである。

なお、下記2文書の執筆者、および出版社・印刷所には、本稿の目的を説明し、論文およびデータを使用することに対する許諾を得ている。

○文書

文書の質の違いが、テキストデータ化に要する時間に与える影響を測定するために、性質の異なる2つの文書を使用した。文書Aは、日本語文字列で構成されてはいるが、文章中に図表、画像、英文が混在している全15ページの論文である。文書Bは、全文のほとんどが日本語文字列で構成されている全17ページの論文である。共にグローバルCOEプログラム「生存学」創成拠点の報告書である《生存学研究センター報告》所収の論文であり、レイアウトは1段組、横書き、1ページ27行×34字を規定としている。ページ数に差があるが、段組や文字サイズ、1ページ文字数を統一することで変数を限定するために、同一雑誌から選出することを優先した。

選出したのは、以下の2文書である。

・文書A：櫻井浩子，2008,「NICUにおいて親と子がどのように関係性を築いていくのか――18トリソミー児の親の語りから」『PTSDと「記憶の歴史」――アラン・ヤング教授を迎えて』（立命館大学生存学研究センター，生存学研究センター報告1）: 139-154.
・文書B：斉藤龍一郎，2009,「スーダンと日本、障害当事者による支援の可能性」青木慎太朗編『視覚障害学生支援技法』（立命館大学生存学研究センター，生存学研究センター報告6）: 110-126.

○データ

上記2文書のそれぞれに、以下に示す3種のデータを準備した。

・データ1：文書を複写機でコピーし、イメージ・スキャナとOCRソフトで作成したデータ。OCRソフトのレイアウトを設定せず、文字認識を行った。
・データ2：データ1と同じ複写物を用い、イメージ・スキャナとOCRソフトで作成したデータ。OCRソフトのレイアウトを設定し、文字認識を行った。
・データ3：印刷用データをtxt形式に変換したのみの段階のデータを、出版社・印刷所から提供を受けたもの。文書Aは田中プリントから、文書Bは生活書院から提供を受けた。

なお、データ1とデータ2のデータ作成に際して、裁断した原本を用いることもできたのだが、本実験では複写物を用いた。これによって、OCRソフトの認識率の低下を招いた可能性がある。しかしそれでも敢えて複写物を用いたのには、以下のような理由がある。今後、DTPで組版された書籍のデータが提供されるようになれば、イメージ・スキャナとOCRソフトでテキストデータ化する必要があるのは、DTPが開発される以前の書籍に限られる。それらは、多くの場合絶版になっていたり、古本であっても高価であるなどの理由によって、裁断することが望ましくないと判断されることがありうる。そうした場合、図書館から借り出したものなどから複写物を作成して用いることとなる。そのような事情も鑑み、本稿では原本を裁断する方式よりも、複写物を用いた方式についての実験結果を提示する方が有益であると考えた。

○機器

上記のデータ１とデータ２の作成に使用した機器は、以下である。イメージ・スキャナは、文書を手で一枚ずつ差し替えるタイプのパーソナル向けスキャナではなく、文書の束をセットすると、自動的に一枚ずつ読み取っていくオート・ドキュメントスキャナであることが特徴である。これによってスキャン作業に要する時間の短縮を図ることができる。

・PC：TOSHIBA DynaBOOK（PAS4275PNHW）
・イメージ・スキャナ：Canon DR-3060/3080CII[6]
・OCR ソフト：Win Reader PRO v.10.0[7]

Ⅲ－Ⅱ　実験結果

２文書それぞれに３つのデータ、合計６つのテキストデータを製作し、その過程であるスキャン作業、および校正作業に要した時間を測定した。校正においては、要した時間とともに誤認識の箇所を数えた。加えて、校正において必要とした作業を記述することで、質の違いによる特色を抽出できるよう試みた。

