

3 テキストデータ作成に関する実証実験に関する評価

テキストデータ作成に関する実証実験では、以下に示す評価を実施した。

- ・テキスト化システムの構築の評価
- ・テキスト化システムの効率化・高度化の評価
- ・テキストデータ作成にかかる作業時間の評価

評価方法として、技術動向との比較、作業時間の測定および作業結果の分析に加えて、テキスト化作業の作業員 8 名、大学生と大学院生からなる一般利用者 11 名、有識者 11 名を評価者としたアンケート調査を実施した。本章では、これらの評価結果を示す。

3. 1 テキスト化システムの構築の評価

テキスト化システムに対して、以下の観点から評価を行った。

- ・日本語対応の OCR 出力フォーマット
- ・構造化メタデータ
- ・OCR 再学習機能
- ・出力フォーマット
- ・出版社データから OCR 出力フォーマットへの変換

3. 1. 1 日本語対応の OCR 出力フォーマットの評価

日本語対応の OCR 出力フォーマットについて、利用可能なフォーマットとその特徴を公開された文書に基づく文献調査により把握し、本実証実験で利用したフォーマットが全文テキスト化作業を適切に実施できるかを検証した。また、技術的な課題を整理した。

(1) OCR 出力フォーマットの調査結果

OCR の出力フォーマットは様々な形式が存在するが、仕様が公開されているものは少ない。そのため、まず現時点で仕様が公開されている hOCR と ALTO を候補として比較を行った。それぞれの形式の特徴を以下に示す。

○hOCR

- ・オープンソースの OCR ソフトウェアである OCRopus で使われているフォーマットであるが、それ以外のソフトウェアでの利用は確認されなかった。
- ・OCR の認識結果の保存よりも構造情報の保存を目指している。

○ALTO

- ・OCR の認識結果を保存するためのフォーマットであり、米国議会図書館でメンテナンスしている。
- ・文字の絶対位置を保存できる。
- ・EU 支援のもとで、ヨーロッパ各国の国立図書館が推進している、テキストのデジタル化プロジェクト「IMPACT」で採用されるなど、広く利用されている。
- ・海外では、ABBYY などの商用 OCR ソフトウェアでも広くサポートされている。
- ・シンプルかつ標準的な規約であるため、将来 JEITA¹⁵などで日本独自の規約を策定した場合でも比較的容易に乗り換えられる可能性がある。

(2) 評価結果

上記の調査結果から、以下に示す理由により、本実証実験では、ALTO を OCR 出力フォーマットのベースとして用いることとした。

- ・広く利用されている。
- ・米国議会図書館がフォーマットをメンテナンスしている。
- ・文字の絶対位置を保存することができる。

ただし、日本語対応のために、ALTO では不足している機能も存在することから、以下に示す独自の拡張を行った。

- ・文字単位の認識結果の保存
- ・文字単位の認識文字候補の保存
- ・縦書きへの対応

なお、日本語特有の構造項目であるルビに関しても ALTO の拡張が必要になるが、本実証実験では独自の拡張は行わず、構造化システムのメタデータにルビの情報を保持することにした。そもそも ALTO には構造情報の記述能力がないため、本実証実験で必要となる構造情報はメタデータ形式で保持することにした。この理由は、IMPACT プロジェクトの成果を踏まえた拡張案において、構造化情報を記述するた

¹⁵ 一般社団法人 電子情報技術産業協会 (JEITA: Japan Electronics and Information Technology Industries Association) の略。IEC (国際電気標準会議) および ISO (国際標準化機構) における国際標準化活動、JIS 規格の作成協力、JEITA 規格類制定などの標準化事業の推進等、標準化を推進する活動を行っている。

めのタグを新たに用意するという動向があり、今後 ALTO の構造情報保持能力が向上する可能性があるため、現時点で独自に拡張を行うことを避けたためである。

これらの変更を加えた形式を「拡張 ALTO」と呼ぶこととし、本実証実験の OCR 出力フォーマットとした。この「拡張 ALTO」は、以下に示す理由により、本実証実験における OCR 出力フォーマットとして適切であると言える。

- ・ XML 形式で文書校正を行うことにより、情報が構造化され、管理が容易になる。
- ・ 校正完了後に目次・索引の自動生成が可能となる。
- ・ 将来予測される外部とのデータ交換や異なるフォーマットへの変換の際のデータの可搬性を確保できる。

(3) 今後の技術的課題

現在、日本の OCR 出力フォーマットは、OCR ベンダが独自に策定し、仕様が公開されていない。OCR データの相互運用性を確保するためには、OCR 出力フォーマットの標準化を進めることが課題となる。本実証実験で利用した拡張 ALTO は、公開されたフォーマットの拡張であることから、標準フォーマットとして適していると考えられるが、ルビの記述ができないなど、日本語に対応した十分な拡張ができていない。

標準化に際しては、日本語に対応した OCR フォーマットとしてどのような拡張が必要となるのかを十分に検討する必要がある。本実証実験では、縦書きとルビの対応について拡張を検討したが、日本語の組版には縦中横、割注など様々な要素がある。書籍の全文テキストデータに必要な情報を出版社や印刷会社等の関係者と検討のうえ、どの要素に対応するかを決める必要がある。

3. 1. 2 構造化メタデータの評価

構造化メタデータに必要な情報について、文章を構成する「文」と「書式」を分離した上で、書式について位置情報と意味情報に分けて整理し、構造化のための位置情報と意味情報の規定方法について検討した。また、技術的な課題を整理した。

(1) 構造化のためのメタデータフォーマットの検討と評価

書籍の構造情報をメタデータとして保持できるよう、構造化のためのメタデータフォーマットの検討を行った。構造情報とは、文字フォントや行間、段落構成、本文、見出し、注といった論理的なカテゴリなどである。一般にメタデータのフォーマットは、以下のような基本情報から構成される。

- ・アドレッシング（メタデータを付加する対象となる部分を指定する位置情報）
- ・セマンティクス（見出し、柱¹⁶など付加する意味情報）

セマンティクスについては、見出しや柱などのような構造化の種類を設けるかを検討する必要があるが、フォーマットとしては意味の種類情報（Data type）と意味の補足情報（Data description）を定義すればよいため、すぐに確定した。一方、アドレッシングについては、メタデータを付与する対象のデータフォーマットによって、指定方法を考える必要がある。本実証実験でメタデータ付与の対象となりうるデータフォーマットとしては、OCR 出力フォーマット、スキャンした画像データ、出版社から提供された電子データファイルの3種類があるが、位置情報を指定するためには、文字校正作業によって変更が加わる可能性が低いものが望ましい。

本実証実験では、上記を踏まえて、スキャンした画像データにメタデータを付与する方式を採用した。具体的には、画像のうち、対象となるテキスト領域を矩形で囲んで指定した。

本実証実験では、構造化作業として、画面に表示された書籍の画像データに、構造情報を付与する方式を採用し、全文テキスト化作業を行うことができた。

（2）今後の技術的課題

文字が位置情報を持たないリフロー型¹⁷の書籍の電子データに対する、メタデータの付与の仕方が課題となる。リフロー型フォーマットは、今後増加が見込まれる書籍の電子データのフォーマットであり、書籍の構造や読み上げ順序を設定するための、基準位置の割り当てが大きな問題となる。本実証実験では、リフロー型の書籍の電子データに対しても、構造情報を付与するために、PDF フォーマットに変換することで、相対的な位置情報を設定した。この方法も参考にした上で、海外の事例調査等を実施しつつ、各種の構造化メタデータの付与方法を決定する必要がある。

3. 1. 3 OCR の再学習機能の評価

1つの書籍を前半と後半に分け、前半の校正結果をOCRに再学習させて、再学習前後における認識率の変化を測定し比較することで、OCRの再学習機能の有効性の検証を行った。また、技術的な課題を整理した。

¹⁶ 書籍や雑誌など冊子形式の印刷物の紙面において、内容が印刷される範囲周囲の余白（マージン）に配置される、書名・章・節・内容の要点などを記した文字列のこと。

¹⁷ 利用者による文字サイズの変更によりページのレイアウトが自動的に変更されるようにした電子書籍の実現方式。これに対して、レイアウトを固定した実現方式をレイアウト型と呼ばれる。

(1) 評価結果

本実証実験では、明治期、大正期など比較的古い書籍を中心に OCR 処理を実施した。これらの時代では、書籍ごとに使用される字体などの特徴が異なるため、書籍ごとに辞書を作成した上で、対象書籍を前半と後半の半分に分け、前半の OCR 処理と校正処理の結果を再学習させて、後半の OCR 処理に活用した。書籍の前半部分で校正された文字は、すべて辞書に登録することとした。書籍ごとの OCR 認識率の変化を図 3-1 に、書籍の時代別の OCR 再学習効果の比較を図 3-2 に示す。図 3-1 では、OCR 再学習前の認識率の低い書籍を左から順番に並べている。なお、グラフ中の全文 ID とは書籍別に固有に振られた番号を示す。「全文 ID」については、21 ページ、表 2-1 を参照のこと。

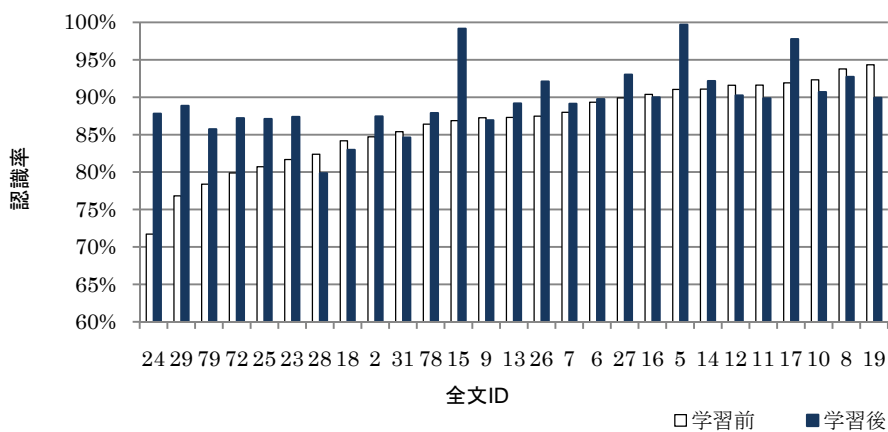


図 3-1 テキスト化システムの OCR 再学習前後の認識率

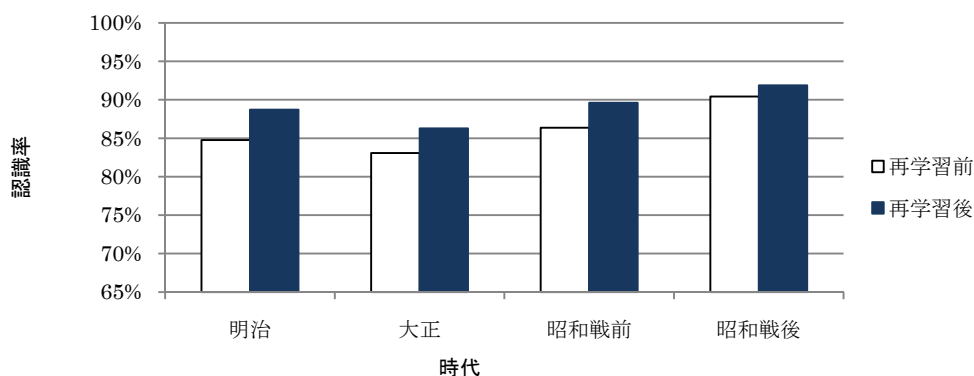


図 3-2 書籍の時代別の OCR 再学習前後の認識率

図 3-1 では、多くの書籍で再学習後の認識率が再学習前を上回っており、OCR

の再学習効果が有効であることを確認した。

また、図 3-2 からは、古い書籍ほど OCR 再学習の効果が高いことがわかる。古い書籍には特殊なフォントが多く、再学習により認識可能な文字が大きく増加したためであると推察される。

(2) 今後の技術的課題

本実証実験では、書籍の前半部分で校正された文字は、すべて辞書に登録することとしたが、OCR 再学習によって認識率が下がる場合があることから、元々 OCR の認識率が高い書籍に対しては、効果がある文字だけ登録する等の工夫が必要であると考えられる。また、本実証実験では、書籍ごとに OCR の再学習をしたが、すべての書籍に対して横断的に学習結果を共有し OCR の辞書を強化していくことも有効である。ただし、辞書の強化にあたっては、単純に校正された文字を蓄積していくのではなく、文字の誤認識の傾向を踏まえながら人手を介したメンテナンスが必要になると考えられる。また、学習用辞書のもととなっている OCR の誤認識に関する情報は、OCR ソフトウェアの辞書を強化するために有効な情報であり、OCR 技術の進展にも寄与するものと考えられる。

3. 1. 4 出力フォーマットの評価

全文テキスト化された書籍の出力フォーマットに、どのフォーマットを採用するかについて、ウェブ等で公開されている既存の書籍の電子フォーマットの仕様等を元に調査し、選定した。また、技術的な課題を整理した。

(1) 出力フォーマットの検討と評価

全文テキスト化された書籍の出力フォーマットを検討するにあたって、まず既存の書籍の電子フォーマットの調査を行った。既存の書籍の電子フォーマットとしては、PDF、EPUB、.book、X MDF、AZW、DAISY など種々のフォーマットが存在する。主な書籍の電子フォーマットの特長を表 3-1 に示す。

表 3-1 主な書籍の電子フォーマットとその特長

フォーマット名称	特長
PDF (Portable Document Format)	Adobe Systems が開発したフォーマット。作成した文書のオリジナルのレイアウトやフォントをコンピュータの環境によらず正確に再現できることが可能。PDF で文字情報を扱うには、OCR により認識した文字をテキストに変換し、画像の文字部分に該当するテキストを重ねる形で配置する。ISO 32000-1 として標準化された。
EPUB (Electronic Publication)	アメリカの電子書籍団体である IDPF (International Digital Publishing Forum) が発表したフォーマット。XML をベースとした規格で、XHTML などで作成したコンテンツを画像や CSS などとともに ZIP 形式で圧縮する。オープンなフォーマットであること、Web 制作との親和性が高いこと、テキスト系コンテンツに適していることが特長。現状では日本語の縦書きやルビに対応していないが、2011 年 5 月に日本語対応の仕様を確定する予定。

フォーマット名称	特長
TTX、.book	ボージャー社が開発した商業出版向けのフォーマット。オーサリングツール18などでの編集に用いられる記述フォーマットが TTX、配信時に用いられるのが.book。TTX は HTML 形式でテキストを中心に扱うことが可能。日本語表現への対応、段組み、ページサイズの変更、画像に対するテキストの回り込み表記、欧文向け機能、リンクジャンプ、しおり、付箋、文字サイズの拡大や資料の読上げ等視覚障がい者への対応などの機能を持ち、原稿イメージに近いレイアウトが再現可能。また、TTX は XMDF に容易に変換可能。
XMDF (ever-eXtending Mobile Document Format)	シャープが、主に PDA (携帯情報端末)、PC 向けのフォーマットとして開発。XML に基づいて作成され、国際標準規格となっている。携帯電話端末用の書籍を中心に日本国内で多くの電子書籍が販売されている。日本語表現への対応、段落、インデント、画像に対するテキストの回り込み表現、欧文向け機能、リンクジャンプ、画像上の特定の場所をクリックすると別のページにリンクするクリッカブルマップ、音声再生などの機能を持ち、原稿イメージに近いレイアウトが再現可能。
AZW	Amazon.com が採用している Amazon Kindle 専用のフォーマット。現状では縦書きやルビなどの日本語表記には対応していない。
DAISY (Digital Accessible Information System)	DAISY コンソーシアムが開発を行っている、主に視覚障がい者等が利用するデジタル録音図書のフォーマット。米国の標準規格 ANSI/NISO Z39.86-2005 として無償で公開。XML をベースにしており、日本で主流となっている規格は DAISY 2.02。コンソーシアム公認のオーサリングツールでデジタル図書を作成可能。音声だけでなく、テキスト、画像を含むマルチメディアに対応しており、世界で共通して使えるユニバーサルデザイン。

