ヨミログ：読書ログによる個人に応じた読了時間推定システム

伊藤雄一[†] 縣 啓治[††] 高嶋和毅[†††]

Kindle や iPad などのタブレット端末の普及に伴い，電子書籍で読書することは一般的になりつつある．通勤途中や休憩中などに読書する際，その時間を活かして効率的に読書をしたいというニーズはあるが，これまで読了時間を推定するようなシステムは存在しなかった．そこで本研究では，読書ログを用いて読了時間を推定する「ヨミログ」を提案し，その実現に向けて本人の既読作品の読書データをもとに読了時間を推定する手法と，本人と他人の読書力の差を用いて読了時間を推定する 2 つの手法を検討する．評価実験の結果，前者は決定係数が 0.81 の重回帰モデルが構築でき，推定誤差の平均は 16.9%であった．また後者はモデルの決定係数が 0.78 で，推定誤差の平均は 12.0%と良好な結果が得られ，ある程度正確な読了時間の測定が可能でありこれら手法の有効性が示された．

YOMILOG: A System to Estimate Reading Completion Time by using Individual Reading Log

Yuichi Itoh,[†] Keiji Agata[†] and Kazuki Takashima[†††]

With spreading tablets such as iPad, Kindle, etc., reading digital books becomes common in our daily life. Sometimes people want to know how it takes a time to complete reading a book during their traveling or break times. Unfortunately, there is not such a system that estimates the reading completion time according to their reading skills. In this paper we propose “YOMILOG” system to estimate users’ reading completion time with their reading history. To realize it, we examined two methodologies (only individual logs, and individual and others’ logs). They were verified by using reading data from 8 participants with 8 books in a controlled experiment. As a result, the first methodology achieved a regression model with high fitness (ave. $R^2 = 0.81$) and an estimation error rate of 16.9%. The second one also provided a high fitness model (ave. $R^2 = 0.78$) and a lower estimation error rate of 12%. These results clearly demonstrated effectiveness of the proposed methodologies.

1. はじめに

2010 年は電子書籍元年と言われているように，電子化された書籍を Kindle や iPad などの端末で閲覧する機会が増加している．例えば，青空文庫☆では，著作権の消滅した作品が無料で公開されている．作品の電子化は現在も進められており，2011 年 2 月現在で約 9900 作品が公開されている．電子書籍の場合は，簡単に大量のデータを大量に持ち運べるため，外出中でもその日の気分に合わせて読む作品を決め，同時に複数の作品を読み進めるといったことも可能となり，読書スタイルが従来に比べ多様化してきている．このように，電子書籍端末が普及し，電子書籍による読書が読書スタイルの一つとして定着しつつある．このような状況において，電子書籍の特性を活かすことで，紙媒体では不可能だった，情報の自動取得・推定などが可能となると考えられる．

ユーザは新しく読む作品を選ぶ際に，他の読者によるレビューやお気に入り度などからその作品の難しさや面白さを見積もるが，レビューを投稿した人とユーザとの感性や理解力が異なり，参考にできないことも多い．また，移動時間や休憩時間などを利用し，効率的に読書したいという要望もあるが，内容が異なり，読みやすさが分からない作品に対する読了時間を推測することは困難である．そのため，個人に応じた読了時間や難易度，面白さなどを推定する必要がある．そこで本研究では，まずは，難易度や面白さ，文字を読

† 大阪大学 大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University
†† 大阪大学 大学院情報科学研究科 現在，富士フイルム株式会社
Graduate School of Information Science and Technology, Osaka University. Currently, Fujifilm Corporation
††† 東北大学 電気通信研究所
Research Institute of Electrical Communication, Tohoku University
☆ http://www.aozora.gr.jp/.

むスピードなど，個人によって異なる様々な要素が影響すると考えられる読了時間に着目し，その推定を可能とするシステム「ヨミログ」の実現を目指す．個人に応じた読了時間を推定することができれば，例えば移動時間に読み終えられる作品を選んだり，ある本のその人に対する難易度を読了時間から推し量るということが可能になると考えられる．

そこで，本論文では，様々な書籍カテゴリのうちから最も広く読まれているであろう小説を対象書籍カテゴリとし，個人の読書履歴から対象とする書籍の読了時間を推定する 2 つの手法を提案する．さらに実際にこれらの手法を用いて読了時間を推定した結果から，提案手法の有用性について考察する．

2. 関 連 研 究

これまで日本語の文章の特徴を解析し，様々な情報を推定する研究がなされている．

馬場らは小説テキストのジャンル推定手法[1] を提案している．この研究では，小説テキストを形態素解析し，TF-IDF 法を用いて各語に重み付けを行い，SVM (Support Vector Machine) を拡張した多値分類器を用いて小説のジャンル（「童話」,「現代物」など）を推定している．