本実験におけるデータ１のスキャン作業は作業者Ａが、データ２のスキャン作業は作業者Ｂが、準備した６種のデータの校正は全て作業者Ｃが行った。校正作業の全てを同一人物が行ったのは、校正者による能力の差を変数から除外するためである。なお作業者Ａ、Ｂ、Ｃは、全員がテキストデータ化の作業に３～４年の経験を有する者である。

実験結果を以下に表示する。**表１**は文書Ａの、**表２**は文書Ｂのテキストデータ化を、データ１、２、３のそれぞれを用いて行った際に要した時間を測定し、示したものである。

○表１：文書Ａ（図表、画像、英文含む）

文書Ａ	スキャン時間（分）	レイアウト・認識時間（分）	総文字数（字）	校正時間（分）	誤字（箇所）	その他
データ　１	1.27	2.52	9,287	72.54	220	157 字（日本語）＋ 718 字（英語）打ち込み、ページ数挿入、図表の説明
データ　２	1.32	5.41	9,049	50.36	177	318 字（英語）打ち込み、ページ数挿入、図表の説明
データ　３			9,091	28.19	6	ページ数挿入、図表の説明、段落はじめの空白削除

・データ１：１ページあたり約 312 秒、１文字あたり約 0.5 秒
・データ２：１ページあたり約 235 秒、１文字あたり約 0.38 秒
・データ３：１ページあたり約 113 秒、１文字あたり約 0.19 秒

○表２：文書Ｂ（図表、画像、英文をほとんど含まない）

文書Ｂ	スキャン時間（分）	レイアウト・認識時間（分）	総文字数（字）	校正時間（分）	誤字（箇所）	その他
データ　１	1.28	3.1	13,534	81.41	132 ＋ 3p	2,466 字（日本語）打ち込み、ページ数挿入
データ　２	1.31	5.39	13,386	41.21	128	ページ数挿入
データ　３			13,429	25.47	0	ページ数挿入、段落はじめの空白削除

・データ１：１ページあたり約 304 秒、１文字あたり約 0.38 秒
・データ２：１ページあたり約 171 秒、１文字あたり約 0.21 秒
・データ３：１ページあたり約 91 秒、１文字あたり約 0.12 秒

Ⅳ　考察

○文書Aと文書Bの比較

文書Aは、文書Bと異なって図表や英文が含まれている。データ1、2、3の全てにおいて、校正作業ではそれに応じた作業を必要とした。図表はテキスト表示できる部分を校正した上で、「校正者注」として内容についての簡単な解説を加えた。写真は、タイトル等キャプションのみ校正した上で、「校正者注」として割愛する旨記載した。

データ1と2において、英文は単語間の半角スペースが省略されていることが多く、原本と照合しながら一文字ずつ追うことになった。さらに日本語と英語が混在している箇所、特に文献表はほとんどが文字化けしており、英文を全て打ち込むことになった。

本実験では文書Aにおいて図表と写真が一点ずつであったこと、英文も少量であったこと、また文書Bにおいて、データ1に3ページにわたる文字化けがあったことから、文書AとBの校正時間に大きな差異は見られなかった。しかし、誤認識の箇所は文書Aに多く、校正に必要な作業量は図表や英文の量に比例すると考えられる。

○データ1とデータ2の比較

校正時間の差について、文書Bでは、データ1はデータ2の2倍近い時間を要した。文書Bのデータ1では、3ページにわたって全文に文字化けが生じており、1ページあたり約800字を全て手作業で打ち込まなければならなかったためである。この作業に要した時間は、1ページあたり10分を超えた。データ1とデータ2でレイアウト・認識に要した時間の差は、約3分であり、それがのちの校正作業においては約40分の差を生じさせている。加えて、1ページ内に文章と画像が併置してレイアウトされている場合も文字化けの頻度が多くなる。この点は、OCRソフトのレイアウトの際に画像を認識対象から除外することで避けられるものである。