電子書籍の構成要素には、図 3-3 のように「テキスト」「構造情報」「スタイル情報」がある。

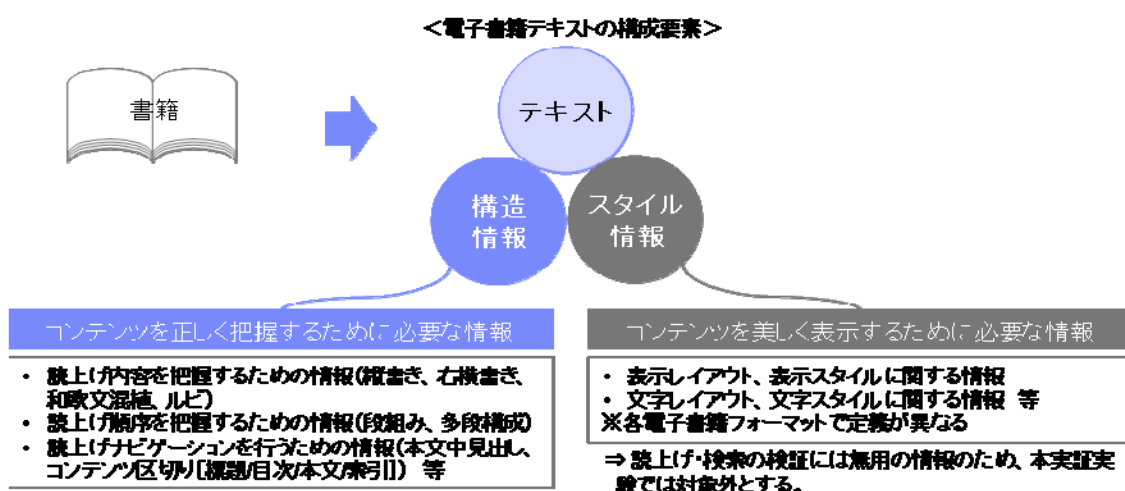


図 3-3 電子書籍テキストの構成要素

このうち、スタイル情報については読み上げ・検索の検証には必要がない情報のため、本実証実験では対象外とした。

18 文字や画像、音声、動画といったデータを編集して一本のソフトウェアや作品を作るためのアプリケーション。

しかし、電子書籍の表示においては、利用者は原文と同じ再現性を求めている。調査対象とした書籍の電子フォーマットのうち、PDF は画像を背景に持っており、原文と同じ再現性を求める利用者のニーズに応えることができる。よって、本実証実験では、本文表示画面での原文の表示用途に PDF フォーマットを採用した。

また、構造情報は、書籍の電子フォーマットにより保持可能な情報が異なる。このうち、DAISY は、テキスト情報、検索・表示と読上げに必要となる構造情報に加え、音声や画像を含むマルチメディアデータにも対応したフォーマットである。このため、視覚障がい者のほかに学習障がい者、知的障がい者、高齢者等を幅広く支援できるという特長がある。欧米では電子教科書や電子書籍のフォーマットとしても採用されている。このほか DAISY には以下の特長がある。

- ・情報の交換や送付が容易である。
- ・リンクジャンプや、しおり機能を備える。
- ・読んでいる場所がハイライト表示されるので、読み飛ばしを防ぐことができる。
- ・頭出し、検索、文字の拡大が容易である。
- ・数式は MathML¹⁹でサポートしており、図表も SVG²⁰での対応を予定している
- ・「標題、目次、本文、索引、その他」といったページ種別を用いて各ページの属性を分類し、構造化することができる。
- ・DAISY コンソーシアムと IDPF は、それぞれ次のバージョンの DAISY 規格と EPUB 規格において、以下のような互換性を持つことを合意している²¹。
 - －出版社が EPUB で出版すれば、すぐにテキスト DAISY と同様に専用端末で読むことができる。
 - －肉声や手話が必要な時は EPUB ファイルに朗読や動画を追加してマルチメディア DAISY として利用できる。

以上をふまえ、本実証実験では、検索・読上げサービスの用途に DAISY フォーマットを採用した。

(2) 今後の課題

書籍の電子フォーマットには、前述のように多くのものが存在する。各フォーマット

¹⁹ Mathematical Markup Language の略。数式を表すためのマークアップ言語。

²⁰ Scalable Vector Graphics の略。XML ベースの画像記述言語。

²¹ “DAISY と EPUB は読書のユニバーサルデザインをどう実現するのか”。障害保健福祉研究システム。

<http://www.dinf.ne.jp/doc/japanese/access/daisy/seminar100709/index.html>,

ットの動向にも留意して、全文テキスト化された書籍の出力フォーマットとして適切なものを比較検討していくことが必要である。

3. 1. 5 出版社データから OCR 出力フォーマットへの変換の評価

出版社から提供された書籍の電子データを、テキスト化システムで利用可能な OCR 出力フォーマットに変換する機能について、処理時間を測定するとともに、技術的な課題を整理した。

(1) 評価結果

出版社から提供された書籍の電子データを、テキスト化システムで利用可能な OCR 出力フォーマットに変換する機能について、まずはどの書籍の電子フォーマットを変換対象にするかを検討した。

本実証実験では、期間内で最大限のテキスト化を行うため、冊数の多い PDF、TEXT (フラットテキスト)、XMDF (TTX 有)、.book (TTX 有) の 4 フォーマットをテキスト化対象とした (XMDF、.book は TTX 有のものを対象にしたので TTX に対応しさえすればよく、実質的には 3 フォーマットといえる)。まず、対象書籍を PDF ファイルに変換した後に OCR 出力フォーマットに変換した。対象書籍 333 タイトル中 1 タイトルについては、編集プロテクトがかかっているため PDF に変換できなかった。また、5 タイトルについては、ほとんど画像のみでテキスト情報が少ない PDF ファイルであったため、最終的に 327 タイトルを OCR 出力フォーマットに変換した。

一部の書籍については、フォーマット変換時に、以下に示すような手作業が必要となった。

- 複数のファイルから構成される書籍について、手作業でファイルを一つにまとめた。
- TTX フォーマットから、HTML フォーマット、PDF フォーマットを経て OCR 出力フォーマットに変換するなど、複数のフォーマット変換ツールを使用する場合、ツール間のデータ受け渡しを手作業で行った。

このような手作業の例として、TTX フォーマットから HTML フォーマットを経て PDF フォーマットに変換した 58 タイトル分の作業には、全体で 310 分を要した。これは 1 タイトルあたりのフォーマット変換処理に約 5 分を要していることを示している。

(2) 今後の技術的課題

本実証実験では、出版社からの提供データのうち、冊数の多い4種類のフォーマットに対してフォーマットを変換した。本実証実験ではテキスト化対象外としたその他のフォーマット（XML、.book（TTX 無）、indd、eps）についても、今後変換の手順やツールを開発することが必要である。ただし、総務省、文部科学省、経済産業省による「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に関する懇談会」の報告書にもあるように、中間（交換）フォーマット²²を共通化する方針もあるため、このフォーマットを介することにより、変換作業の簡素化が期待できる。

一方、本実証実験で使用した変換機能では、一部の書籍においてフォーマット変換時に手作業が必要になる等人間の判断が関与する部分が残っており、この部分を自動化することで効率化を図ることができる。

²² 印刷会社が保有する最終データをもとにして、様々なプラットフォーム、端末が採用する多様な閲覧ファイルフォーマットに変換対応が容易に可能となるフォーマットのこと。

3. 2 テキスト化システムを用いた作業の効率化、高度化の評価

OCR、校正、構造化を効率化、高度化するための機能等について、以下に示す評価を行った。

- ・レイアウト校正作業
- ・共同校正機能
- ・OCRの再学習機能
- ・共同構造化機能
- ・構造情報推論機能
- ・読上げ順序の編集機能

3. 2. 1 レイアウト校正作業の効率化、高度化の評価

OCRの読取精度向上のためのレイアウト校正作業について、作業時間を測定し、実用化に向けた課題を整理した。

(1) 評価結果

OCR技術は日々進歩しているが、現状では、レイアウト校正など、人手による作業が不可欠である。本実証実験では、できるだけ人手を介さず、自動的に認識精度を向上させるために専用のレイアウト校正ツールを使用し、その効果を評価した。

書籍ごとのレイアウト校正作業時間を図3-4に、時代ごとの平均レイアウト校正作業時間を図3-5にそれぞれ示す。図3-4ではレイアウト校正1万字あたりの作業時間が短い書籍を左から順番に並べている。なお、グラフ中の全文IDとは書籍別に固有に振られた番号を示す。「全文ID」については、21ページ、表2-1を参照のこと。

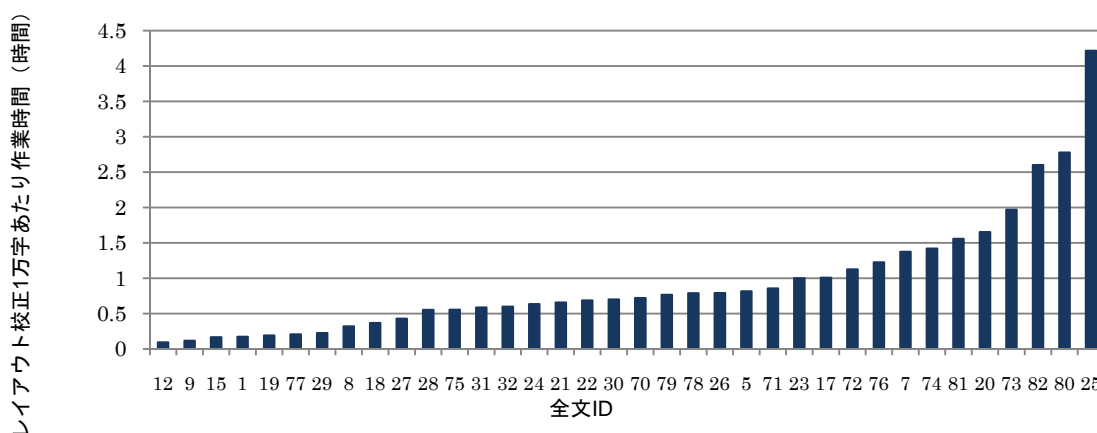


図 3-4 書籍ごとのレイアウト校正作業時間

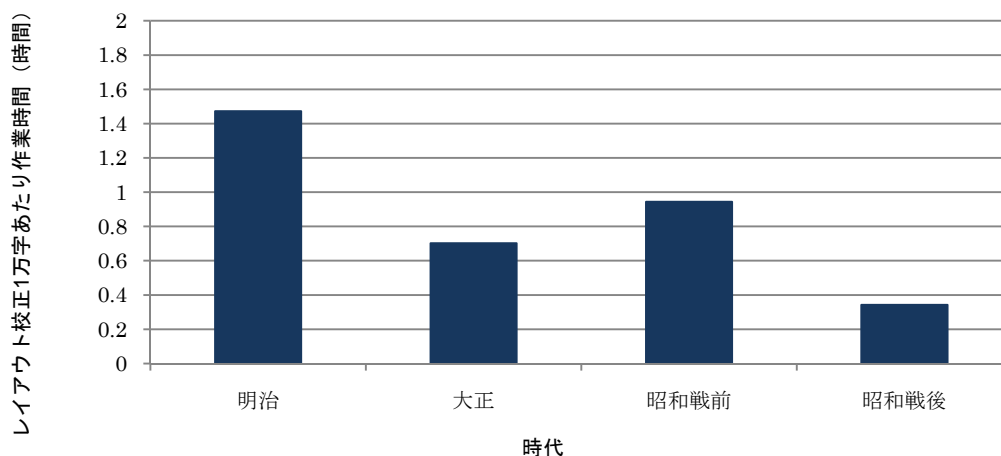


図 3-5 時代ごとのレイアウト校正作業時間

図 3-4 で、特定の書籍について、作業時間が極端に長くなっている。これは、OCR で読み取る画像の画質が悪いことに加えて、目次ページが上下 2 段組みになっている特殊なレイアウトの書籍であったことが影響していると考えられる。また、図 3-5 から、時代が古くなるほど、レイアウト校正作業に時間がかかることがわかる。これも、OCR で読み取る画像の画質の影響と、レイアウトの複雑さに起因すると考えられる。

(2) 今後の技術的課題

レイアウト校正を自動化できれば、校正作業を短縮できる。しかし、本実証実験の対象書籍である明治期～昭和戦前期の書籍では、ほぼすべてのページにおいて、校正ツールを使い手作業でレイアウト校正を行う必要があった。図 3-5 から、明治期の書籍のレイアウト校正時間が特に長いことがわかる。

本実証実験では、レイアウト校正に要する時間や手間が当初想定していた以上に大きかったことから、全文テキスト化作業を効率的に進めるためにはレイアウト校正作業を軽減する必要があることがわかった。また、古い書籍や複雑なレイアウトの書籍において、手作業でのレイアウト校正が特に必要となった。これらのことから、レイアウト認識率を向上させるための技術開発が求められるとともに、古い書籍や複雑なレイアウトの書籍専用のレイアウト認識技術の検討が必要になる。

3. 2. 2 共同校正機能による効率化、高度化の評価

OCR で読み取った文字情報を修正するための共同校正機能について、作業時間を測定するとともに、評価者にアンケート調査を行い、実用化に向けた課題を整理した。

(1) 評価結果

OCR で読み取った文字情報を修正するための共同校正機能について、作業時間を測定した。各書籍の全ページに対する共同校正作業の所要時間をアクセスログから算出し、各書籍内の文字数から 1 万文字あたりの作業時間に置き換えた結果を図 3-6 に示す。また、この結果をもとに、書籍の時代ごとに共同校正作業の平均所要時間を算出した結果を図 3-7 に示す。図 3-6 では共同文字校正 1 万字あたりの作業時間が短い書籍を左から順番に並べている。なお、グラフ中の全文 ID とは書籍別に固有に振られた番号を示す。「全文 ID」については、21 ページ、表 2-1 を参照のこと。

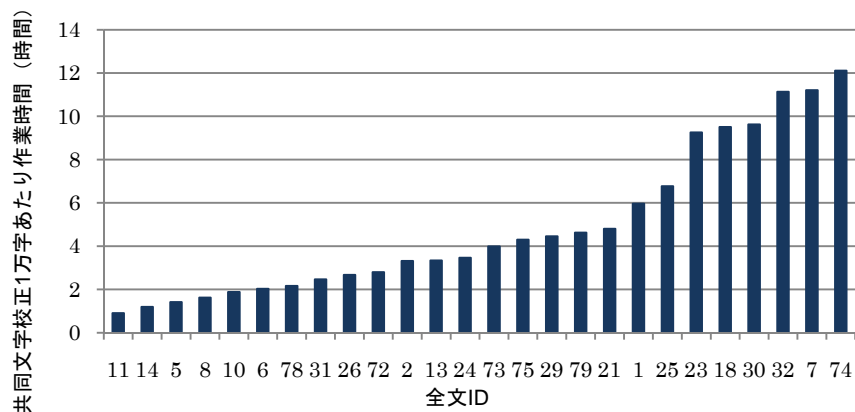


図 3-6 書籍ごとの共同文字校正作業の所要時間

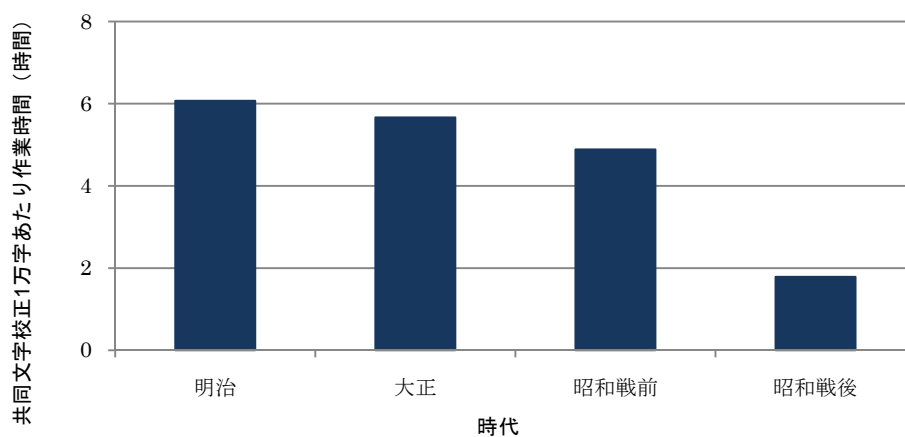


図 3-7 書籍の時代ごとの共同文字校正作業の平均所要時間

図 3-7 からは、明治～昭和戦前期の、時代が比較的古い書籍の文字校正作業に

時間がかかるとともに、書籍ごとの作業時間のばらつきが大きいことがわかる。一方、昭和戦後期の比較的新しい書籍は、作業時間も短く、ばらつきも小さい。また、複数の作業員で1冊の書籍の校正作業を進められるため、短時間で作業を完了できるようになることは自明であるが、これも共同校正機能の効果の表れである。