川村は，学習および教育のための支援システムとして，テキスト中の語彙の難易度を判定する「語彙チェッカー」を実装し，日本語教科書の分析を行っている[2]．語彙チェッカーは，テキストに含まれるすべての語彙のレベルを日本語能力試験の 1 級から 4 級までの語彙レベルを規準に自動判定するものであり，学習者の「わからなさ」を疑似体験することもできる．松吉らは，文体と難易度を制御しつつ，日本語を言い換える手法を提案している[3]．例えば，日本語初級者用や日本語中級者用のように難易度指定をすることで，難解な文章を言い換えるといったことが可能になっている．中條らは，日英パラレルコーパスを構成している日本語テキストおよび英語のテキストの難易度を，教育的視点に立った客観的な指標を用いて計測することによって，難易度によるテキスト分類の可能性を確認している[4]．特に日本語テキストについては,「日本語能力試験の語彙 1，2 級の割合」,「漢字含有率」というテキストの内容理解に直接影響する指標を用いて，おり，物語文は説明文より一般的に易しい傾向があるということなどが明らかにされている．建石らは，読みやすさに関係のある項目として，文字数，各文字種についてその文字種の連の相対頻度，文字種ごとの連の平均の長さ，読点の数の句点の数に対する比などを用い，構文や意味によらずにその文章の読みやすさを評価する式を求めている[5]．近藤らは，テキストの難易度のレベルが直感的に分かりやすくなるように，小学 1 年から高校 3 年までの 12 学年に大学を加えた 13 の区分で，英語を除く全教科の教科書をテキスト収集源として作成した規準コーパスを利用して，テキストの難易度を表している[6]．一方で，柴崎らは，小学校の学年を予測したリーダビリティーの公式を生成している[7]．この公式には，文中の一つの項と一つの述部からなる単位を IU と定義し，文章中のひらがなの割合，一文中の IU の割合，文章中の全内容の漢語の割合を利用している．しかし，これらの手法では，日本語文章の一般的なリーダビリティを推定することにとどまっており，個人に応じた文章の難しさや読みやすさの推定には至っていない．

3. 電子書籍によるデータの取得実験

3.1 概　　要

図 1 に示すインタフェースを用い，8 名のユーザ (男性 7 名，女性 1 名，平均年齢 23.6 歳の大学院生) にランダムな順番で 8 作の中長編小説を読んでもらい，実際に各ユーザの読了時間等のデータを取得する実験を実施する．実験参加者には

- 日頃 PC でテキストを読むようなイメージで読書すること
- 理解しながらできるだけ集中して読むこと
- 読書は途中で中断しても良い

と教示する．この 8 名のユーザを以下では，ユーザ A〜H とする．このうち 6 名は短編作品を用いた予備実験にも参加している．読書した作品は，**表 1** に示す青空文庫で公開されていて，実験参加者全員が読んだことのない，8 作品であり，小説の代表カテゴリとしてミステリー系 4 作，文学系 4 作とした．

3.2 実 験 環 境

3.2.1 電子書籍インタフェース

実験に利用したインタフェースの詳細について述べる．図 1 にインタフェースの概観，**表 2** にインタフェースの仕様について示す．短編作品を読書してもらった予備実験 (文章は横書き表示) において,「縦書きの方が読みやすい」,「小説は縦書きで読みたい」といった意見が多かったため，文章の表示形式を縦書きとする．インタフェースのサイズは，13.3 インチの LCD (1,024 × 768 ピクセル，96 dpi) とする．一般的な文庫本の文字量は 40 文字× 17 行 (680 文字前後) であるが，これは縦書き文章の場合，横書きの 1.5 倍程度の許容量があり，一行が 40 文字程度であれば，読む

表 1　実験で用いた 8 作品の概要
Table 1　Overview of 8 stories used in experiment

作品	タイトル	文字数 [字]	ページ数	章	著者	ジャンル
作品 1	金狼	90895	200	10	久生 十蘭	ミステリー
作品 2	透明人間	89830	200	5	H・G・ウエルズ	ミステリー
作品 3	恐怖城	80810	189	9	佐左木 俊郎	ミステリー
作品 4	黄鳥の嘆き	42329	93	6	甲賀 三郎	ミステリー
作品 5	坊っちゃん	103426	203	11	夏目 漱石	文学
作品 6	斜陽	93920	207	7	太宰 治	文学
作品 7	人間失格	75036	156	4	太宰 治	文学
作品 8	水晶の栓	64843	139	7	モウリス・ルブラン	文学