誤認識の箇所について、データ2で文字化けしていた箇所はデータ1においても同じ箇所に見られた。たとえばカタカナの「カ」と漢字の「力」、数字の「0」とアルファベットの「O」の混在、「く」が「ぐ」と認識されるなど、誤認識される文字列の傾向は共通していた。これはコピーの状態や、スキャナの精度に依るだろう。ただし、データ2はOCRソフトのレイアウトの段階で、余白に記されているページ番号やタイトル、汚れや埃に対して確定された部分の削除もなされていたため、校正作業に要した時間は短縮された。

以上からOCRソフトで文字認識するにあたって、レイアウト機能によって文字認識する対象、除外する対象を設定することが、その後の校正作業に要する時間に大きな影響をもたらすことが確認された。

○データ1、2とデータ3の比較

本実験で最も明確な違いが見られたのは、この比較である。まず誤認識・文字化けの箇所の数に大きな差がある。それゆえ、文書Bでは、データ1の校正に要した時間がデータ3の約3倍という大きな差に繋がった。スキャナとOCRソフトによるデータでは、HPアドレスや文献表のように、日本語の文中に数字やアルファベットが混在する不規則文字列が正確に認識されないという傾向があるとはいうものの、どこに誤認識・文字化けがあるか予想できない。そのため校正者は常に高い注意力を維持していなければならない。しかしデータ3においては、誤認識・文字化けはほとんど認められず、校正者が行った作業はⅡ－ⅢのBに詳述したテキストデータ用のレイアウトの修正のみであった。データの正確さに一定の信頼があることは、校正者の注意力にかかる負荷の軽減に繋がる。

このことから、印刷用データをtxt形式でエクスポートしたデータが提供されることによって、テキストデータ化に要する時間、および校正者にかかる負担の大幅な軽減が期待できる。

Ⅴ　結論

本稿では、OCRソフトで文字認識するにあたって、レイアウト機能を丁寧に行うことによって、その後の校正作業に要する時間を短縮できることが確認された。また、校正作業に要する作業量は、文書に含まれる図表や英文の量に比例すると考えられる。これらのことから、書籍のテキストデータ化を、イメージ・スキャナとOCRソフトを

用いて行う場合には、OCR を丁寧に行うことによって、総体としての作業量を軽減できるといえる。また、校正作業に対する対価は、ページ数や文字数を単位とした算出より、時給による算出の方が、妥当と考えられる。

さらに、書籍のテキストデータ化は、DTP で組版された印刷用データを用いて作成する方法が、最も低コストであることを明らかにした。しかも、印刷用データからエクスポートした txt 形式のデータには、正確さに一定の信頼があることに加えて、文字化けする箇所に規則性があることから、その発見が容易であった。このことは、校正者にかかる負担の軽減に繋がるのみならず、作成したテキストデータの正確さの向上に繋がるものでもある。

書籍のテキストデータ化にかかるコストは、ページ数単位、文字数単位、時給のどれで算出したとしても、原本価格を大幅に上回る額になることに違いはない[8]。元を正せば、それは、出版社・印刷所が拒否したコストであり、それが読者の一人ひとりに降りかかるときには、およそ 3 倍にまで膨張するのである。出版社・印刷所は、このことを重く受け止めるべきである。

視覚障害などによって読みに困難を生じている者の読書環境の改善のためには、印刷用データからエクスポートした txt 形式のデータは、有用な資源となる。出版社・印刷所は、このことを強く認識し、データの提供に積極的に取り組むべきである。

＜注＞

1　以下に、主だったもののみ記す。

・2008 年 2 月 17 日、関東聴覚障害学生サポートセンター主催による 2007 年度研修会「第 2 回「講義保障を見よう、体験しよう」」（日本財団ビル）

・2008 年 2 月 23 日、立命館大学人間科学研究所主催による「障害学生支援の新しいビジョン――学生も職員も教員も＜研究者＞である」（立命館大学衣笠キャンパス）

・2008 年 6 月 21 日、（特活）アフリカ日本協議会主催による座談会「大学における視覚障害者支援の現状と課題　スーダンで今求められていること」（立命館大学衣笠キャンパス）