評価者に対して実施した、共同校正機能のアンケート結果は以下のとおりである。

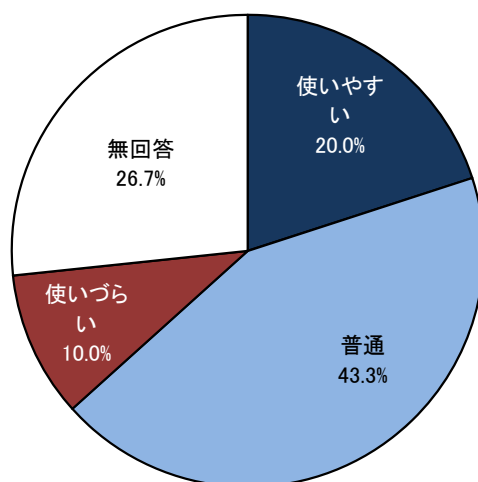


図 3-8 共同校正機能に対するアンケート結果

「使いやすい」と「普通」と回答した評価者が60%以上を占めている。すべての評価者にとって、テキスト化システムの共同構成機能は、初めて利用する機能であるが、簡単な操作説明を行えば、ほぼ問題なく利用できることが明らかとなった。

共同校正機能に対しては、以下のような評価を得た。

○共同文字校正機能

- ・校正する文字にマウスカーソルを重ねることで、前後の文字を含む画像を表示する機能は便利であり効率化に資する。
- ・共同文字校正作業は機能がわかりやすく、作業も簡単である。

○共同仕上げ校正機能

- ・レイアウトがわかりやすい。
- ・元の書籍と同じ表示状態で、校正作業ができる点が良い。
- ・前工程の共同校正機能での校正ステータスごとに、文字の囲みの色が異なる点わかりやすい。

(2) 今後の技術的課題

共同校正機能に関する主な指摘・改善事項を以下に示す。

○共同文字校正機能

- ・マウスカーソルを重ねて表示する範囲を拡大し、1行単位で表示すれば、前後の文脈を見ながら判断することが可能となり、作業効率が一層向上する。
- ・OCR 出力フォーマットの持つ文字認識の確信度を活用して、校正対象の文字を減らし、作業を効率化できるのではないか。
- ・作業している書籍の作業がどれだけ進んでいるのかわからない。

○共同仕上げ校正機能

- ・それぞれの画面の機能の説明が不足している。
- ・認識されていない文字が修正できない。

前述のアンケート結果からも明らかなように、共同校正機能は分かりやすい操作を採用しており、校正スキルの少ない一般の利用者であっても、十分に必要な作業を行うことができた。特に、共同文字校正機能は、OCRにより同じ文字として認識した文字の画像を一覧表示し、誤認識している場合には正しい文字を入力するという、きわめてシンプルな操作であった。ただし、操作がシンプルであるが故に、作業が単調になり、長時間作業をしていると、飽きが来ることも考えられる。評価者の指摘にもあるように、作業の進捗状況を合わせて表示することで、作業者のモチベーションを維持する効果が期待できる。

ただし、操作がシンプルであったとしても、最低限の機能の説明を画面に表示して、利用者が操作に迷わないようにする配慮は必要である。また、同じ文字や似たような文字を繰り返し見ること、意識が散漫になり、校正ミスを起こす可能性も考えられる。書籍の1行単位など、作業者が普通に文章を読むのと同じような感覚で文字を校正するようなインタフェースを実現することで、作業の単調さを解消し、校正ミスも防止できる可能性がある。

3. 2. 3 OCRの再学習機能による効率化、高度化の評価

1つの書籍を前半と後半に分け、前半の校正結果をOCRに再学習させて、OCRの認識率を向上させた場合に、後工程の校正作業時間を測定し、実用化に向けた課題を整理した。

(1) 評価結果

3つの書籍に対して、作業者を1人に特定した上で、書籍を前半30ページと後半30ページに分け、書籍の前半の校正結果を学習させて後半の校正作業を行い、前半

と後半とでシステムのアクセスログの集計から得られる作業時間の違いを比較した。作業時間については、1 ページあたりの作業時間と 1 万字あたりの作業時間を算出した。その結果を表 3-2 に示す。なお、「年代 ID」については、21 ページ、表 2-1 を参照のこと。

表 3-2 書籍ごとの共同文字校正作業時間

[年代 ID] 書籍名		ページ数	文字数	作業時間 (分)	時間/1 万字
[明治 07] 初等代数学	前半	30	8,248	115	2.32
	後半	30	8,772	80	1.52
[大正 07] 政治学研究	前半	30	7,131	90	2.10
	後半	30	10,127	120	1.97
[大正 08] 変態心理学講義録. 第 1 篇	前半	30	7,888	210	4.44
	後半	30	11,101	150	2.25

1 万字あたりの作業時間の前後半を比較すると、いずれの書籍においても後半の処理時間の方が短くなっており、OCR の再学習効果があると言える。

(2) 技術的課題

本実証実験では実現できなかったが、1 つの書籍だけでなく、複数の書籍に対して再学習機能が有効に機能するかどうかを検証することが望まれる。

3. 2. 4 共同構造化機能による効率化、高度化の評価

ページ種別やページ内の見出し、ページ番号などの構造情報を付与するための、共同構造化機能について、作業時間を測定するとともに、評価者にアンケート調査を行い、実用化に向けた課題を整理した。

(1) 評価結果

各書籍の全ページに対する、構造化作業の所要時間をアクセスログから算出し、各書籍内の文字数から 1 万文字あたりの作業時間を算出した。国立国会図書館所蔵資料に対する処理結果を図 3-9 に、出版社提供データに対する処理結果を図 3-10 にそれぞれ示す。なお、国立国会図書館所蔵資料の構造化作業には読上げ順序設定作業を含むが、出版社提供データの構造化作業には読上げ順序設定作業を含まない。この図では共同構造化 1 万字あたりの作業時間が短い書籍を左から順番に並べている。グラフ中の全文 ID とは書籍別に固有に振られた番号を示す。「全文 ID」については、21 ページ、表 2-1 を参照のこと。(ただし、全文 ID には出版社提供データの ID (3-4、83-104) を含む。以下同じ。)

また、この結果をもとに、書籍の時代ごとに共同構造化作業の平均所要時間を算

出した結果を図 3-11 に示す。

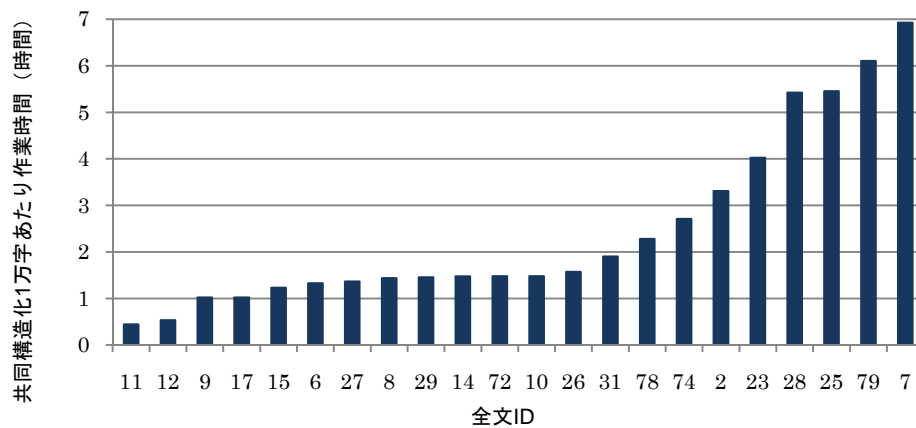


図 3-9 国立国会図書館所蔵資料に対する書籍ごとの共同構造化作業の所要時間

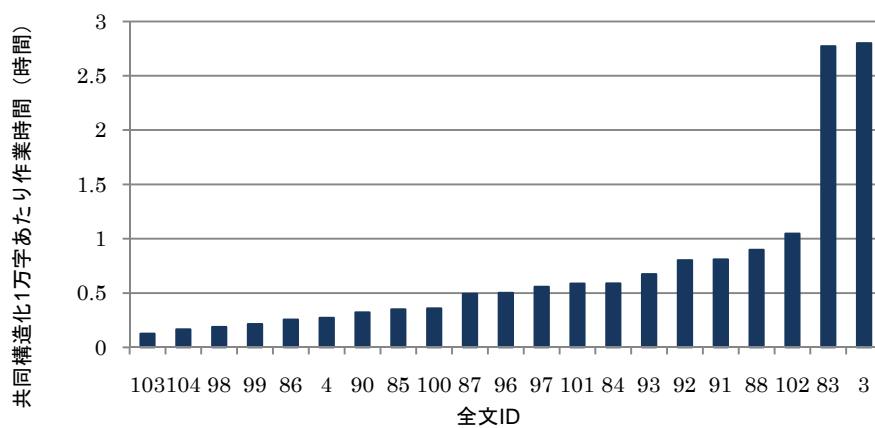


図 3-10 出版社提供データに対する書籍ごとの共同構造化作業の所要時間

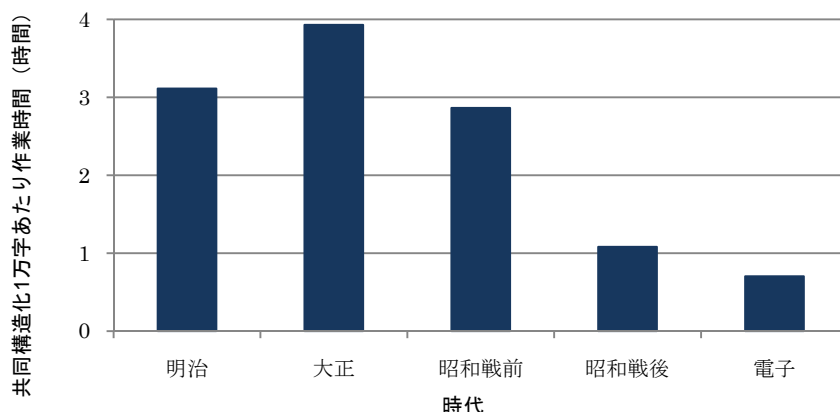


図 3-1 1 書籍の時代ごとの共同構造化作業の平均所要時間

共同構造化作業に関しても、図 3-1 1 から明らかなように、明治～昭和戦前期の比較的古い書籍に対する作業時間が、昭和戦後期の比較的新しい書籍よりも長くなっている。明治期の書籍が大正期の書籍よりも作業時間が短い理由は、書籍の構造が平易なものが存在したことによるものと考えられる。一方、昭和戦後期の比較的新しい書籍については、作業時間も短く、作業時間の分散も大きくないことから、構造が複雑でなく、作業が容易であったと考えられる。

出版社提供データのうち、一部の書籍では作業時間が長くなったが、その他の書籍については、おおむね 1 時間以内で作業を終えており、昭和戦後期の所蔵資料と比べても短時間で作業が完了している。

さらに、4 つの書籍に対して、作業者を 1 人に特定した上で、構造化項目別の 100 ページあたりの作業時間を算出した。この結果を表 3-2 に示す。なお、「年代 ID」については、21 ページ、表 2-1 を参照のこと。

表 3-3 100 ページあたりの構造化項目別の作業時間

構造化項目 \ 年代 ID	[大正 03] (hh:mm:ss)	[大正 04] (hh:mm:ss)	[明治 09] (hh:mm:ss)	[明治 11] (hh:mm:ss)	平均 (hh:mm:ss)
ページ種別	0:03:42	0:08:30	0:08:13	0:05:20	0:06:25.25
柱	—	0:24:02	0:19:33	—	0:21:47.50
ページ番号	0:20:02	0:25:32	0:16:10	0:15:10	0:19:13.50
目次項目	0:16:04	0:19:13	—	—	0:17:38.50
見出し	0:14:49	0:13:22	0:10:27	—	0:12:52.67
目次リンク	0:10:39	0:24:28	—	—	0:17:33.50
その他	0:02:47	0:15:48	0:22:00	—	0:13:31.67
合計	1:08:03	2:10:55	1:16:23	0:20:30	1:13:57.75
合計 (その他除く)	1:05:16	1:55:07	0:54:23	0:20:30	1:03:49.00

構造化項目の中で最も平均作業時間が短いページ種別の設定は、リストから種別を選択するというシンプルな方式で実装した。これにより、構造化の知識の少ない一般の作業者であっても、事前に操作説明を実施することで、作業が可能であると考えられる。

評価者に対して実施した、共同構造化機能のアンケート結果は以下のとおりである。

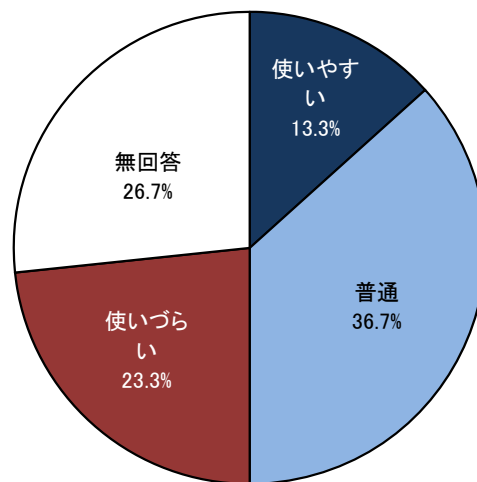


図 3-1 2 共同構造化機能に対するアンケート結果

アンケート結果から、「使いづらい」と「普通」と回答した評価者を合わせると60%となっている。詳細な機能でみた場合には、目次と見出しの構造化に対する評価が低い。本文に出現する見出しの文字と、目次の文字列を関連付ける作業については、テキスト化システムの中でも操作が難しく、作業者の思いどおりにならないケースが散見されたため、評価が低くなったと考えられる。

(2) 技術的課題

共同構造化機能に関する技術的課題を以下に示す。

- ・構造化作業のルール化
- ・利便性の向上（大型ディスプレイの採用、動作の高速化など）

共同構造化作業のルールについては、目次の階層化をどこまで行うのか、柱の概念の定義、ページ種別の判断基準の設定など、作業上の取決めを明確にした上で作

業を進めないと、作業者の負荷が上がり、結果がばらついてしまう。共同校正作業と同じく、作業上のルールを定義することが必要である。

本実証実験ではノート PC を利用して作業を行ったことから、ディスプレイが小さく文字表示も小さかったため、結果として構造化作業が煩雑で、作業の全体像が把握しづらかったと想定される。また、テキスト化システムが Web ブラウザで操作する方式であり、サーバに負荷が集中したため、動作が遅くなったと考えられる。実運用時には、大きなディスプレイを用意したり、十分な通信回線を確保するなど、作業環境にも配慮が必要である。

3. 2. 5 構造情報推論機能による効率化、高度化の評価

構造化作業の効率化を図るための構造情報推論機能について、テキスト化システムが生成した構造情報の割合を測定するとともに、評価者にアンケート調査を行い、実用化に向けた課題を整理した。

(1) 評価結果

構造情報推論機能の効果を計測するために、構造化メタデータ全体の中で推論機能によって生成されたメタデータが占める割合（以下、「構造推論メタデータ生成率」という）を算出した。この結果を図 3-1 3 に示す。この図では推論が生成されたメタデータの割合が低い書籍を左から順番に並べている。なお、グラフ中の全文 ID とは書籍別に固有に振られた番号を示す。「全文 ID」については、21 ページ、表 2-1 を参照のこと。

この結果をもとに、書籍の時代ごとに構造推論メタデータ生成率の平均を算出した結果を図 3-1 4 に示す。

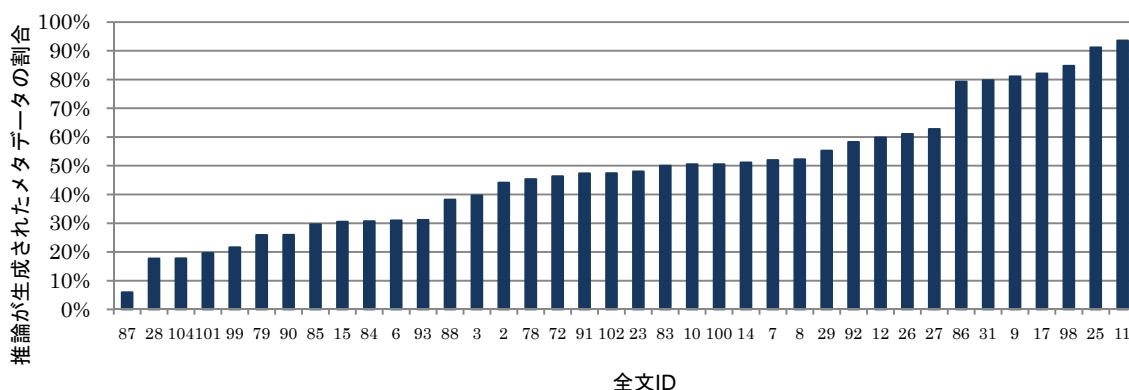


図 3-1 3 書籍ごとの構造推論メタデータ生成率

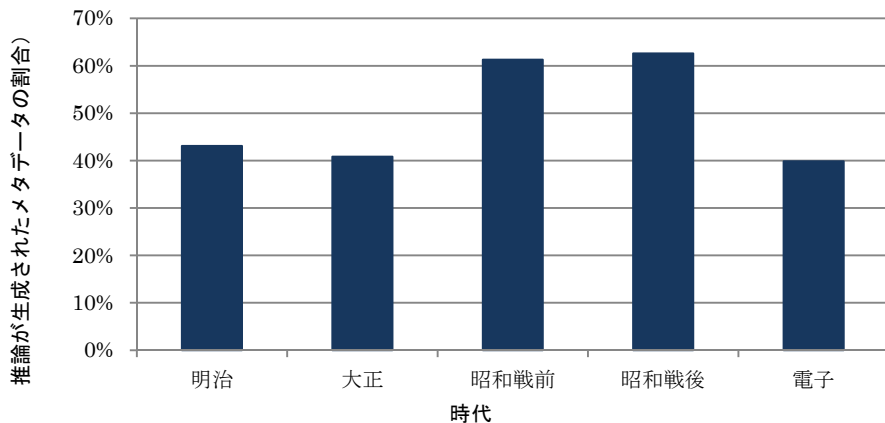


図 3-1 4 書籍の時代ごとの構造推論メタデータ生成率の平均

図 3-1 3 および図 3-1 4 から、明治～大正の古い書籍は、構造推論メタデータ生成率が低い。また分散が小さいことから、全般的に構造推論メタデータ生成率が低い値にとどまっていることがわかる。一方、昭和戦前～戦後期の書籍、および出版社の電子データは、構造化メタデータ生成率が高い場合と低い場合とのばらつきがある。構造推論機能が合致する場合とそうでない場合の差が大きいことがわかる。