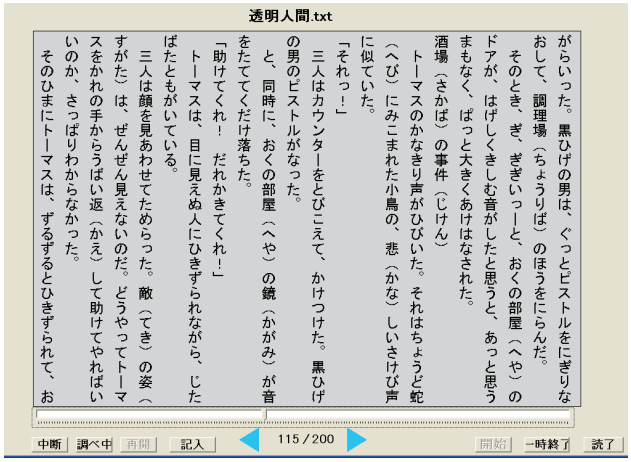

図 1　電子書籍インタフェースの概観
Fig. 1　Overview of interface

ことに苦痛を感じないといわれているためである．そこで，一行は 40 文字前後を超えないように実装する．フォントについては，ディスプレイに文章を表示する場合，文字サイズに関係なくゴシック体の方が明朝体よりもユーザの理解度が高い[8] こと，また，18pt の文字サイズ (この実験環境において約 24 ピクセル相当) のゴシック体では，液晶ディスプレイ (LCD) に表示した場合と印刷した場合の理解度に有意差がなく，どちらも高い理解度が得られるという予備実験の結果から，フォントは 24 ピクセルのゴシック体（MS ゴシック）とした．行間については，適度な行間は 0.4〜0.9 文字分の範囲内で最適行間は 0.7 文字分ということが検証されており[9]，さらに「Design Basic Book-初めて学ぶデザインの法則」[10] によると，大量の文字を組む場合の行間は，0.5〜1.0 文字分が目安となると述べられていることから，行間は 0.7 文字分とする．また，標準的なレイアウトは 35 文字× 15 行の 525 文字[11] といわれており，文庫本では 680 字前後となっているため，上記で述べた行間などを考慮に入れながら，その範囲内の文字量を 1 ページに表示するようにした．文字の表示領域における背景色は可読性が高いことが

表 2　電子書籍インタフェースの仕様
Table 2　Specification of interface

項目	詳細
画面サイズ	13.3 インチ（1,024 × 768 ピクセル，96 dpi）
表示形式	縦書き
文字量	25 文字× 22 行 (1 ページ：550 字)
フォント	MS ゴシック，24 ピクセル，黒色
行間	0.7 文字分
背景色	ライトグレー

実証されている[12] ライトグレーを使用する．読書作品のタイトルは，どちらも本文の表示画面の上部に表示している．

3.2.2　読 書 手 順

一般的にユーザが読書中に行うと考えられる操作を予備実験などから選び出し，それらを記録するために次に示す操作を GUI によって可能とする．

- 開始ボタン：作品の文章，ページ数を表示する (読書を開始する際に使用)．
- 左向き/右向き三角マーク：1 ページ進む/戻る．
- TrackBar：大きくページを移動する (読み返し時に使用)．
- 中断ボタン：読書を中断する（休憩時などに使用）．
- 調べ中ボタン：分からない言葉などを調べる．
- 再開ボタン：中断，調べ中から読書を再開する．
- 記入ボタン：評価を記入する（今回は使用しない）．
- 一時終了ボタン：プログラムを一時的に終了する (日をまたいで読書する際や長く時間を空ける際などに使用)．
- 読了ボタン：プログラムを終了する (作品の読了時に使用)．

上記のインタフェースによるユーザの読書の流れについて述べる．ユーザはまず最初に，「開始ボタン」をクリックする．すると，1 ページ目の文章と，全ページ数，現在閲覧しているページ番号が表示される．そこから，ユーザは基本的に，「左向き三角マーク」をクリックし，1 ページずつページを読み進めていく．前のページに戻って読み返す際は，「右向き三角マーク」

をクリックし，1 ページずつ戻るか，大きくページを移動したい際には,「TrackBar」のツマミをドラッグして左右に動かすことで移動する．ドラッグ移動中にはページ番号に対応したページの文章が表示されるため，その内容を確認しながら読み返したいページを探すことができる．読書を中断したいと思った際は,「中断ボタン」をクリックする．作品中に自分の知らない語句が出てきて，意味を知りたいと思った際には,「調べ中ボタン」をクリックし，語句の意味を調べる．中断や調べものが終わり，読書を再開する際には,「再開ボタン」をクリックし，通常の操作で読書を続けてもらう．中長編作品は一度に読み切ることが難しいと思われるため，一旦終了した後，その時点から読書を再開できる「一時終了ボタン」を用意した．なお、「中断ボタン」「調べ中ボタン「一時終了ボタン」を押下した際には対象となる文章は非表示とした．作品全体を読み終えた際には，読了ボタンをクリックすることで，プログラムが終了し，一つの作品を読み終えたことになる．