・2009 年 2 月 6 日 7 日、支援技術開発機構主催による国際シンポジウム「地域における障害者のインクルーシブな情報支援」（京都市国際交流会館）

・2009 年 2 月 8 日、NPO 支援技術開発機構主催による高等教育における障害学生支援に関するセミナー「日米のネットワーク構築をめざして」（東京大学）

・2009 年 2 月 21 日、日本福祉大学「障害学生支援フォーラム 2009」実行委員会主催による「障害学生支援フォーラム 2009」（日本福祉大学名古屋キャンパス）

・2009 年 2 月 21 日、特定非営利活動法人バリアフリー資料リソースセンター主催によるシンポジウム「障害のある人への読書支援におけるデータ活用の現状と課題」（大阪市立中央図書館）

・2009 年 2 月 28 日、障害学生支援に関する公開研究会として「理系の大学院の障害学生支援を、今、変える――富山大学生命融合科学教育部が発信する世界への提言」（名鉄トヤマホテル）

・2009 年 10 月 3 日、講演会「ウィーン大学での障害学生への配慮とは？」（日本点字図書館）

2　主だったものとして、日本障害者高等教育支援センター問題研究会（2001）、佐野（藤田）・吉原（2004）、独立行政法人国立特別支援教育総合研究所（2005）、関西学院大学教務部キャンパス自立支援課 KSC コーディネーター室（2008）などがある。

3　植村らは、立命館大学障害学生支援室が障害をもつ大学院生に対して実施した支援において、予算不足から事後的にその院生が 20 万円を超える自費による支払いを請求された事例を報告した（植村・青木・伊藤・山口 2007）。これに対して、その院生の在籍する研究科院生会から障害学生支援室に対する要望書が提出されていることを報告した（植村・青木・韓 2008）。

4　大きな改善には違いないが、そこには大きな問題点もある。2009 年 7 月に、日本図書館協会から文化庁長官宛に改善の要望書が提出されている（日本図書館協会 2009）。

5　経験的にいえば、複写物の場合、コピーがよりきれいであれば、認識率は上がる。しかし、コピーの文字濃度が濃すぎると逆に認識率は下がる。もっとも認識率が高いのは本を裁断してスキャンした場合である。

6　「DR-3060 機種仕様」http://cweb.canon.jp/e-support/qa/1055/app/servlet/qadoc?qa=041868　アクセス日：2009 年 10 月 8 日。価格は 398,000 円となっている。

7　「活字文書 OCR ソフト：WinReader PRO v.12.0」 http://mediadrive.jp/products/wrp/index.html　アクセス日：2009 年 10 月 8 日。2009 年 10 月 8 日現在発売されている Win Reader PRO はバージョン 12.0（198,000 円）である。

8　Ⅳの「データ 1、2 とデータ 3 の比較」の考察においても記したことであるが、さらに原本価格との比較が明確になるよう、表 3 を示す。これは、1 冊の書籍をテキストデータ化した際に要した時間、およびその作業に対する対価を時間単位、ページ単位で算定した場合の金額を示したものである。時間単位では、グローバル COE プログラム「生存学」創成拠点の採用する基準である 1 時間 1500 円で算定した。ページ単位では、立命館大学障害学生支援室の採用する基準である目次や文献表を 1 ページ 200 円、それ以外の本文を 1 ページ 80 円で算定した。なお、**表 3** は、実際のニーズに応じて書籍のテキストデータ化を情報保障として行ってきた際の記録であり、変数を制御した本稿の実験下で行った校正作業ではない。そのため、ここに補足的に示すにとどめる。