また、各書籍における、見出し、柱、ページ番号、目次項目の各構造化項目に対する構造推論メタデータ生成率をそれぞれ図 3-1 5、図 3-1 6、図 3-1 7、図 3-1 8 に示す。この図では推論が生成されたメタデータの割合が低い書籍を左から順番に並べている。なお、グラフ中の全文 ID とは書籍別に固有に振られた番号を示す。「全文 ID」については、21 ページ、表 2-1 を参照のこと。

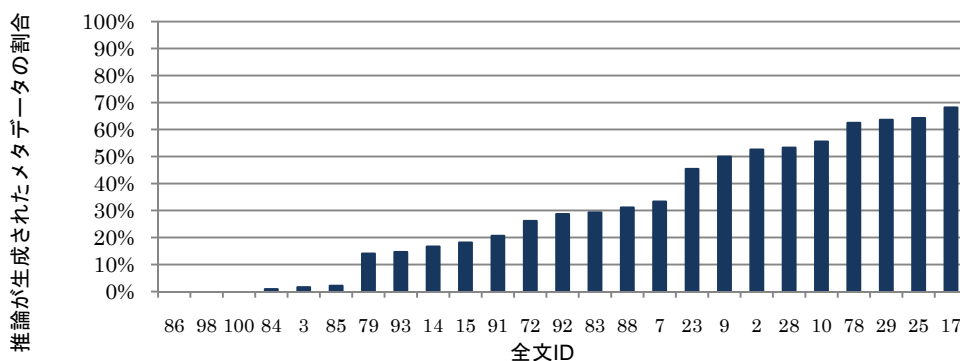


図 3-1 5 見出しに対する構造推論メタデータ生成率

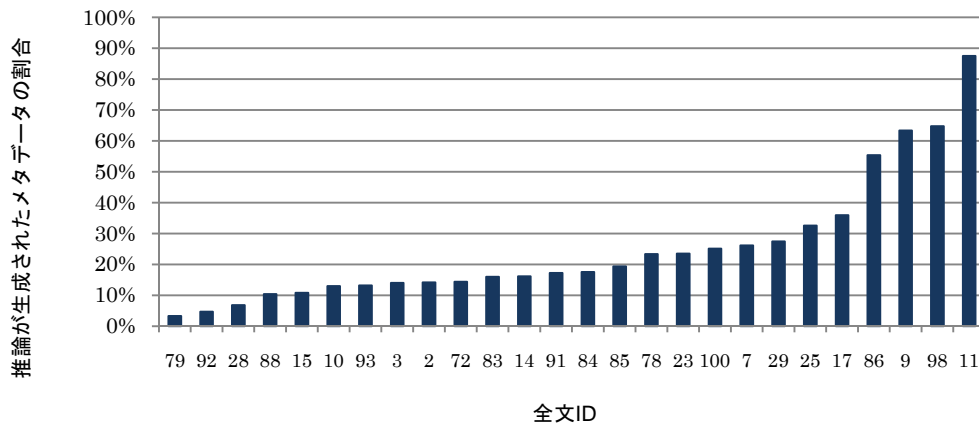


図 3-16 柱に対する構造推論メタデータ生成率

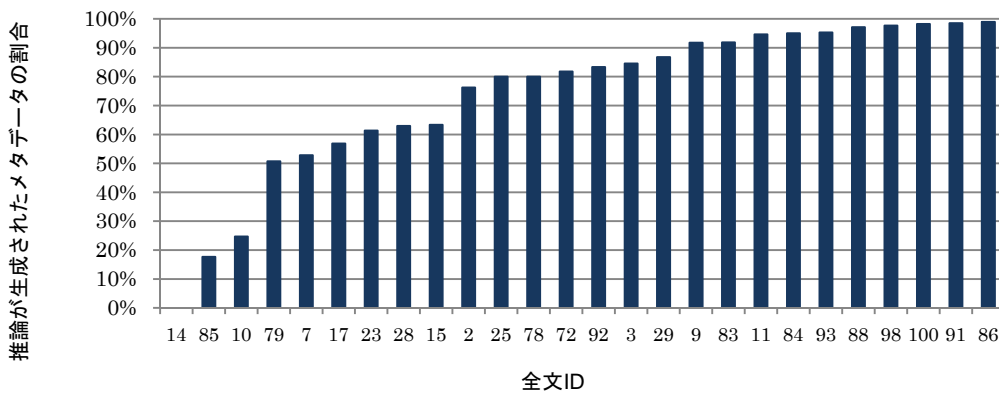


図 3-17 ページ番号に対する構造推論メタデータ生成率

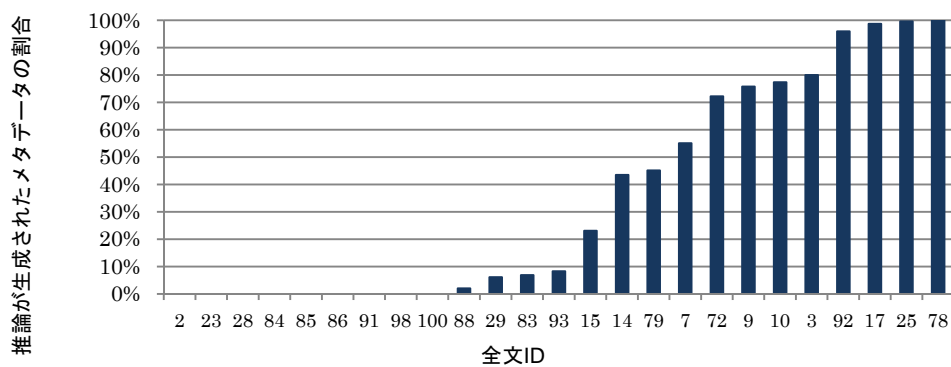


図 3-18 目次項目に対する構造推論メタデータ生成率

なお、テキスト化システムでは、構造情報推論機能の1つとして読上げ順序の推

論機能も実装した。この機能の効果を測定するため、1人の作業員に対して、3つの書籍の各5ページ分を、読上げ順序の推論機能を用いない場合と用いた場合とで2回作業を行い、その作業時間を測定した。なお、「年代ID」については、21ページ、表2-1を参照のこと。

表 3-4 読上げ順序の推論機能の利用有無による作業時間の比較

[年代ID] 書籍名	段組数	作業ページ	手動 [h:mm:ss]	推論利用 [h:mm:ss]	削減率 [%]
[明治01] 一元哲学	1	P.21-25	0:40:27	0:01:26	96.5
[昭和戦後06] 資料日本現代史.2	2	P.46-50	0:08:27	0:05:42	32.5
[雑誌01] 文芸春秋	3	P.21-25	1:17:10	0:27:05	64.9

作業時間の比較から、読上げ推論機能が、読上げ順序の構造化作業の効率化に効果的であることが明らかとなった。

評価者に対して実施した、構造情報推論機能のアンケート結果は以下のとおりである。

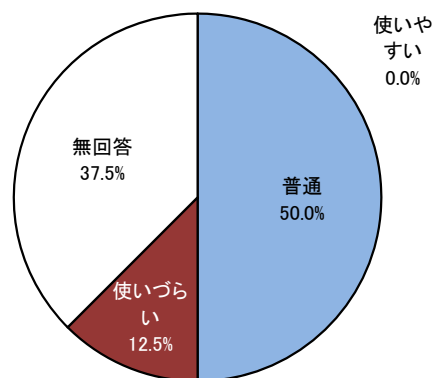


図 3-19 見出しの構造情報推論機能に関するアンケート結果

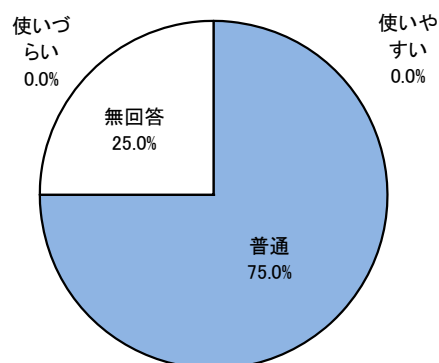


図 3-20 柱の構造情報推論機能に関するアンケート結果

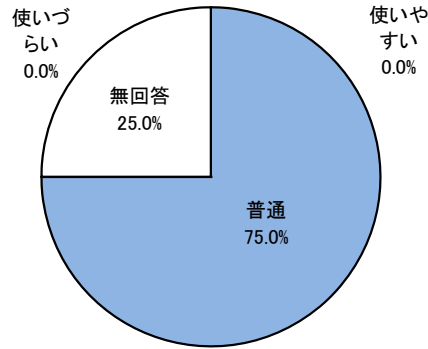


図 3-2 1 ページ番号の構造情報推論機能に関するアンケート結果

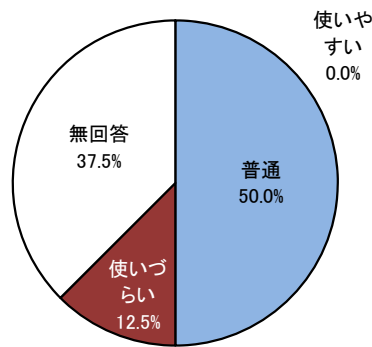


図 3-2 2 目次に関する構造情報推論機能に関するアンケート結果

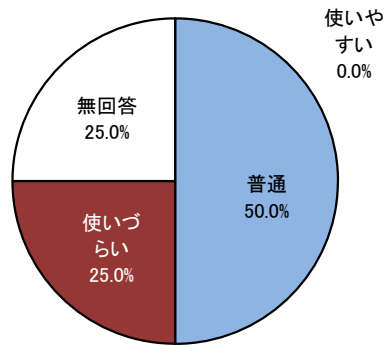


図 3-2 3 ページ種別に対する構造情報推論機能に関するアンケート結果

このアンケート結果からは、構造情報推論機能に対しては「普通」という評価がいずれについても 50%前後であった。テキスト化作業者からは、目次、見出し、柱、ページ番号の推論が効果的であったことや、柱を一度指定すると、その作業結果をもとにしてほかのページの柱も推定する点が便利であることを評価する意見があった。

(2) 技術的課題

画像データよりも電子データの方が、構造推論メタデータ生成率が高い傾向がある。電子データの方が、ページ番号の位置ずれが少ないことなどが原因として考えられる。構造推論メタデータ生成率のさらなる向上のため、画像データと電子データとで推論アルゴリズムを分けることも考えられる。

構造情報については、ページ番号であればページ余白部分に数字が配置される、柱であればページ余白の左右交互に縦書きでタイトルや章・節の見出しが配置されるというように、構造項目ごとに傾向がある。この特徴をふまえて、効果的に推論アルゴリズムを改良していくことが必要である。また、検索・表示や読上げといった、全文テキストデータの活用方法も考慮した上で、どの構造項目必要かということも、今回の実験結果をふまえて検討し、システムに組み込むことが望まれる。

3. 2. 6 読上げ順序編集機能による効率化、高度化の評価

構造化作業において、本文の文字の読上げ順序を設定する際に、視覚的な手法を採用した場合の作業時間を測定するとともに、評価者にアンケート調査を行い、実用化に向けた課題を整理した。

(1) 評価結果

まず、読上げ順序編集作業に要する時間の傾向を分析するために、本文が一段組の6つの書籍を選び、作業者をそれぞれの書籍について1人に特定して、読上げ順序編集作業を実施し、作業時間を測定した。この結果を表 3-5に示す。なお、「年代ID」については、21ページ、表 2-1を参照のこと。

表 3-5 読上げ順序編集作業の所要時間

[年代 ID] 書籍名	作業時間 (分)	ページ数	文字数	分/ ページ	時間/ 1万字
[大正 03] 能率的販賣經營法	47	102	38,639	0.46	0.20
[大正 04] 内国保険会社信用録…	19	60	22,059	0.32	0.14
[大正 07] 政治学研究	40	60	17,258	0.67	0.39
[明治 07] 初等代数学	25	60	17,020	0.42	0.24
[明治 09] 東京印刷同業組合名鑑…	19	30	6,625	0.63	0.48
[明治 11] 育嬰草	46	60	18,947	0.77	0.40
合計	196	372	120,548	0.53	0.27

評価者に対して実施した、読上げ順序編集機能のアンケート結果は以下のとおりである。

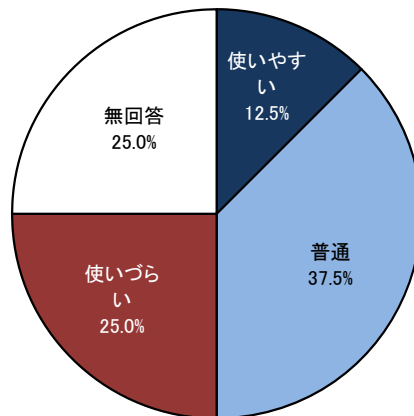


図 3-2 4 読上げ順序編集機能に関するアンケート結果

表 3-5 に示された、1 ページあたりの作業時間を見ると、平均して 30 秒程度で作業が完了しており、読上げ編集機能が作業の効率的に資するものであると言える。ただし、アンケート結果からは「使いやすい」とする意見は少なく、本機能には改善の余地がある。

(2) 今後の技術的課題

読上げ順序編集機能に対しては、縦書き、横書き混在のものを対象にした時の操作が難しいこと、読上げ順序の編集は思いとおりに操作ができず、使いづらいことが指摘された。前述の推論機能を利用することで、作業を大幅に省力化できる場合がある一方で、レイアウトが複雑なものについては、不自然な推論結果が目立ち、かえって手間がかかる場合もあった。推論アルゴリズムのさらなる改善が望まれる。

3. 3 テキストデータ作成にかかる作業時間の評価

OCR、校正、構造化の一連の作業時間を測定し、作業上の課題を整理した。

(1) 評価結果

各書籍に対して、テキスト化システムを用いてレイアウト校正、共同文字校正、仕上げ校正、共同構造化の各作業の所要時間をアクセスログから算出し、作業で処理した文字数から1万文字あたりの作業時間に置き換えた結果を図 3-25 に示す。

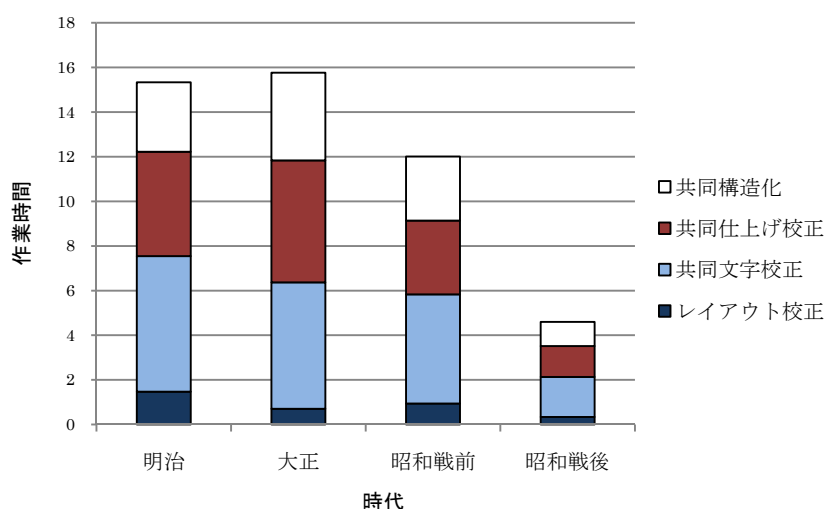


図 3-25 書籍の時代別、作業別の作業時間の積算

レイアウト校正以外の作業は、昭和戦後期の書籍が極端に短い。これは、OCRの元画像が鮮明であることにも起因すると考えられる。一方、文字校正および仕上げ校正は、昭和戦前期より古い書籍で作業時間が長い。これは、旧字が増えること、OCRの処理結果にノイズが多く含まれることが原因として考えられる。