3.2.3 取得データ

この実装環境において取得するデータは，一般的な読書中のユーザの行動を考慮した読書データと，作品の文章構造を表す固有データとする．

読書データ

読書データは次の 5 項目である．

読了時間　読書を開始してから，最後に読了ボタンを押すまでの時間

スピード　読了時間を文字数で割って算出する

読み返し回数　読書中に以前のページに戻った回数

調べ回数　分からない語句などが出てきた際に調べ中ボタンを押して調べ物をした回数

中断回数　読書を中断ボタンを押して中断した回数

固有データ

作品の文章構造を表す固有データとして，日本語文章のリーダビリティ（難易度）の算出などに利用される**「文字数」,「平均文長」,「漢字含有率」,「2 級以上の語彙の割合」**の 4 項目に，小説特有の要素として**「セリフ割合」**を加えた 5 項目とする．これは，予備実験に参加したユーザから「セリフ部分は速く読めた」という意見が得られ，セリフ割合が読了時間に影響すると考えられたためである．平均文長，漢字含有率，2 級以上の語彙含有率に関しては，日本語教科書の分析によると，学年が上がるのに応じて (難易度が上がるごとに応じて) これらの指標が大きくなるという結果が得られておりこれらの項目が日本語文章のリーダビリティに関係するとされている[13] ためである．ここで，2 級以上の語彙とは，日本語能力試験 (JLPT)☆☆において 2 級および 1 級試験に使用されている語彙のことである．

4. 読了時間の推定手法

実験により得られた 8 名 × 8 作品分のデータを用いて，個人ごとに未読の書籍の読了時間を推定する手法を検討する．書籍の固有データは 3.2.3 節で述べたように作品ごとに取得できるが，読書データは既読の書籍からしか取得できない．そこで，本人の既読の読書履歴をもとに読了時間を推定する手法（手法 1）と，本人と他人の読書履歴をもとに推定する手法（手法 2）の 2 つを検討する．手法 1 も手法 2 も，読了時間は書籍の固有データと個人の読書力に依存するという考え方に基づいており，手法 1 は他の本がこれくらいの時間で読めればこの本はこれくらいの時間で読めるはずと推定する手法であり，手法 2 は本人と他人のある本の読了時間の差が読書力の差であるとみなし，対象となる本を他人が読んだ読了時間と読書力の差から本人の読了時間を推定する手法である．

以下では，ユーザ A の作品 1 に対する読書データを A_1 と表記する．

4.1 手法 1：本人の読書履歴からの推定

あるユーザによる未読作品の読了時間を，そのユーザ本人の既読作品における書籍の固有データと読了時間を用いて推定する．本研究で定義した書籍の固有データの 5 つの指標「文字数」,「平均文長」,「漢字含有率」,「2 級以上の語彙の割合」,「セリフ割合」のうち，時間を推定するのに最適な組み合わせが何かを検証する．8 作品の固有データの各指標を表 3 に示す．

まずは，これらの 5 つの固有データを用いて，ステップワイズ法により，各ユーザの各作品に対する読了時間についての重回帰分析を行った．その結果，得られた 8 名 ×8 作品分の 64 個の重回帰式において，文字数が 59 回，漢字含有率が 37 回，2 級以上の語彙が 21 回，セリフ割合が 8 回，平均文長が 2 回使用されており，特に文字数と漢字含有率が，全員に共通して多く利用されていることがわかった．したがって，文字数，漢字含有率を組み合わせた式 (1) に示す重回帰式で未読作品の読了時間をモデル化する．ここで，a_1，b_1，c_1 はユーザごとの各作品に対して決まる定数であり，ユーザの読書力によって変化する定数である．

$$\text{読了時間 [秒]} = a_1 \times (\text{文字数}) + b_1 \times (\text{漢字含有率}) + c_1 \qquad (1)$$

☆☆ http://www.jlpt.jp/about/index.html.