○表 3：テキストデータ化に要した対価

文献	文字数（字）	ページ数（枚）	時間（H）	金額（ページ）	金額（時間）	書籍の定価
A	187,813	284	65	￥24,040	￥97,500	￥2,415
B	262,852	286	71	￥30,560	￥106,500	￥2,940
C	185,711	234	26	￥19,440	￥39,000	￥3,045
D	182,816	279	47	￥24,360	￥70,500	￥2,940
E	80,934	205	4	￥16,880	￥6,000	￥400
F	115,297	198	7	￥16,320	￥10,500	￥1,680
G	603,248	626	69	￥52,000	￥103,500	￥2,956
H	130,121	272	52	￥22,960	￥78,000	￥1,890
I	338,101	375	37	￥30,360	￥55,500	￥5,775
J	424,107	438	37	￥35,120	￥55,500	￥5,040

＜文献＞

独立行政法人国立特別支援教育総合研究所，2005,『大学における支援体制の構築のために発達障害のある学生支援ガイドブック――確かな学びと充実した生活をめざして』ジアース教育新社 .

石川准，2006,「アクセシビリティはユニバーサルデザインと支援技術の共同作業により実現する」村田純一編『共生のための技術哲学――「ユニバーサルデザイン」という思想』（UTCP 叢書）未来社 : 124-138.

石川准，2008,「本を読む権利はみんなにある」上野千鶴子・大熊由紀子・大沢真理・神野直彦・副田義也編『ケアという思想』（ケアその思想と実践 1）岩波書店 .

岩井和彦，2007,「フォーラム 2007 著作権法の改正と視覚障害者の情報保障」『ノーマライゼーション』27（10）（通号 315）: 40-42.

金子和弘，2005,「大活字出版をブックオンデマンドで――プログラム「T-bridge」を活用した試み」『出版ニュース』（通号 2045）: 6-9.

関西学院大学教務部キャンパス自立支援課 KSC コーディネーター室，2008,『ボーダーをなくすために――視聴覚に障害がある学生への学習支援』関西学院大学総合政策学部ユニバーサルデザイン教育研究センター .

金智鉉，2006,「どのように視覚障害者は読書環境を獲得してきたのか――点字図書館、公立図書館、読書権運動の関係を中心として」『京都大学大学院教育学研究科紀要』52:108-121.

マルチメディア振興センター（FMMC），2009,「書籍デジタルコンテンツ流通に関する研究会報告書」（http://www.fmmc.or.jp/shoseki/090818/sho090818.html, 2009.10.05.）.

日本学生支援機構学生生活部特別支援課，2009,「平成 20 年度（2008 年度）大学、短期大学及び高等専門学校における障害のある学生の修学支援に関する実態調査結果報告書」（http://www.jasso.go.jp/tokubetsu_shien/documents/zixtutaichousa2008_1.pdf, 2009.10.05.）.

日本障害者高等教育支援センター問題研究会，2001,『大学における障害学生支援のあり方』星の環会 .

日本図書館協会，2009,「著作権法改正に伴う図書館の障害者サービスの充実に係る法運用について」（文化庁長官宛要望書）（http://www.jla.or.jp/kenkai/20090724.html, 2009.10.05.）.

渾大防三恵，2004,「IT 時代の「読書権」視覚障害者が求める出版ユニバーサル・デザイン」『朝日総研リポート』169:78-90.

斉藤龍一郎，2009,「スーダンと日本、障害当事者による支援の可能性」青木慎太朗編『視覚障害学生支援技法』（立命館大学生存学研究センター，生存学研究センター報告 6）: 110-126.

櫻井浩子，2008,「NICU において親と子がどのように関係性を築いていくのか――18 トリソミー児の親の語りから」『PTSD と「記憶の歴史」――アラン・ヤング教授を迎えて』（立命館大学生存学研究センター，生存学研究センター報告 1）: 139-154.

佐野（藤田）真理子・吉原正治編，2004,『高等教育のユニバーサルデザイン化――障害のある学生の自立と共存を目指して』大学教育出版 .

障害学研究編集委員会，2008,『障害学研究』(3)（特集　障害学生支援の障害学――入学障壁、学習障壁、就職障壁の過去と現在を問う）明石書店 .