また、本実証実験の結果を用いて、1 ページあたりの平均作業時間を試算した。この結果を表 3-6 に示す。

表 3-6 1 ページあたり概算作業時間 (試算値)

時代	レイアウト校正 (分)	共同文字校正 (分)	仕上げ校正 (分)	構造化 (分)	合計 (分)
明治	2.8	11.6	9.0	6.0	29.4
大正	1.9	15.1	14.5	10.5	42.0
昭和戦前	2.5	13.0	8.8	6.5	30.8
昭和戦後	1.5	7.9	6.1	4.8	20.3

この結果より、OCR、校正、構造化の一連の作業は、約 20 分～40 分という結果となった。

(2) 今後の課題

国立国会図書館の所蔵資料に対して、全文テキスト化を実施することを想定した場合、1 ページあたりの作業に 20～40 分の作業時間を要するのでは、実用的とは言いがたく、改善の余地が大きく残されている。これまでも述べたように、OCR による処理、文字校正、構造化の各作業に対して、OCR の読取精度向上、校正作業の一層の効率化、構造情報推論機能の高度化など、全文テキスト化の効率化の実現に資する基礎的な技術開発が望まれる。

また、本実証実験では、校正・構造化作業を複数人で行うことによる効率化の可能性の検証を目的に、テキスト化システムに共同校正機能、共同構造化機能を実装した。OCR やソフトウェアの機能向上だけでなく、インターネットの特性を生かして「人手による入力」を組み合わせることで校正・構造化作業を効率的に行う必要がある。例えば、オーストラリア国立図書館 (NLA) が行っている新聞電子化プロジェクト (Australian Newspapers Digitisation Program、ANDP)²³では、OCR で処理したデータをインターネットに公開し、これを閲覧者が修正している²⁴。また、米国議会図書館や米国国立公文書館と提携して電子化を進めている Footnote 社では、ユーザが画像内の手書き文字などを読み取り、画像内の該当する箇所に付箋を貼れるようなサービスを提供している。全文テキスト化作業にこのような手法を採用することも考えられる。

²³ <http://www.nla.gov.au/ndp/>

²⁴ Holley, Rose. “Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers”. National Library of Australia. http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf

4 全文テキストデータの検索・表示に関する実証実験に関する評価

全文テキストデータの検索と表示に関する実証実験では、以下に示す評価を実施した。また、視覚障がい者等向けの読上げサービス等については、視覚障がい者等 18 名に対して、利便性に関するアンケートおよびヒアリング調査を実施した。

- ・ 検索結果表示における全文テキストデータの活用の評価

- 目次表示、スニペット表示等

- ・ ナビゲーション・リコメンドなど高度な検索機能への全文テキストの活用の評価

- キーワード検索（構造指定検索機能・難易度検索機能）、サジェスション、自然文検索、連想検索、文脈検索、固有名表示、タグクラウド等

- ・ 全文テキストを検索対象とする場合の検索結果のランキング・表示方法の評価

- 検索結果のランキング、検索語ハイライト表示、書籍本文における検索語出現数の表示、書籍本文のテキスト表示、PDF ファイル表示、目次・本文リンク等

- ・ 視覚障がい者等向け読上げサービス等の有効性および高度化の評価

- 全文テキストデータの読上げ、全文テキスト化データの品質、OCR 認識率が読上げサービスに与える影響、DAISY ファイルの読上げ等

- ・ その他

- 全文テキストデータの処理時間の評価、文字コード対応の評価

本章では、これらの評価結果を示す。なお、上記の評価事項は評価の観点により分類しているため、以下の節からは、検索表示システムの画面構成に沿って説明する。

4. 1 検索画面における機能の評価

評価方法として、一般利用者 11 名、有識者 8 名に出版社を加えた評価者に対して、機能や表示の利便性に関するアンケートおよびヒアリング調査を実施した。検索画面の評価として、全文テキストデータを活用した、以下の機能や表示の利便性を評価した。

- ・ キーワード検索（構造指定検索機能・難易度検索機能）
- ・ 自然文検索
- ・ サジェスション

4. 1. 1 キーワード検索（構造指定検索機能・難易度検索機能）の評価

（1）評価結果

キーワード検索の詳細検索に実装した、書籍の目次・本文・索引という構造単位で検索対象を限定して検索できる構造指定検索機能、書籍の内容の難しさを小学 1～3 年、小学 4～6 年、中学、高校、大学・一般の 5 つのレベルから指定して検索できる難易度検索機能について、ヒアリングでは以下の評価を受けた。

- ・本文だけでなく、目次や索引なども検索対象として指定できることが、国立国会図書館の既存の検索サービスと比較して新しいものと感じる。
- ・難易度検索は、検索結果が適切でなく、検索条件にはふさわしくない。

構造指定検索機能については新規性を評価されたが、難易度検索は、大学・一般向けの難易度の書籍であるにもかかわらず小学 4～6 年向けの難易度と表示されるなど、難易度の推定が不適切と評価された。

（2）実用化に向けた課題

検索表示システムで試行した「難易度検索」は、現代語で記述された小学校、中学校、高校、大学の教科書から抽出した語彙を元にして難易度を推定するものである。本実証実験では、難易度検索を試験的に導入したが、難易度を設定する書籍に明治期～昭和戦前期の古いものが多く、現代教科書の語彙に基づく難易度の推定が十分には機能しなかったと考えられる。難易度を推定する語彙のまとまりを、書籍の年代に応じて変えるなど、根本的な改善が必要である。

4. 1. 2 自然文検索の評価

（1）評価結果

自然文検索に対しては、機能を評価する意見は少なく、改善・指摘事項が大半であった。自然文検索機能に対する、ヒアリングにおける改善・指摘事項を以下に示す。

- ・「自然文検索」という機能の意味がわからない。
- ・キーワード検索と自然文検索の違いがわからない。
- ・同じ検索語をキーワード検索と自然文検索に入力したが、検索結果が異なった。
- ・自然文検索で入力できる文字数が 200 字に制限されている点が不自由である。

(2) 実用化に向けた課題

本実証実験では、校正・構造化処理を行った全文テキストデータが全体の 2%程度であり、正しく文字認識された文章を含む書籍が少ない一方で、文字を誤認識した全文テキストデータを多く含んでいた。また、本実証実験では、TF-IDF²⁵という計算方法を用いて検索結果を出力したが、検索ボックスに入力された検索語（自然文）を形態素解析した結果をすべて入力値として用いた。そのため、検索語に含まれる数字や 1~2 文字では意味をなさない単語も入力値になり、検索結果が不適切であるという評価を受けた。例えば、検索語として「～しております」という自然文を入力した場合、形態素解析の結果である「し」「て」「おり」「ます」が検索エンジンの入力値となり、これらの文字がテキストデータに多く含まれている書籍がヒットする。また、スニペットではこれらの文字がハイライトされる。

設定した検索語に対する検索結果として不適切と思われる書籍やスニペットが表示されたり、誤認識した文字を含むデータが検索結果として表示されたことが自然文検索に対する低評価につながったと考えられる。

検索結果の精度を向上させるためには、今後多くのテキストデータを蓄積していくとともに、多くの利用者の試行を受けて、検索結果がより適切なものになるよう、TF-IDF のチューニングを行うことが必要である。

また、機能の名称を、一般になじみの薄い「自然文検索」とした場合、名称から実装されている機能をイメージすることが難しくなる。入力した文章と似ている内容を探し出すことを、説明文として画面に明示するなどの工夫をすることで、利用者にとっても理解しやすく、使いやすい機能になると考えられる。

4. 1. 3 サジェスチョンの評価

(1) 評価結果

ヒアリングによると、サジェスチョンに対して以下の評価を受けた。

- ・ 検索語の候補が書籍にある語から選ばれる点が便利である

既存の検索サービスでも、同様の機能が提供されているが、図書館で提供している検索サービスで実装しているところは少ないこともあり、ある程度利便性が評価された。

サジェスチョンに対するヒアリングにおける指摘・改善事項を以下に示す。

²⁵ 文章内でキーワードがどれだけ多く使用されているのかを示す指標 TF (term frequency) と、そのキーワードがどれだけ数の文章で使用されているかを示す指標 IDF (inverse document frequency) を用いたアルゴリズム。

- ・単純な文字からのサジェストだけでなく、カタカナ、ひらがな、ローマ字の入力から漢字をサジェストできるのが望ましい。

(2) 実用化に向けた課題

検索語の候補を、書籍にある語から選択している点は、書籍の全文検索サービスとしてふさわしい機能であるといえる。一方で、指摘・改善事項として挙げられた点は、既存の検索サービスがすでに一般的に提供している機能である。サジェスチョン機能についても、基本的なレベルにとどまらず、最新の技術にキャッチアップすることが望ましい。

4. 2 検索結果一覧画面における機能の評価

評価方法として、一般利用者 11 名、有識者 8 名に出版社を加えた評価者に対して、機能や表示の利便性に関するアンケートおよびヒアリング調査を実施した。検索結果一覧画面の評価として、全文テキストデータを活用した、以下の機能や表示の利便性を評価した。

- ・ランキング
- ・スニペット
- ・連想検索

4. 2. 1 ランキングの評価

(1) 評価結果

ランキングに対しては、評価する意見は少なく、大半が改善・指摘事項であった。

検索表示システムによるランキングの結果に対しては、違和感があるという意見があった。

(2) 実用化に向けた課題

本実証実験では、全文検索エンジン Solr の機能を利用して、検索結果を検索語との関連の強さでスコアを付け、関連度順に表示した。スコアの重み付けは、目次・本文・索引のデータの特徴を考慮し、以下のように重み付けした。

○【キーワード検索の簡易検索】

書誌：2,000 倍、本文目次：1,000 倍、本文全文：1 倍

○【キーワード検索の詳細検索】

書誌のタイトル：5,000 倍、本文目次：1,000 倍、本文索引：1,000 倍、
本文全文：1 倍

ただし本実証実験では評価期間が短かったこともあり、評価者の意見を受けて重み付けを変更することはできなかったため、ランキングの結果が適切でないという評価を受けた。実運用に際しては、検索対象のデータの増加を図るとともに、より多くの利用者による試用を受け、その結果をフィードバックして各種パラメータの調整を行うことが必要である。

また、ランキングの方法を検討する上では、既存の検索サービスのランキング手法も参考にすることが望ましい。例えば、Google で採用されている PageRank は「多くの web ページからリンクされているページは重要である、重要なページからのリンクは重要である」といったアイデアである。これを書籍に応用して「多くの書籍

から引用されている書籍は重要である、重要な書籍で引用されている書籍は重要である」という考え方に沿ったランキングが考えられる。これは一例であり、このようなランキングを実現するためには、全文テキストデータに対して「引用書籍」という構造を付与する必要があるものの、適切なランキングの作成の一助になる可能性がある。

また、現在の検索結果の単位は書籍 1 冊であるが、より小さな単位、例えば章や節などを単位とすべきという意見がある。アンソロジー²⁶やオムニバス²⁷など、複数の著者による複数の作品が納められた書籍では検索結果として、書籍内の作品単位で区別する方が有効である。このような場合には、著作としての単位である「作品」と物理的な単位である「書籍」とを区別して取り扱うことが重要である。また、著作の単位に留まらず、利用者が求める知識・情報そのものに効率的にたどり着けるように、それらが記述されている章・節・パラグラフ等の任意の単位で、必要な部分を特定し、取り出せる機能が求められる。

さらに、ランキングを算出する際に、利用者のこれまでの検索履歴や閲覧履歴を用いることも、利用者の求める検索結果を提供する有効な手法と考えられる。最も単純な履歴の活用方法として、利用者の閲覧回数が高い書籍を上位にランク付けするというものがある。また、ある検索語を選んだ際の閲覧回数というように、検索語と閲覧履歴を組み合わせた情報を利用する方法もある。個人の行動履歴の活用には慎重に行わなければならないが、履歴と個人情報の結びつきを断ち、匿名化された統計情報として活用することについては、利用者の検索結果に対する満足度の向上という観点から有効であると思われる。

4. 2. 2 スニペットの評価

(1) 評価結果

評価者に対して実施したアンケートの結果を図 4-1 に示す。

²⁶ 異なる作者による詩を集めたもの。

²⁷ 短い数編の独立した作品（主に短編）または楽曲を集め、ひとつにまとめて一作品としたもの。

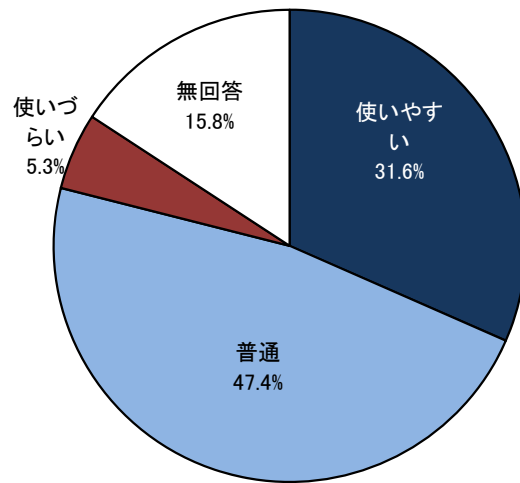


図 4-1 スニペット検索に関するアンケート結果

「適切である」「普通」と回答した評価者の合計が全体の約 80%であった。ヒアリングの回答によると、スニペットに対しては以下の評価があった。

- ・スニペット中の検索語をハイライト表示することにより、検索性の向上につながる。
- ・Google や Amazon 等の既存の検索サービスと類似した表示はなじみやすい。

スニペットに対するヒアリングにおける指摘・改善事項を以下に示す。

- ・スニペットとして表示する文章の位置が妥当でない。
- ・1つの書籍に対して複数のスニペットを表示したい。
- ・スニペットから直接本文の該当箇所にリンクしたい。

(2) 実用化に向けた課題

検索表示システムでは、表示するスニペットを選ぶ際に、検索語の出現頻度をスコア化して、高いスコアの文章を選んでいる。スコアの算出にあたっては、タイトルや目次、索引に含まれる検索語に対して、それ以外の検索語より重み付けを大きくした。そのために、スニペットとして表示される文章として、目次近辺が優先され、表示するスニペットが妥当でないという指摘を受けたものと考えられる。スニペットとして表示する文章の位置を選択する方法には改良の余地がある。

また、スニペットの候補が複数存在する場合もあることから、スニペットの複数表示も検討が必要である。ただし、複数のスニペットを表示することで、一画面で

の視認性が低下する恐れもあるため、レイアウトを工夫しなければならない。

さらに、検索結果から本文を閲覧する手段の1つとして、利便性を向上する観点から、目次から本文へのリンクと同様の技術により、スニペットから本文にリンクすることも検討することが必要である。

4. 2. 3 連想検索の評価

(1) 評価結果

ヒアリングの回答から、連想検索機能に対して以下に示す評価を受けた。

- ・表示される検索結果が適切でないと感じる。
- ・検索結果のスニペットに誤認識された文字が多数含まれているため、正しい評価ができない。
- ・興味深くはあったが、検索結果画面から積極的に用いることはないと考える。
- ・連想検索は簡単に検索ができて便利である。

本実証実験においては、検索対象として、OCR処理のみを施し校正を行っていないために、文字を誤って認識した書籍のデータを多数含んでいた。全文検索の対象とするデータを短時間で大量に用意する必要があったために、やむを得ず取った処置であったが、連想検索機能の結果として、これらの書籍のデータが表示されるケースがあり、その結果、機能自体が低評価を受けることになったと考えられる。

連想検索は、検索結果一覧画面に表示されるチェックボックスをチェックして「連想検索」ボタンを押すだけの簡便な操作であるが、利用者にとって、連想検索機能の具体的な利用イメージが浮かばないことに起因して、機能としては面白いものの使われないという指摘もあった。

(2) 実用化に向けた課題

本実証実験では、検索表示システムに蓄積された全文テキストデータのうち、校正・構造化された全文テキストデータが少なく、大半が誤認識した文字を含むテキストデータであったため、想定した結果が得られないケースがあったと考えられる。今後、多くの全文テキストデータを蓄積し、検索結果の重みづけに関するパラメータをチューニングすることで、この点が改善することが期待される。また、連想検索に関する技術が進展することで、人間が書籍の内容から連想した結果に近い検索結果が、システムの連想検索機能から得られるような改善が望まれる。