表 3　実験で用いた 8 作品の固有データ
Table 3　Detailed data of 8 stories used in experiment

作品	文字数 [字]	平均文長 [字]	漢字含有率 [%]	セリフ割合 [%]	2 級以上の語彙の割合 [%]
作品 1	90895	24.0	20.8	41	13.5
作品 2	89830	27.8	10.1	45	16.6
作品 3	80810	21.2	23.9	43	37.4
作品 4	42329	35.4	30.6	17	13.3
作品 5	103426	31.7	24.4	15	14.8
作品 6	93920	29.4	20.7	20	13.5
作品 7	75036	40.6	24.4	12	14.0
作品 8	64843	28.4	27.0	41	14.4

4.2　手法 2：本人と他人の読書履歴からの推定

ユーザ本人と他人の既読作品の読書データから，ユーザごとの読書力を算出し，他のユーザの対象作品の読了時間をもとに，そのユーザの対象作品の読了時間を推定する．3.2.3 節で述べた取得できる読書データのうち，中断回数および読了時間以外のスピード，読み返し回数，調べ回数をもとに読書力を算出する．これは中断回数にはトイレ休憩や電話対応など，ユーザがやむを得ず読書を中断せざるを得ない状況により，ユーザ自身の読書力に関係なく中断してしまう場合が多く存在すると考えられるためである．読書データと読書力の関係について，次のことが言える．

- スピード：速いほど読書力は大きい
- 読み返し回数：少ないほど読書力は大きい
- 調べ回数：少ないほど読書力は大きい

そこで，既読作品におけるこれらの 3 項目を用いて，読書力を式 (2) のように定義する．SPEED，REVERSE，SEARCH はそれぞれ，スピード，読み返し回数，調べ回数を，同じ作品を読んだユーザのものと比較し正規化した値である．j は既読作品を，n は既読作品の総数を表す．予備実験の結果から読了時間にスピードが大きく影響すると考えられること，読み返す，調べるといった行為も実験参加者ごとに異なって観察されたことから，経験則的に $\alpha = 0.5$，$\beta = 0.25$，$\gamma = 0.25$ とする．

$$\text{読書力} = \sum_{j=1}^{n} (\alpha \times SPEED_j + \beta \times (1 - REVERSE_j) + \gamma \times (1 - SEARCH_j)) \ / \ n \qquad (2)$$

あるユーザの未読作品に対して，その作品を読んだ他のユーザの読書力と読了時間をもとに，未読作品の読了時間を重回帰分析によりモデル化する．これにより，式 (3) に示すモデルで読了時間を推定できる．a_2，b_2 はユーザごとの各作品に対して決まる定数である．

$$\text{読了時間 [秒]} = a_2 \times (\text{読書力}) + b_2 \qquad (3)$$

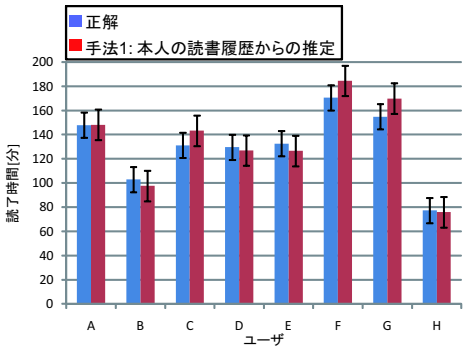

図 2　ユーザごとの読了時間推定の結果 (手法 1)
Fig. 2　Result for each user by methodology 1

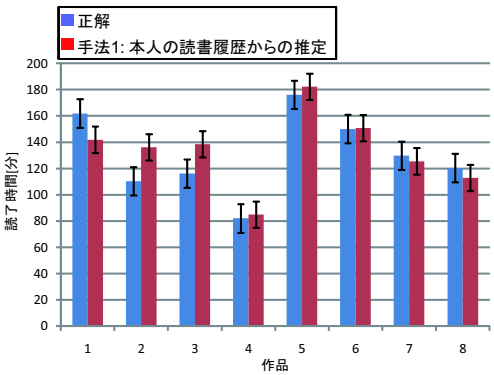

図 3　作品ごとの読了時間推定の結果 (手法 1)
Fig. 3　Result for each story by methodology 1

5. 結　　果

5.1　手法 1：本人の読書履歴からの推定

本人の読書履歴から，書籍の固有データを用いて読了時間を推定する手法 1 では，A_1〜H_8 の読了時間を推定する重回帰式において，平均の決定係数 R^2 は 0.81 であった．例えば，ユーザ A の未読作品が作品 1 である場合は，**表 4** に示すようなユーザ A の既読作品の読了時間，文字数，漢字含有率をもとに重回帰分析を行い，式 (4) に示す重回帰式が得られた．これを全ユーザごとに繰り返す交差検定を行った．

表 4　A_1 の読了時間を推定する際に利用するデータ
Table 4　Data used to estimate A_1's reading completion time

	A_2	A_3	A_4	A_5	A_6	A_7	A_8
読了時間 [秒]	7403	8664	6112	12134	9645	8919	7264
文字数 [字]	89830	80810	42329	103426	93920	75036	64843
漢字含有率 [%]	10.1	23.9	30.6	24.4	20.7	24.4	27.0