高畑由起夫・星かおり・小野田弘之・植田幸利・久保田哲夫・細見和志・中條道雄・窪田誠・渡部律子・井垣伸子，2007,「障がいを持つ学生への学習支援（4）――関西学院大学総合政策学部における教材点訳のシステムについて」『総合政策研究』25: 125-139.

立命館大学障害学生支援室，2009,「資料編　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』（立命館大学生存学研究センター，生存学研究センター報告 6）: 150-178.

植村要，2008,「出版社から読者へ、書籍テキストデータの提供を困難にしている背景について」『Core Ethics』立命館大学大学院先端総合学術研究科，4: 13-24.

植村要・青木慎太朗・伊藤実知子・山口真紀，2007,「視覚障害学生支援の技法・2――立命館大学における視覚障害のある大学院生への支援についての一事例」障害学会第 4 回大会ポスター報告 .

植村要・青木慎太朗・韓星民，2008,「スーダン視覚障害学生支援の現状と課題――立命館大学における支援の現状からスーダンでの支援を考える」障害学会第 5 回大会ポスター報告 .

吉田次男，2006,「視覚障害学生に対する教材提供について――高等教育機関における臨床医学教育」『教育学研究』6: 25-35.

An Empirical Study on the Cost of Making Text Data of Books: Aiming to Improve the Reading Environment for the Visually Impaired

UEMURA Kaname, YAMAGUCHI Maki, SAKURAI Satoshi, KASHIMA Moeko

Abstract:

For the visually impaired to read books, information accessibility through methods such as Braille transcription, transliteration and production of books' text data must be guaranteed. However, information accessibility is currently insufficient for visually impaired students in higher education. To improve this situation, we compare methods of making books' text data in this paper. We demonstrate (a) how text data can be made accurately and at low cost through preformed desktop publishing (DTP) data and (b) the cost of making text data using image scanners and optical character reader (OCR) software. We tested (a) and (b) on two research papers of about the same length, one with text only, the other with illustrations. Regarding (b), we also experimented with the use or not of the OCR software's layout function, and we compared the time and cost of making text data relative to the type of document and the use or not of the layout function. In conclusion, we proved that the working hours and proofreading load for making text data can be greatly reduced by obtaining DTP printing-data from publishers or print shops. Also, when text is transcribed with OCR software, proofreading work can be reduced by using the layout function carefully.

Keywords: reading environment for the visually impaired, desktop publishing (DTP), optical character reader (OCR), making of text data, cost comparison

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――

植　村　　要・山　口　真　紀・櫻　井　悟　史・鹿　島　萌　子

要旨：

視覚障害者が書籍を読む場合、点訳･音訳･書籍のテキストデータ化などの情報保障環境が不可欠である。しかし、大学や大学院に在籍する視覚障害者にとって、現在の情報保障環境はかならずしも必要を充足するものにはなっていない。

そこで本稿では、この状況を改善するため、書籍のテキストデータ化に注目し、以下の目的を設定する。まず、(a) DTP で組版された印刷用データを用いて書籍のテキストデータ化を行うことで、低コストかつ正確にデータ化できることを実証的に示す。次に、(b) イメージ・スキャナと OCR ソフトを用いたテキストデータ化作業に要するコストも実証的に明らかにする。

この目的を達成するため、(a) と (b) の方法を用いて、性質の異なるほぼ同量の二つの論文をデータ化する実験を行なった。(b) については、OCR ソフトでレイアウトするか , しないかの差異も設けた。そこからデータ化する文書の質的差異、データ化方法の差異が、データの製作に要する時間やコストに与える影響を測定した。

結論として、出版社・印刷所から、印刷用データを提供してもらうことによって、データ化に要する時間、および校正者にかかる負担の大幅な軽減が期待できることが明らかになった。また OCR ソフトで文字認識するにあたって、レイアウト機能を丁寧に行うことにより、その後の校正作業に要する時間を短縮できることが確認された。