なお、本機能を利用した場合、連想検索の元となった書籍も連想検索の結果に含まれて表示される。本来であれば、連想検索の元となった書籍は表示から除外されるべきである。

連想検索と類似した機能として、既存のショッピングサイトなどにみられる、「この商品を見た人は、ほかにもこのようなものを見ています」という形の「お勧め機能」の実現を求める意見も挙げられた。図書館における利用では、「この書籍を検索・閲覧した人は、ほかにもこのような書籍を閲覧しています」という形で、利用者に対して検索の幅を広げるような機能が実現できる。この機能の実現には、検索・表示のログデータを分析・活用する必要がある。高度な検索の実現には、ログデータにも着目が必要である。

4. 3 書誌詳細表示画面の評価

評価方法として、一般利用者 11 名、有識者 8 名に出版社を加えた評価者に対して、機能や表示の利便性に関するアンケートおよびヒアリング調査を実施した。書誌詳細表示画面の評価として、全文テキストデータを活用した以下の機能や表示の利便性を評価した。

- ・目次
- ・文脈検索
- ・固有名表示
- ・タグクラウド

4. 3. 1 目次の評価

(1) 評価結果

ヒアリングの回答では、目次があることにより、書籍の情報がよりわかりやすくなるという点が評価された。

一方、目次の表示レイアウトについてはさらなる改善が求められた。具体的な指摘・改善事項を以下に示す。

- ・目次情報が複数の画面に表示されるが、どちらか一方でよい。
- ・目次がインデントされておらず読みづらい。
- ・目次情報として、当該の章・節・項の開始ページ番号や章・節・項の総ページ数などの情報も表示すべきである。

(2) 実用化に向けた課題

検索表示システムでは、書誌詳細画面と本文表示の 2 つの画面に目次情報を表示している。書誌詳細画面では本文の構造を示すために目次を表示する一方、本文表示では、本文の構造を示すとともに、本文へのリンクも実現している。このように、それぞれ異なる目的で目次を表示していることから、目次は現状のように 2 つの画面に表示することが妥当である。

一方、検索表示システムでは、テキスト化システムで生成した目次の階層に関する情報を、目次の表示には活用せずに、図 4-2 のように文字を列挙して表示した。目次の表示にも構造情報を活用して、インデント等の表示上の工夫を行い、視認性の向上を図ることが望ましい。

<div style="background-color: #cccccc; padding: 2px; margin-bottom: 5px;">目次</div> <p>緒言 第一章電化の要素 第一項電気の応用 一、発電水力と火力 二、電熱及照明 三、電熱 四、電力 五、電気鐵道 六、電気化學及冶金 七、電気通信 八、電気治療 九、其の他の應用 第二項經濟 第二章電化の實現 第一項工業の電化 第二項家庭の電化 第三項鐵道の電化 第四項農業の電化 第三章結論</p>	<p>大きさ、容量等 60p ; 23cm 責任表示 書柳榮司 [著] 責任表示 越佐教育雜誌社 編輯 JP番号 21342560 校了日 2008-01-09 最終更新年月日 2010-03-23T15:46:24Z 目録規則 NCR レコードの状態 新規 タイトル 電化問題 タイトル(読み) デンカ モンダイ NDC 543 対象利用者 一般 資料の種別 図書(日本語) 資料の種別 = 本文の言語 jpn : 日本語 本文の言語 jpn : 日本語</p>
---	---

図 4-2 検索表示システムにおける目次の表示例

現在の検索表示システムの目次には、ページが表示されていない。これは、テキスト化システムの出力フォーマットが、ページ番号に関する情報を持たなかったことも影響している。したがって、ページ範囲を把握するためには、元画像を表示させるしかない。一般に、目次に表示されるページ数が章や節の分量の目安となる場合もあることから、ページ数の表示についても検討が必要である。なお、電子書籍の一形式であるリフロー型の書籍データは、Web ページと同様に、ページサイズが可変である。このような書籍に対してどのように「ページ」という単位を与えるかなど、様々なケースを想定することが必要である。

4. 3. 2 文脈検索の評価

(1) 評価結果

ヒアリングの回答によると、短時間で知りたい箇所を閲覧できる点を評価する意見があった。

文脈検索に対する、ヒアリングにおける課題・改善事項を以下に示す。

- ・文脈の検索結果の文字列から、本文の該当箇所にリンクできるとよい。
- ・「書籍内の固有名」や「タグクラウド」をクリックして、文脈検索の検索語を

- 入力できるようだがわかりづらい。
- ・新たな検索語を入力する方法が不便である。

(2) 実用化に向けた課題

検索表示システムでは、検索結果と本文とのリンクは、後述する目次と本文のリンクのみである。利用者にとっては、本文に直接アクセスできる手段が多く用意されているほど利便性が向上すると考えられる。本文へのアクセス手段を多く用意するという観点から、文脈検索の結果からも、本文にリンクすることが望まれる。

検索表示システムは、基本的にはシンプルな画面構成であり、詳細な説明がなくても大抵の機能はすぐに使うことができる。ただし、書籍内の固有名やタグクラウドとして表示した文字をクリックすると、文脈検索の検索語として入力されることは、気づきづらい。このような機能については、画面に説明を記述するなどして、利用者にわかりやすくすることが必要である。

なお、本機能の名称である「文脈」という言葉の意味が難しく、「文脈検索」という言葉自体になじみがないために、機能をイメージすることが難しいという意見がいくつか挙がっていた。機能の名称についても、利用者の視点から、よりわかりやすく、機能をイメージしやすいものであることが望ましい。本機能についても、「検索語周辺文表示」など、実態に合った名称とすべきと考える。

4. 3. 3 固有名表示の評価

(1) 評価結果

全文テキストデータから固有名を抽出するというアイデアは、従来の検索サービスには見られなかった取り組みとしては評価できる。しかしながら、固有名が適切でないために評価できないという意見がアンケートの自由回答で挙げられた。

(2) 実用化に向けた課題

テキストデータの中から固有名を抽出するためには、文章の意味や構造を解析する必要がある。しかし、検索表示システムでは、意味や構造の解析を行っておらず、「月日」の固有名として「28月」と表示されるケースがあった。また、「東京都立図書館」という固有名詞の文字列が、「京都」という固有名詞を含むと判断する場合もあり、本機能は実用的なレベルには達していない。文章の意味による解析や構文の構造による分析など、自然言語処理における研究成果を活用して、表示される固有名から本文の内容が類推できるレベルになることが期待される。なお、この機能に関しては、固有名で文章を分割することの意味を十分に検討した上で実装の可否を考慮する必要がある。

4. 3. 4 タグクラウドの評価

(1) 評価結果

タグクラウド表示に関しても、本文内の頻出単語を出現頻度に合わせて文字サイズと文字色を変更して表示するというアイデアに対しては、ある程度の評価が得られた。しかしながら、「この」「その」というような指示語が抽出されるなど不要な語が多いという指摘があった。

また、「タグクラウド」という言葉も一般にはまだなじみがなく、名前から機能をイメージしづらいという意見もあった。

(2) 実用化に向けた課題

本機能の狙いは、表示された語句から本文の概要を理解することであるが、現状ではそれが実現できていない。頻出単語のスコアを算出して表示するだけでなく、「する」「ある」「いる」などの重要な意味を持たない動詞は除外する、「この」「その」などの指示語は除外するなど、表示前に頻出単語の選別を行うことで、効果が改善される可能性がある。

なお、タグクラウドとして表示された語句は、クリックすると文脈検索の検索語として入力される。頻出単語という、本文と密接に関連する単語であるので、本文とリンクする、本文中の語句を検索してハイライト表示するなど、タグクラウドを活用した機能拡張も検討に値する。

4. 4 本文表示画面の評価

評価方法として、一般利用者 11 名、有識者 8 名に出版社を加えた評価者に対して、機能や表示の利便性に関するアンケートおよびヒアリング調査を実施した。本文表示画面の評価として、全文テキストデータを活用した、以下の機能や表示の利便性を評価した。

- ・書籍本文のテキスト表示
- ・目次・本文リンク
- ・検索語出現数表示
- ・検索語ハイライト

4. 4. 1 書籍本文のテキスト表示の評価

(1) 評価結果

書籍本文のテキスト表示に対するアンケート結果を図 4-3 に示す。

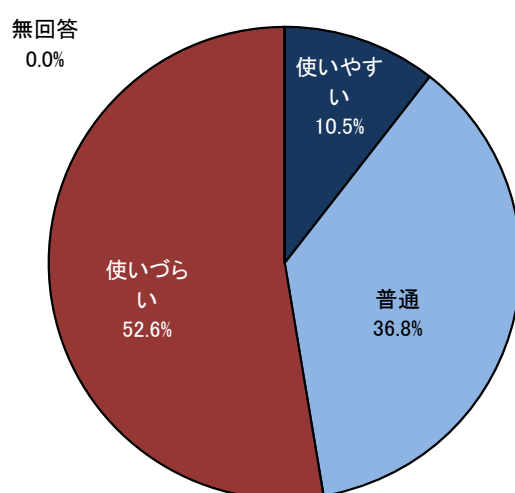


図 4-3 書籍本文のテキスト表示に関するアンケート結果

「使いづらい」とした評価者が 50%以上であった。特に、リンク先として表示する書籍本文に対して、現状の実現方法では問題があるとされた。

書籍本文のテキスト表示に対する、ヒアリングにおける指摘・改善事項を以下に示す。

- ・本文表示はプレーンテキストでなく、元画像のデータや元画像のレイアウト情報を残すような形で表示すべきである。

(2) 実用化に向けた課題

検索表示システムの本文表示では、プレーンテキストを表示した。その際、本文内の見出し項目はフォントサイズをほかの文字より大きくするなどの処理はしていたものの、行間が詰まった形で表示され、章・節・項などの文章の区切りがわかりづらくなった。検索のための基本的な情報としてプレーンテキストの情報は必要であるが、本文を表示する場合には、できる限り元の書籍に近い形が望ましい。しかしながら、本文表示には元画像を用いるのか、全文テキストデータに対してレイアウト情報を付与して表示させるのかについては、便益を受ける利用者の観点だけでなく、便益を提供する、書籍データの提供者としての観点からも十分な検討が必要である。

4. 4. 2 目次・本文リンクの評価

(1) 評価結果

目次・本文リンクに対するアンケート結果を図 4-4 に示す。

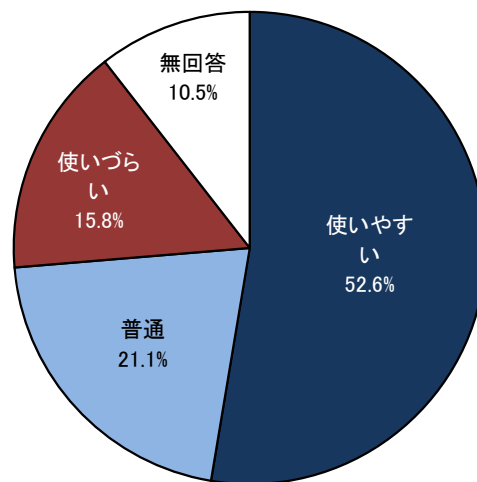


図 4-4 目次・本文リンクに関するアンケート結果

「適切である」「普通」とした評価者の合計が 70%以上であった。本文表示の目次から、本文中の該当箇所にリンクされており、目次をクリックすることで本文の該当箇所を表示できることの利便性が評価された。

目次・本文リンクに関する、ヒアリングにおける指摘・改善事項を以下に示す。

- ・目次と本文のリンクは、本文表示部分をスクロールして本文中の位置を表示す

るのではなく、目次とリンクした章・節・項の部分だけを表示すべきである。

(2) 実用化に向けた課題

検索表示システムの本文表示は、本文の全データを表示する形とし、リンクもスクロールによる移動で実現した。テキストデータが構造化されていることから、この構造情報を活用して、指定した章・節・項だけを抜き出して表示することも考慮すべきである。

本文へのリンク方法については、検索結果一覧画面に表示されたスニペットをクリックしたり、書誌詳細表示画面の文脈検索の文字列をクリックすることで、本文の該当箇所が表示されるなど、リンクの実現方法については今回実装した以外の方法の実現を求める意見が挙げられた。検索表示システムが実用化される際には、これらの機能についても実装を検討することが望まれる。

4. 4. 3 検索語出現数表示の評価

(1) 評価結果

ヒアリングによると、検索語出現数表示に対して以下の評価を受けた。

- ・検索語が多く出現する章や節がわかるので、興味がある箇所、関係の強い箇所に早くたどりつける。

検索語出現数表示に対するヒアリングにおける指摘・改善事項を以下に示す。

- ・アイデアとしては面白いものの、この表示によりかえって目次そのものの視認性が低下しては本末転倒である。

(2) 実用化に向けた課題

本実証実験で使用した全文テキストデータは、単純に OCR 処理を施しただけのもの、OCR 処理後に文字の校正を行わずに構造化作業だけを行ったもの、文字校正から構造化までのすべての作業を行ったものなど、構造化の品質の観点において様々なデータが混在していた。そのため、検索結果として表示された書籍の中には、文字の校正が不完全であったり、構造化されていないものがあった。構造化されていない書籍では、目次が表示されず、検索語出現数を表示する場所がないために検索語出現数が表示されなかった。これは検索表示システムの問題ではないものの、機能の評価には大きな影響を及ぼす。検索表示システムの機能確認においても、可能な限り多くの校正・構造化された全文テキストデータを対象とし、データの品質が

機能の評価に影響しないよう、留意する必要がある。

なお、「4. 6. 1 全文テキストデータの読上げの評価」で実施した読上げサービス評価において、読上げソフトウェアがブラウザに表示された内容を読上げる際に、目次の項目の後に本機能による検索語の出現数が読上げられるため、出現数をページ数と誤解される例があった。晴眼者にとっては一目でわかる機能であっても、視覚障がい者にとっては誤解を与える情報となりかねないため、視覚障がい者の利用も考慮して表示位置や表示方法等を検討する必要がある。

4. 4. 4 検索語ハイライトの評価

(1) 評価結果

検索語ハイライトに対するアンケート結果を図 4-5 に示す。

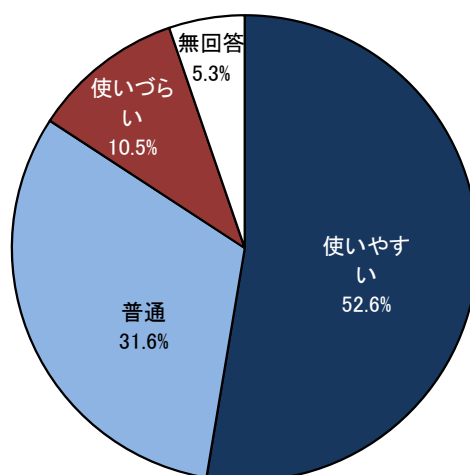


図 4-5 検索語ハイライト表示に関するアンケート結果

「適切である」「普通」とした評価者の合計が 80%以上であった。全文テキストデータに含まれる検索語をハイライト表示することで、検索語と合致する本文中の位置を一目で把握でき、検索語の視認性が向上した。

検索語ハイライトに対する、ヒアリングにおける指摘・改善事項を以下に示す。

- ・複数の検索語を設定した際に、それぞれの検索語で色を変えてハイライトすべきである。

(2) 実用化に向けた課題

検索語ハイライト表示は、検索結果の視認性を向上させる上では欠かすことので

きない機能であり、検索表示システムにおいても、本文表示だけでなく、検索結果のスニペット表示でも実現している機能である。ただし、複数の検索語を入力した場合でも、同じ背景色でハイライト表示しているため、今後はこれを区別してハイライト表示するなど、さらに視認性を向上することが望まれる。

4. 5 ページ構成の評価

評価方法として、一般利用者 11 名、有識者 8 名に出版社を加えた評価者に対して、検索表示システムの各画面における、検索語入力エリアや検索ボタン、文脈検索などの検索機能の配置、および検索結果一覧画面や本文閲覧画面における目次や本文などの表示に関するページ構成に対して、アンケートおよびヒアリング調査を実施した。

(1) 評価結果

ページ構成に対するアンケート結果を図 4-6、図 4-7 に示す。

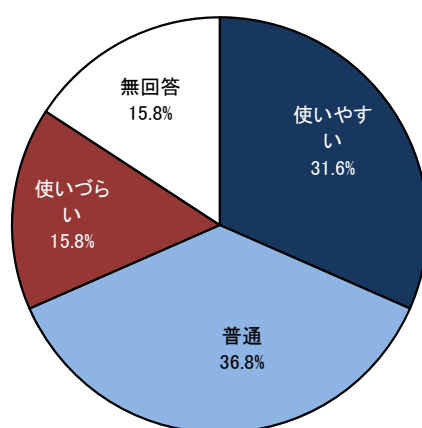


図 4-6 各機能の配置に関するアンケート結果

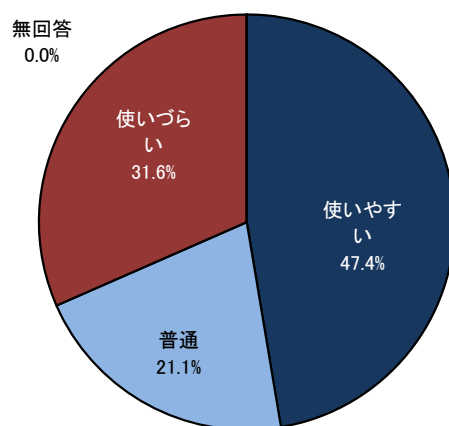


図 4-7 ページ全体の構成に関するアンケート結果

各検索機能の配置、ページ全体の構成とも「使いやすい」「普通」とした評価者の合計が約 70%であった。ただし、ページ全体の構成については、「使いづらい」とした評価者も約 30%を占めており、インタフェースの改善を求める指摘が多く見