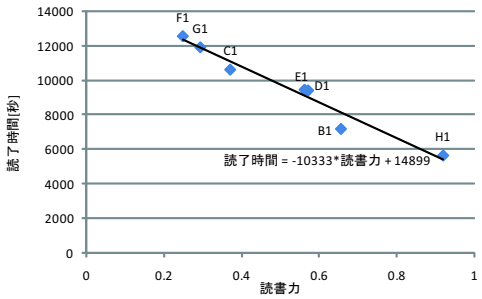

図 4　A_1 の読了時間を推定する重回帰式
Fig. 4　Multiple regression formula to estimate A_1's reading time

$$\text{読了時間 [秒]} = 0.118 \times (\text{文字数}) + 203 \times (\text{漢字含有率}) - 5366 \quad (R^2 = 0.96) \qquad (4)$$

ユーザごとに 8 作品の実際の読了時間を平均した結果と推定した 8 作品の読了時間を平均した結果を**図 2** に示す．なお，正解と推定誤差の誤差率の絶対値の平均は 16.9 %であった．図 2 のユーザ A とユーザ H を見ると，読了時間が個人によって大きく異なっている場合にも，個人に対応した推定ができていることがわかる．また，作品ごとに 8 名のユーザの実際の読了時間を平均した結果と推定した 8 名のユーザの読了時間を平均した結果を**図 3** に示す．図 3 からわかるように，ジャンル（今回はミステリーと文学小説）や文章の構造が異なる作品に対してでも，読了時間を推定することが可能であると言える．

5.2　手法 2：本人と他人の読書履歴からの推定

ユーザ本人と他人の既読作品の読書データから，ユーザごとの読書力を算出し，他のユーザの対象作品の読了時間をもとに，そのユーザの対象作品の読了時間を推定する手法 2 では，A_1〜H_8 の読了時間を推定する重回帰式において，平均の決定係数 R^2 は 0.78 であった．

例えば，A_1 の読了時間を推定する場合は，**表 5** に示すような他人の読書力と読了時間をもとに重回帰分析を行い，式 (5) に示す重回帰式が得られた．この処理を 8 ユーザごとに行う交差検定を実施した．

$$\text{読了時間 [秒]} = -10333 \times (\text{読書力}) + 14899 \quad (R^2 = 0.96) \qquad (5)$$

この重回帰式と推定に用いたユーザ B〜H の読書力と B_1〜H_1 における読了時間の関係は**図 4** に示すようになっており，読書力が高いほど，読了時間が短くなっているということがわかる．

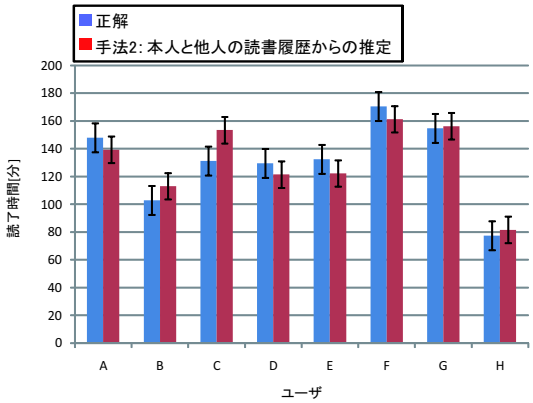

図 5　ユーザごとの読了時間推定の結果 (手法 2)
Fig. 5　Result for each user by methodology 2

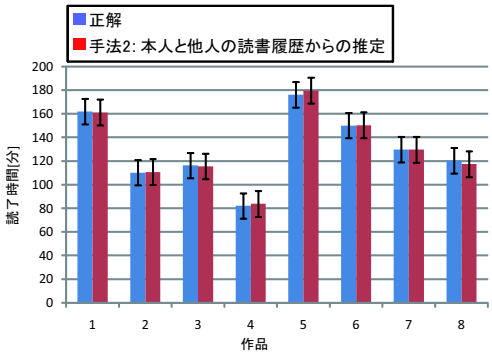

図 6　作品ごとの読了時間推定の結果 (手法 2)
Fig. 6　Result for each story by methodology 2

ユーザごとに 8 作品の実際の読了時間を平均した結果と推定した 8 作品の読了時間を平均した結果を**図 5** に示す．また，作品ごとに 8 名のユーザの実際の読了時間を平均した結果と推定した 8 名のユーザの読了時間を平均した結果を**図 6** に示す．

このようにして推定した読了時間と実際の時間との誤差率の絶対値の平均は 12%であった．図 5 のユーザ B と F を見ると，読了時間が個人によって大きく異なっている場合にも，個人に対応した推定ができていることがわかる．しかし，ユーザ C のように誤差が平均して約 25 分程度と大きくなってしまうこともある．これは個人により，読了時間と相関のある項目に

表 5　A_1 の読了時間を推定する際に利用するデータ
Table 5　Data used to estimate A_1's reading completion time

	B_1	C_1	D_1	E_1	F_1	G_1	H_1
読了時間 [秒]	7197	10627	9422	9464	12566	11921	5657
読書力	0.656	0.371	0.572	0.562	0.249	0.294	0.919

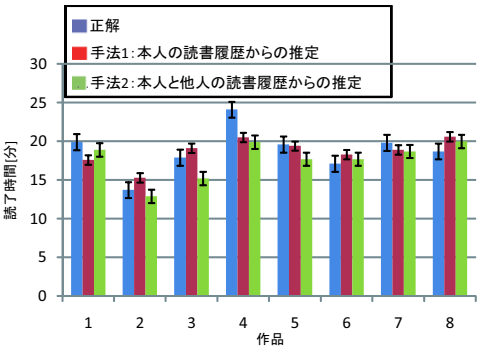

図 7　各作品での 1 万字に対する読了時間
Fig. 7　Reading completion time for 10000 letters with each story

差があるためであると考えられ，読書力の算出時に，相関が高い項目の重みを高くするといった改善が必要であると思われる．

5.3　1 万字あたりの推定

作品によって，読了時間が異なっていることはわかったが，同じ分量に対してでも読了時間に差があるかを確認するために，各作品における 1 万字に対してかかった実際の読了時間を求めた．また，1 万字に対する実際の読了時間と 2 手法により推定した読了時間を**図 7** に示す．図 7 から，同じ分量でも作品によって，読み終えるのにかかる時間が異なっていることがわかり，実際に文字数や漢字含有率だけではなく，作品の内容や難易度に応じて読了時間が異なることがわかる．また，2 手法とも各作品に対して異なる読了時間を推定出来ていることがわかる．

表 6 に，ユーザごとの各作品での 1 万字に対する読了時間を示す．この表 6 からユーザ A は作品によって，1 万字での読了時間に最大で 10 分程度の差がある（作品 2: 13.7 分，作品 4 :24.1 分）が，ユーザ H は 4 分程度の差（作品 2: 8.2 分，作品 8: 12.0 分）となっており，個人によって，作品の内容や難易度が読了時間へ与える影響度も違うことがわかった．

6. 考　　察

本研究で提案した「ヨミログ」のための，本人の読書履歴からの読了時間の推定手法（手法 1）と，本人と他人の読書履歴からの推定手法（手法 2）の 2 手法とも，個人に対応した推定ができているといえる結果となった．手法 1 は個人の履歴のみを参照するだけで読了時間を推定できるため，ほとんどの人が読んでいないマイナーな作品や新しく出版された作品を読書する際などにも適用できる．一方手法 2 は，ベストセラーなど，他人が多く読んだ作品の読了時間の推定に応用できる．しかし，誤差が大きい被験者の例もあり，またモデルについても改良の余地がある．今回，モデルの構成要素 (式 (1) や (2)) は全体の傾向から決めており，ユーザごとに係数は変更する形ではあったが完全にパーソナライズされたものではない．これが，個人差や誤差が大きい事例の主な原因だと考えられる．また，式 (5) に示したように，推定式の係数がマイナスを取るなど，ユーザによっては少し扱いづらい推定式が算出されることなどもあった．推定結果により，モデルや推定アプローチの有用性は示せたが，サンプル数を増やして，提案モデルの適用範囲を明らかにすることや，パラメータの逐次選定などモデルの改良が求められる．

手法 1 と手法 2 において，推定の精度に大きな差は見られなかった．これは，読了時間が作品の固有データに依存しやすいためと考えられる．また本実験ではジャンルを変数として調査はしていないが，少なくとも今回の結果からはミステリーと文学小説に顕著な差は見られなかった．しかし，技術書や専門書など，読解に読者の知識を必要とするような書籍についてはジャンルによって大きな差が出ると考えられるが，今後の課題である．一方，誤差について，全体では，読了時間が約 2 時間程度の作品に対し，15 分程度となっている．その多寡はユーザの状況に依存するものの，読書は継続的に行うものであることから，データセットが増えればこの誤差は改善できると考えられる．また，違う作品の同じ分量の読了時間の比較においては，文字数と漢字含有率以外の要素（内容や難易度）が影響を与えていることが明らかになり，このような要素をいかに取得するかも今後の課題である．

本実験では，20 代の学生を対象に読書データを取得したが，幅広い年齢層で読書データを取得し，年齢層が違うことで，読書データに差が見られるかどうかなども，今後検証していきたい．