受けられた。

ページ構成に対する、ヒアリングにおける指摘・改善事項を以下に示す。

- ・検索結果一覧表示画面の表示件数を、利用者が変更できるようにすべきである。
- ・画面の文字の大きさを変更できるようにすべきである。
- ・すべての機能が常に表示されていてわかりにくい。
- ・スクロールが必要な部分が多くて見づらい。

(2) 実用化に向けた課題

上記のいずれの指摘事項についても、ページ構成を検索系の Web デザインの基本に沿って作成することで解決できるものであるため、専門家を起用するなどにより、ページ構成の改善を図ることが望ましい。

4. 6 視覚障がい者等向けの読上げサービス等の評価

評価方法として、視覚障がい者等 18 名からなる評価者に対して、機能や読上げサービス等の利便性に関するアンケートおよびヒアリング調査を実施した。視覚障がい者等向けの読上げサービス等の評価として、全文テキストデータを活用した、以下の機能や表示の利便性を評価した。

4. 6. 1 全文テキストデータの読上げの評価

(1) 評価結果

目次で指定した位置から読上げが開始できる点について、利便性が高いと認められた。

ただし、目次表示でページ数の情報があるとよい、また、現在読上げられているページ番号がわかるとよい、という意見が得られた。

また、トップ画面から、検索、書誌詳細表示、本文表示にいたる各画面の構成については、1 ページあたりの情報量を減らし、情報量が多い時は別ページとする、重要な情報をページの先頭よりに位置するなど、シンプルで分かりやすい構造とすべきという意見も多数あった。

(2) 実用化に向けた課題

検索表示システムに対しては、通常システムと視覚障がい者用のシステムに 2 つに分けてはどうかという意見もあったが、その 2 つを分けずに全体の構造をわかりやすくすべきとする意見が大半を占めた。2 つのシステムに分けることにより、それぞれの利用者に向けた適切な対応が取れる一方、メンテナンス作業が 2 倍になり、一方を改修した際に、もう一方の改修が滞る恐れがあるということから、1 つのシステムとすべきという意見があった。同じシステムを利用する場合でも、視覚障がい者向けの情報を、画面には表示されない形で埋め込むことが可能である。このような対応を様々な画面で採用することで、一般の利用者にも、視覚障がい者にも利用しやすいシステムを構築することが必要である。

4. 6. 2 全文テキストデータの品質の評価

(1) 評価結果

検索表示システムを用いた、視覚障がい者向けの検索・読上げサービスでは、テキスト化システムで校正・構造化作業を行った全文テキストデータを利用した。市販の読上げソフトウェアを用いた画面表示の読上げに対する聞き取り結果はおおむね実用的であるという評価が得られたが、正しく読み上げられないケースが少なからず発生し、課題が明らかとなった。

(2) 実用化に向けた課題

校正・構造化作業時の文字修正のミス、読上げソフトウェア自体の語句の認識ミスなど正しく読み上げられない要因として様々な事項が想定されるが、大きな問題の1つは、読上げソフトウェアが旧字体の漢字に対応していないことである。特に本実証実験では、明治期～昭和戦前期の比較的古い書籍も検索対象であり、その中に旧字体の漢字を含んでいた。

旧字体の漢字の読上げに対しては、以下の対応策が考えられる。

- ・読上げソフトウェア側で旧字体の読上げに対応する。
- ・全文テキスト化する際に旧字体をすべて新字体にする。
- ・テキスト化システムで作成されたデータを検索表示システムに蓄積する際に、読上げソフトウェア用に旧字体を新字体に変換したテキストを作成する。

本質的な対応策は、一番目の読上げソフトウェア側での対応である。しかし、現状では旧字体を読み上げるニーズも少なく、実現が困難であると考えられる。また、国立国会図書館の提供するサービスとして考えると、できるだけ多くの読上げソフトウェアに対応することが望ましく、読上げソフトウェアに依存して読上げ結果が変わるべきではない。

二番目の全文テキスト化作業での対応についても、全文テキスト化作業において、どこまで元の書籍からの変更を許すのかという点が重要になる。今回の全文テキスト化作業においても、OCRの処理がJIS第2水準までしか対応していないことに対して疑問を呈する意見やJIS第3、第4水準までは少なくとも対応すべきという意見が出された。

全文テキスト化作業において、JIS第3、第4水準の文字まで対応した上で、読上げのためにJIS第2水準までの文字との対応関係も持たせるためには、全文テキスト化データにおいて、1つの文字に対して新字体と旧字体の複数の文字情報を持たせるような改修が必要になる。また、検索表示システムでも、改修した全文テキスト化データを読み込んで、新字体と旧字体を区別して、表示は新字体、読上げは旧字体で行うというような処理が必要になる。

検索表示システムに全文テキストデータを蓄積する際に、変換処理を行う方法については、旧字体読上げ対応の処理として独立して行うことが可能である。後段の「4. 8 文字コード対応の評価」でも述べるが、新字体／旧字体については検索の際にも問題となる。検索の場合は、検索表示システムの持つ文字の統制機能を活用することで、旧字体を新字体に読み替えることが可能である。しかし読上げソフ

トウェアの場合には、表示された文字列をそのまま読み上げるため、あらかじめ旧字体を新字体に置き換えたテキストデータが必要となる。

このためには、旧字体の漢字を新字体の漢字に一括して置換するソフトウェアを開発することで、容易に実現できる。ただし、検索表示システムが元の全文テキストデータと、新字体に置き換えた読上げ用の全文テキストデータの2種類を持ち、表示用と読上げ用に使い分けをしなければならないので、体系的な検討が必要となる。

4. 6. 3 OCR 認識率が読上げサービスに与える影響の評価

(1) 評価結果

評価では、参加した視覚障がい者等全 18 名のうち、14 名から回答を得た。

本評価に使用した書籍と OCR の認識率などの特徴を表 4-1 に示す。

表 4-1 対象とした書籍

	特徴	
書籍 1	OCR 認識率 98%、新仮名遣い(発刊年度も書籍 2 より新しいもの)	
書籍 2	OCR 認識率 98%	
	書籍 2-1	旧仮名遣い
	書籍 2-2	新仮名遣い(「旧仮名遣い→新仮名遣い」変換を行うソフトウェアを通したもの)
書籍 3	OCR 認識率 93%	
	書籍 3-1	旧仮名遣い
	書籍 3-2	新仮名遣い(「旧仮名遣い→新仮名遣い」変換を行うソフトウェアを通したもの)
書籍 4	OCR 認識率 90%、新仮名遣い	
書籍 5	OCR 認識率 70%、新仮名遣い	

また、書籍の内容の理解度は、以下の選択肢から選択することとした。

- 1 読み間違いにより理解できない箇所が多くあった。
- 2 正しい読みが想像できない読み間違いが多く、理解できない箇所がいくつかあった。
- 3 読み間違いと思われる箇所も正しい読みが想像できたので、正確に理解できた。
- 4 ほとんど正確に読めていた。

各書籍についての評価参加者の理解度を表 4-2 にまとめる。

表 4-2 OCR 認識率の異なる書籍に対する理解度

書籍 (文字認識率、 仮名遣い)	理解度の分布					理解度の平均値
	1 読み間違えにより理解できない箇所が多かった。	2 正しい読みが想像できない読み間違えが多く、理解できない箇所がいくつかあった。	3 読み間違えと思われる箇所も正しい読みが想像できたので、正確に理解できた。	4 ほとんど正確に読めていた。	0 未実施	
書籍 1 (98%、 新仮名)	1 人	3 人	6 人	1 人	3 人	2.6
書籍 2-1 (98%、 旧仮名)	0 人	2 人	4 人	1 人	7 人	2.9
書籍 2-2 (98%、 新仮名)	0 人	2 人	4 人	1 人	7 人	2.9
書籍 3-1 (93%、 旧仮名)	2 人	3 人	3 人	0 人	6 人	2.1
書籍 3-2 (93%、 新仮名)	0 人	2 人	2 人	0 人	10 人	2.5
書籍 4 (90%、 新仮名)	7 人	2 人	1 人	0 人	4 人	1.4
書籍 5 (70%、 新仮名)	3 人	3 人	0 人	0 人	8 人	1.5

この結果から、OCR の認識率の違いにより、読上げサービスに影響があることが明らかとなった。また、認識率が低くなるにつれ、理解度が下がることがデータからもほぼ明らかとなった。以下、書籍ごとの評価結果から導出される結論を述べる。

OCR 認識率が 98%である書籍 1、2については、いずれも 3 程度の評価であり、相当程度、正確に理解できたことがわかる。また、同じ書籍で旧仮名遣いと新仮名遣いで異なる書籍 2-1 と書籍 2-2 については、明らかな差異が見受けられなかった。今回の検証では、OCR 認識率が高いテキストデータの場合には、旧仮名遣いで読めない部分があったとしても、文章の内容を理解する上では大きな影響とはならないことを示す。

OCR 認識率が 93%である書籍 3については、およそ 2 程度の評価であり、書籍 1、2 に比べて理解が難しいことがわかる。なお、新仮名遣いで記載された書籍 3-2 に

比べて、旧仮名遣いで記載された書籍 3-1 は理解度が低い。これは、新仮名遣いでも理解が難しいテキストデータに対して、旧仮名遣いであることがさらに理解を難しくしていると考えられる。OCR 認識率の低下と仮名遣いの違いの 2 つの要因により、同じ OCR 認識率であっても、新旧仮名遣いの違いが書籍 2 よりも大きくなったと言える。

OCR 認識率が 90% の書籍 4 および 70% の書籍 5 に至っては、評価がおおよそ 1.5 程度であり、「1. 読み間違いにより理解できない箇所が多くあった。」という評価が増え、理解が困難になることがわかる。

(2) 実用化に向けた課題

以上の結果から、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言える。ただし、OCR 認識率が低い書籍についても、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挙げられた。ただしその場合には、OCR で読み取ったままのもので、誤認識部分がある旨を明示する必要があるとのことであった²⁸。

4. 6. 4 DAISY ファイルの読上げの評価

(1) 評価結果

視覚障がい者等向けのデジタル録音図書の国際標準規格である DAISY フォーマットで出力された全文テキスト化ファイルを対象に、読上げの評価を行った。DAISY ファイルの構造情報を利用して、章・節単位の読上げ位置の移動が可能な点などの利便性が認められるとともに、DAISY ファイルが正常に読上げられていることが確認できた。

(2) 実用化に向けた課題

DAISY フォーマットは、視覚障がい者向けの読上げフォーマットとして確立された技術であり、既に実用化されていることから、現状の DAISY フォーマットについては大きな課題は存在しないといえる。ただし、現在の欧米での電子書籍フォーマットの主流である EPUB 形式の次期規格が、DAISY 形式の次期規格と統合されるという動きもあることから、関連する規格の動向も注視し、主流となる規格に対応すべきである。

²⁸ 米国の Bookshare (<http://www.bookshare.org/>) では、提供しているテキストについて、OCR で読み取っただけのものなのか、手を加えてあるのか、といったランク付けがされている。

4. 7 全文テキストデータのインデキシング²⁹処理時間の評価

評価方法として、対象全文テキストのインデキシングに要する投入ページ数、所要時間の記録を収集し、累積投入ページ数の影響などを分析し、実用化に向けた問題がないかなど技術課題を抽出した。

(1) 評価結果

全文テキストの蓄積書籍が多くなっていったときの新規のインデキシング処理時間を計測・分析した。データ蓄積を4回行った際のインデックスサイズ、インデキシング処理時間などの計測結果を表4-3に示す。

表 4-3 インデキシング処理時間

				蓄積1回目	蓄積2回目	蓄積3回目	蓄積4回目
入力条件	OCR・出版社データ	書誌	作成	49件	18件	29件	337件
		本件数	追加	8件	7件	14件	322件
			更新	1件	6件	15件	15件
		インデックスサイズ	合計	1,748KB	1,836KB	5,371KB	258,657KB
平均	36KB		102KB	185KB	768KB		
計測結果	処理性能 (合計)		インデキシング処理時間	3.0s	3.0s	9.0s	424.0s

蓄積の各回において、「インデックスサイズ (合計) / インデキシング処理時間」を求めると各蓄積回とも、約 600(KB/s)となり安定している。

(2) 実用化に向けた課題

前述の結果から、蓄積書籍が増えてくる場合でも処理時間的には問題なく書籍の蓄積が行えることが確かめられた。ただし、本実証実験においても、書籍の蓄積後に校正漏れの誤りが見つかることは少なからず発生した。書籍の蓄積後に修正が入ることは、将来的にも少なくないと考えられる。このような場合に、インデックスの削除、追加、修正を高速に行えることが重要である。

本実証実験では、全文テキスト化データとして校正・構造化されていないものも含めて約 20,000 タイトル分を蓄積し、インデキシング処理を行った。これは、国立国会図書館が所蔵する書籍と比較した場合、ごくわずかな分量でしかない。国立国会図書館が所蔵する膨大な書籍を全文テキスト化データとして蓄積し、検索表示する場合には、インデキシング処理にどれだけ要するのか、現実的な処理となるのかについても、シミュレーションを行うなどして明らかにする必要がある。

また、本システムが実用化されて、実証実験の対象冊数を大きく上回る書籍が既に蓄積された状態に置いては、インデキシング処理時間が実験結果よりも長くなる可能性もあることに十分留意する必要がある。

²⁹ 高速な検索を可能とするために事前に索引ファイルを作成すること。

4. 8 文字コード対応の評価

全文テキストデータ、および検索表示における文字コード対応の評価結果、および実用化に向けた課題を以下に示す。

(1) 評価結果

評価の結果、検索表示における文字コードの対応が十分ではないことが明らかとなった。具体的には、旧字体と新字体とを区別して検索表示する文字と、区別せずに検索表示する文字とが混在していることが挙げられる。これは、検索表示システムにおける新字体と旧字体の統制機能が十分でないことだけでなく、テキスト化システムにおける全文テキスト化作業において、旧字体と新字体の扱いを明確に規定しなかったことにも起因する。

(2) 実用化に向けた課題

現在一般的に利用されている OCR は JIS 第 2 水準の漢字までしか対応しておらず、本実証実験でも OCR は JIS 第 2 水準までの対応とした。その一方で、テキスト化の校正作業では明確なルールを定めなかったため、旧字体と新字体、JIS 第 1、第 2 水準の文字とそれ以外の文字とが混在したデータが作成された。さらに、検索表示システムでも、旧字体と新字体との統制を行っていなかったために、新字体で検索しても旧字体でテキスト化された文字がヒットしないケースや、その逆のケースが生じる結果となった。

この課題は以下のように 3 つの観点で整理することができる。

- ・ OCR で処理対象とする文字の範囲をどこまでにするか。
- ・ 文字校正作業における文字修正のルールをどうするか。
- ・ 校正後のテキストデータに対して、どのようなルールにしたがって検索を行うか。

第一の観点については、全文テキスト化システムにおける OCR が処理対象とする文字コードの範囲に依存する部分である。現在の実用化レベルが JIS 第 2 水準であるものの、本来の書籍の文字を忠実に再現することを重視した場合には、JIS 第 3、第 4 水準などにも対応することが望まれる。また、処理対象文字が広がり、認識精度も向上すれば、全文テキスト化における校正作業の効率化も期待できる。

第二の観点については、全文テキスト化システムにおける校正作業時のルールに依存する部分である。今回の文字校正作業では、作業者が修正できる範囲で対応するということとしたため、同じ言葉であっても旧字で記載されたものと新字に置き換えて記載されたものが混在した状態となった。このような混在が生じないように、

JIS 第4水準までの漢字については、校正作業で入力する、対象範囲外の文字については「■」で置き換えるというような校正作業のルールを設定したり、ルール外の入力を防止するツールを導入するなどの対応が必要である。

最後の観点は、全文検索・表示システムの機能に依存する部分である。検索語のテキスト中の表現が旧字体であれ新字体であれ、同じようにヒットすることが望まれる。一方で、文字の字体に着目して、検索語と正確に一致する情報を求めるニーズもある。そのため、検索の際に旧字体と新字体を「区別する」のか「区別しない」のかを利用者が設定できるような機能が必要であると考えられる。