従来の電子市場や書店での書籍の整理方法は，書籍のジャンルや学年別などであるが，本提案手法により，

表 6 ユーザごとの各作品での 1 万字に対する読了時間 [分]
Table 6 Reading completion time for 10000 letters for each user with each story

ユーザ	作品 1	作品 2	作品 3	作品 4	作品 5	作品 6	作品 7	作品 8
A	19.9	13.7	17.9	24.1	19.6	17.1	19.8	18.7
B	13.2	9.0	13.5	6.6	13.3	12.5	13.4	14.5
C	19.5	12.6	15.3	19.8	15.3	16.3	17.8	16.5
D	17.3	12.9	12.2	18.0	17.9	15.8	16.3	20.7
E	17.4	12.3	13.5	20.5	19.1	14.8	19.0	18.0
F	23.0	14.2	19.1	22.8	21.4	21.7	23.5	27.1
G	21.9	14.6	17.4	19.8	18.3	19.5	20.3	24.5
H	10.4	8.2	8.3	11.3	10.1	8.7	9.4	12.0

個人専用の読了時間別の本棚により書籍を分類することが可能になり，移動時間や休憩時間などを利用し，効率的に読書することが可能になると考えられる．また，本提案手法により，作品全体だけでなく，章ごとなど細かい単位での読了時間を推定することもできると考えられるため，長編作品を数回に分けて，切りのいいところまで読むという場合においても，本手法が適用できると考えられる．

7. おわりに

本研究では，ユーザの読書履歴を用いて書籍の読了時間を推定する「ヨミログ」を提案し，その実現手法として，本人の読書履歴から読了時間を推定する手法（手法 1）と，本人と他人の読書履歴から推定する手法（手法 2）の 2 手法について検討した．次に，中長編の小説 8 作品を，8 名の参加者にランダムな順番で読んでもらい，得られた 8 名 ×8 作品分のデータを用いて，2 つの手法の有効性を検証した．その結果，決定係数および誤差が手法 1 では 0.81 と 16.9%，手法 2 では 0.78 と 12.0% となった．全体では，約 2 時間程度の作品に対し，15 分程度の誤差で推定できており，個人に応じた読了時間をある程度正確に推定できたと言える．

今後は，読書力の算出に用いる新たな指標の検討や重みづけなどについてさらに検討を進めていきたい．

参 考 文 献

1) 馬場こづえ，藤井敦，石川徹也：小説テキスト自動分類のためのジャンル推定と人物抽出，第 4 回情報科学技術フォーラム講演論文集，pp. 67–70（2005）．
2) 川村よし子：語彙チェッカーを用いた日本語教科書の分析，in *proc. of CASTEL/J '99*，pp. 132–137（1999）．
3) 松吉俊，佐藤理史：文体と難易度を制御可能な日本語機能表現の言い換え， 自然言語処理，Vol. 15, No. 2, pp. 75–99（2008）．
4) 中條清美，白井篤義，内山将夫，西垣知佳子，長谷川修治：日英パラレルコーパスを構成するテキストの難易度分類に関する研究，日本大学生産工学部研究報告，pp. 57–68（2004）．
5) 建石由佳，小野芳彦，山田尚勇：日本文の読みやすさの評価式，情報処理学会研究報告，Vol. 88, No. 25, pp. 1–8（1988）．
6) 近藤陽介，松吉俊，佐藤理史：教科書コーパスを用いた日本語テキストの難易度推定，言語処理学会第 14 回年次大会発表論文集，Vol. 14, pp. 1113–1116（2008）．
7) 柴崎秀子，沢井康孝：国語教科書コーパスを応用した日本語リーダビリティー構築のための基礎研究，電子情報通信学会技術報告，Vol. 2007, No. 32, pp. 19–24（2007）．
8) 清原一暁，中山実，木村博茂，清水英夫，清水康敬：文章の表示メディアと表示形式が文章理解に与える影響，日本教育工学会論文，pp. 117–126（2003）．
9) 坂本留美，岸田保：操作表示における読みやすい行間の定量化，松下電工技法，Vol. 52, No.3, pp. 52–59（2004）．
10) 生田信一：Design Basic Book - はじめて学ぶデザインの法則，BNN 出版社（2007）．
11) ジャムハウス：電子書籍の作り方ハンドブック，アスキー・メディアワークス（2010）．
12) 李明姫：ウェブデザインにおけるテキストの制御方法，日本デザイン学会，pp. 234–235（2001）．
13) 柴崎秀子，原信一郎：12 学年を難易尺度とする日本語リーダビリティー判定式， 計量国語学，Vol. 27, No. 6, pp.215-232（2010）．