5 実証実験の成果と課題

本実証実験では、全文テキストデータの整備および当該データを活用した全文検索・表示サービスの提供のために解決すべき技術的課題について、現時点での到達点を確認するとともに、今後必要とされる基盤技術の確認および一層の取組みが求められる課題の整理ができた。以下では、本実証実験から得られた成果および課題をまとめる。

5. 1 テキストデータ作成に関する実証実験の成果と課題

(1) テキスト化システムのフォーマットに関する成果と課題

本実証実験では、OCR 出力フォーマットに採用した ALTO について、日本語表示固有の表現（縦書き、右横書き、和欧文混植等）に対応するため、ALTO の拡張を行った。この拡張により、後続の校正・構造化作業に必要となる日本語特有の情報を記述することができた。

OCR 出力フォーマットは、基本的には OCR ベンダが独自に策定しており、仕様が公開されていない。また、本実証実験において日本語対応のために ALTO の拡張を行ったが、ルビは ALTO に記述するのではなく構造化メタデータとして保持する対応としたなど、日本語表示固有の表現に十分な対応ができたわけではない。

本実証実験では、日本語表示固有の表現に対応した OCR 出力フォーマットの必要性を認識するとともに、OCR データの相互運用性確保のためにも OCR 出力フォーマットの標準化が必要であると認識した。

次にテキスト化システムの出力フォーマットについてであるが、本実証実験では代表的な書籍の電子フォーマットの中から、用途ごとに適切なフォーマットを利用することにした。具体的には、表示用のフォーマットとして、原本の再現性の観点から透明テキスト付 PDF を、検索・読上げ用のフォーマットとして、サーチャビリティ、アクセシビリティの観点から DAISY フォーマットを採用した。

DAISY フォーマットは、資料の要素分解（標題紙、目次、本文、章、節、索引、参考文献、引用文献等）が可能であり、本実証実験では、①縦書き、②右横書き、③和欧文混植、④ルビ、⑤段組みや多段構成、⑥目次から見出しへのリンク、⑦本文中見出し、⑧コンテンツ区切り、⑨柱、⑩ページ番号、⑪圈点や傍線等の強調記号、⑫図、⑬表、⑭参照文献、⑮文字位置情報の計 15 項目を構造化の対象とした。これらの項目を用いて資料を構造化することにより、検索の際にキーワードを重み付けしてスコアを算出したり、読上げの際に読上げをスキップしたりすることが可能となった。

ただし、書籍の電子フォーマットには、透明テキスト付 PDF や DAISY フォーマットのほかにも様々な形式が存在する。今後は、それぞれのフォーマットの動向に

留意して、全文テキスト化された書籍の出力フォーマットとして適切なものを比較検討していくことが必要である。

また、本実証実験では、国立国会図書館所蔵書籍のほかに、出版社が提供する書籍の電子データも全文検索・表示の対象とした。出版社提供データは、校正の必要はないが、検索や読上げのためにテキスト化システムで構造化処理が行われなければならない。また、表示に用いるために PDF フォーマットも必要であるため、本実証実験では、出版社から提供された電子データを PDF に変換した後、OCR 出力フォーマットに変換し、テキスト化システムに投入することにした。

出版社からは様々なフォーマットで書籍の電子データが提供されたが、本実証実験では、そのうち PDF、TEXT (フラットテキスト)、XMDF (TTX 有)、.book (TTX 有) の 4 フォーマットを対象に上記の処理を実施した。

しかし、これらの電子データの中には、複数の電子ファイルにより 1 冊の書籍データが構成されているものが存在し、手作業で 1 つのファイルにまとめるなどの対応が必要となった。今後は、このような作業を自動化する変換ツールの開発が課題となる。

(2) 共同作業支援システムによる効率化等の効果検証の成果と課題

本実証実験で実装した共同校正・共同構造化機能のインターフェースは、比較的単純だったため、短時間のガイドさえ受ければ、機能に関する予備知識がなくても作業が可能だった。また、文字の校正を単純作業に分解し複数人に分配する、クラウドソーシングの技術が活用できる可能性が高いことも確認できた。クラウドソーシングにより多量の人員を導入してテキスト化作業を行う場合には、機能の予備知識や特別なスキルが必要ない点は大きなメリットといえる。

また、共同校正・共同構造化にかかる作業時間を、書籍別、書籍の時代別、構造化項目別に計測し、テキスト化作業全体の所要時間を算出することができた。作業時間の短縮という課題はあるものの、共同作業支援システムを用いて複数人でテキスト化を行った場合の作業時間が計測できたことは意義が大きい。

しかし、より効率的にテキスト化作業を行うためにも、インターフェースには更なる改善が必要である。本実証実験で実装した校正のインターフェースは、個々の文字に対して修正を行うものだったため、文章全体が把握しづらくなり、作業効率が低下するケースが発生した。また、構造化作業については、「ノート PC で作業するには文字が小さくて読みづらい」、「画面が小さいため文章全体が把握しづらい」という指摘もあった。このような課題を解決するためには、例えば、行単位などある程度意味を持った文字の集合に対して校正をしたり、小さな画面でも作業を行いやすいように画面レイアウトを変更するなどの改善が必要である。

また、テキスト化作業時のルールの設定と遵守も大きな課題である。入力する文字の新字・旧字対応や、目次の階層化をどこまで行うのか、柱の概念の定義、ページ種別の判断基準の設定など、作業上の取決めを明確にした上で作業を進めないと、作業者の負荷が上がったり、作業者によって入力ルールが異なるためにできあがった全文テキストデータの品質にばらつきが生じる恐れがある。

本実証実験では、システム全体で取り扱う文字のルールなどを明確に規定せずに作業を進めることになった。旧字体の文字があった時に、旧字体のままテキスト化するのか、新字体に置き換えてテキスト化するのかというように、文字コードの取り扱いに関するルールがなければ、作業者によってテキスト化の結果が異なることになる。また、利用者にとっても、同じ文字でありながら検索結果に含まれていなかったり、旧字体の文字を使った語句だけを検索したいのに、新字体の語句も検索されてしまうために区別ができないなどの問題が生じる可能性がある。

また、ルールを徹底するためには、人手による対応だけでなく、ツールによる支援も不可欠である。例えば、旧字体を含む対応文字だけの文字変換を行う機能や、使用が許されない文字を入力した場合に警告する機能などが必要となると考えられる。

さらに、文字コードの範囲内で表現できないものをどのように取り扱うかもルール化が求められる。こうした表現の例としては、強調記号や漢文の読み下し記号が挙げられる。青空文庫では、これら例外事項を「注記」として記録する方式をとっている³⁰。このような事例を参考として、国立国会図書館としての対応ルールを規定することが望ましい。

元の書籍に、誤植や誤記と考えられる表現が出てきた場合にはどうするのか、ルビや縦中横、傍点などの組版に対してどこまでの情報を構造化するのか、新しい組版が出てきたときにどこまで対応するのかなど、全文テキスト化作業を進める上では様々なルールを規定し、それに沿って作業を進めることが重要である。本実証実験でも、検討課題はいくつも挙げられているが、ルール化すべき事項をすべて洗い出すには至っておらず、今後、時間をかけて十分に議論することが必要である。

以上で述べたように、本実証実験では作業ルールが明確化されていなかったことから、データの品質を保つための品質・作業管理が発生した。具体的には、管理台帳によるデータの品質管理、作業者へのガイド、目視でのデータの最終チェック等である。このような品質チェックを作業者の注意力だけで解決することは現実的ではなく、作業の品質をチェックするための体系的な機能が必要不可欠であるといえる。また、クラウドソーシングによる共同作業を実現するには、作業者からの

³⁰ <http://www.aozora.gr.jp/annotation/>

質問に対応するため、作業の途中で同じ画面から他の作業や管理者に問合せができる機能や、判断に迷った部位をすぐに他の作業や管理者に報告できるようにする仕組みを実装するなどの対応が考えられる。効率的にテキスト化作業ができるよう、共同作業支援の仕組みを充実させていかなければならない。

最後に、最適なアーキテクチャの構築も課題である。本実証実験の共同校正・共同構造化機能は、1つのシステムに対して複数の作業者が並行して作業することを想定したものである。このような集中作業方式をとるためには、作業要員と対象書籍のデータ量等を明確にし、適切なアーキテクチャを構築することが必要である。

(3) 推論機能による効率化等の効果検証の成果と課題

本実証実験では、推論機能による効率化等の検証のため、1つの書籍を前半と後半に分け、前半の校正結果をOCRに再学習させるOCR再学習機能、構造情報を自動的に推論する構造情報推論機能、読上げ順序を自動的に推論する読上げ順序推論機能の3つの機能を実装した。OCRの文字認識、構造情報の設定、読上げ順序の編集の各作業について、システムによる再学習や推論を用いた場合・用いない場合の作業時間を比較したところ、再学習や推論を用いた場合のほうが作業時間が短くなり、効果が実証できた。

今後は、これらの推論機能の精度向上が課題となる。例えば、OCRの再学習については、学習効果のある文字だけを登録するなどの工夫が必要である。また、構造推論機能については、ページ番号であればページ余白部分に数字が配置される、柱であればページ余白の左右交互に縦書きでタイトルや章・節の見出しが配置されるというように、構造項目ごとに特徴があるので、それらの特徴をふまえて推論を行うことで精度の向上が期待できる。推論機能を効果的に利用し、作業負荷を軽減していくことが必要である。また、本実証実験で対象とした形式やレイアウトの構造化に加えて、意味や内容面での構造化を行うためには、推論機能のさらなる高度化を始めとして意味解析に基づく構造化手法を確立する必要がある、これらに必要となる自然言語処理の研究が重要になる。

(4) その他

OCRの精度を高めることは、後続の校正作業の時間を短縮することにつながる。有識者による評価においても、文字や単語のつながり方の規則や辞書を利用した処理を行う、簡単な形態素解析を行うことにより、OCR精度が向上する可能性があるという意見があった。例えば、OCRで読み取る前に、ひらがな、カタカナ、漢字、記号などの文字種類のつながり具合を形態素解析によりチェックするという方法も考えられる。これにより、人間の視覚でも区別しづらい、ひらがなの「へ」とカタ

カナの「へ」、カタカナの「カ」と漢字の「力」などの識別ができるようになり、その後の作業の省力化が期待できる。ただし、いかなる書籍に対しても効果があるかがわからないため、OCRの精度向上と共に検討していく必要がある。

なお、今回の実証実験で使用したOCRは、市販のOCRと同じくJIS第2水準までの対応だったため、JIS第3、第4水準の文字を正確に読み取ることができなかった。テキスト化作業における校正作業を効率的に進める上では、OCRで認識できる文字の範囲は広いほど望ましい。読取精度の向上とともに、OCRで認識できる文字の範囲が拡大することが、今後の技術的な課題である。

5. 2 全文テキストデータの検索・表示に関する実証実験の成果と課題

(1) 全文テキストデータの検索・表示に関する検証の成果と課題

本実証実験では、テキスト化システムにより生成された構造情報を利用し、書籍の目次・本文・索引という構造単位で検索対象を限定して検索できる機能を実装した。また、全文テキストデータを活用して書籍の内容を把握するため、文脈検索、タグクラウド、固有名表示、引用書籍表示等の機能を実装した。このうち、書籍で頻出する単語を表示するタグクラウド機能に対しては、本文表示前の段階で書籍の内容が把握できるというアイデアが評価された。また、文脈検索機能は、タグクラウドや固有名表示に表示される単語が本文でどのように使われているかを短時間で調べることができるという評価を得た。

以上のことから、構造情報を活用した検索機能と、全文テキストデータを活用した書籍の内容を把握するための機能が有効であることが実証できた。

全文テキストデータの表示では、構造情報を用いて書籍の目次を表示し、目次から本文へのリンクを実現した。また、目次ごとに検索語が何回出現したかを表示できる機能について、検索語が多く出現する章や節がわかるため、興味がある箇所、関係の強い箇所に早くたどりつけるという評価を得た。

しかしながら、テキスト化システムで生成されたデータの特徴である構造情報については、検索や表示において、更なる活用の可能性があると考えられる。

検索の面では、まず基本的な検索機能自体についての改善が必要である。検索に関する技術分野には以下のようなものがあり、本実証実験ではそのいずれの分野においても改善できるポイントがあった。

- ・インデキシングの最適化
- ・クエリ³¹の切り分けなど日本語処理の高精度化
- ・揺らぎ、ミス、代替クエリの提案などの検索語の前処理

³¹ 検索の際にユーザが入力する単語やフレーズのこと。

- ・同一のクエリでも目的や意図が違う場合の判定
- ・検索結果のランキングアルゴリズム自体の改善
- ・クエリやクリックログを活用したチューニング
- ・第三者的な検索結果の品質診断
- ・検索のユーザインタフェースの最適化
- ・パケットテスト³²などのテスト手法の導入

また、ユーザの行動履歴によるフィードバックによって検索のロジックをチューニングすることは、多くの一般ユーザによる利用が想定される検索サー

ビスとして必須であり、そのためのログ取得機能が上流設計時に組み込まれている必要がある。

さらに、本実証実験の対象データの特徴である構造情報を活かして、利用者が求める情報に的確にナビゲートすることが必要である。書誌情報を対象としたメタデータ検索ではヒットしなかった書籍が全文テキストデータを対象とすることで見つかる可能性が高くなるのが全文検索のメリットだが、一方で、検索対象が膨大になることによりノイズが多くなるというデメリットもある。このデメリットの解消のため、構造化された情報を活用することによって、有効な検索結果を得られるようにすることが求められる。

構造情報の活用方法として、目次だけでなく、見出しのみを対象とした検索も可能にしたり、書籍単位ではなく、章・節・項などの単位で検索できるようにすることなどがあげられる。また、検索対象が膨大になることから、検索結果の絞り込み機能も必要である。今回実装した、本文すべて・目次・索引の構造指定検索に加えて、見出しの階層を指定して絞り込み、より有用な情報にたどりつけるようにすることが求められる。

一方、表示の面では、本文全体を理解するのに役立つように、見出しの階層を組み合わせて表示できるようにすることが必要である。また、著作権等により本文が表示できない場合でも、今回実装した「検索語出現数表示」のようなイメージで、階層化された見出しとしての目次など、構造化された情報を視覚的に把握できるようにする工夫が求められる。

(2) 視覚障がい者等向けの読上げサービス等に関する検証の成果と課題

本実証実験における、視覚障がい者を対象にした読上げサービスの検証においては、目次で指定した位置から読上げが開始できる点について、利便性が高いと認められた。また、OCRの認識率が高くなく、誤認識された文字を多く含む全文テキストデータであっても、読上げによりおおその内容を把握できるため、十分に価値

³² ユーザーのアクセスを幾つかに分割してそれぞれ別々のユーザインタフェースを提示し、それぞれのクリックログなどから仕様の優劣を判定するテスト手法

があるという意見も得た。

一方で、OCR 認識率が低い書籍ほど、読上げられても正確に理解できないという結果も明らかとなった。これに対しては、5. 1 (4) に示したように、OCR 認識率を高度化することによる対応が必要である。

また、本実証実験で使用した書籍に、国立国会図書館の所蔵する明治時代から昭和戦前期の書籍が含まれており、読上げの際に旧字体の漢字や旧仮名遣いが正しく読み上げられないというケースがあった。これは、現在市販されている読上げソフトウェアが旧字体の漢字や旧仮名遣いに対応していないことに起因する。検索用のデータとしては新字体が必要であるものの、表示用のデータとしては、原本と同じ文字を表示してほしいという利用者のニーズがある。また、アクセシビリティの観点からは、視覚障がい者等が読上げで正確に理解できるようなデータが求められている。このように、用途により利用者が求めるデータが異なることから、検索表示システムにおいては検索表示用のテキストに加えて読上げ用テキストを用意するなど、用途に合わせて適切なデータを用意することが必要である。

校正が完璧に行われていない不完全なデータについては、注記つきで公開するなどの方法も考えられる。例えば、米国の **Bookshare** では、提供しているテキストデータについて、OCR で読み取っただけのものなのか、人手による校正が行われているのか、といったランク付けがされている。このような方法を採用する場合には、どこまでの品質を保証するのかということや、データを公開する際の基準の設定が課題となる。

最後に、今後これらの課題に取り組んでいくに当たっては、国立国会図書館単独ではなく、外部の研究者や開発者等との連携・協力を一層進めることが重要であることを指摘しておきたい。研究開発や実証実験に必要なデータやプラットフォームを国立国会図書館が提供し、当該領域に関心を有する研究者や開発者が知恵を出し合える場を用意するなど、技術的課題の解決に向けた取り組みを加速させる必要がある。