

ライトノベルの余白分析方法の検討

坂東 将光* 瓦谷 嵐**

Investigation of Margin Analysis Method of Light Novels

Masamitsu BANDO *, Arashi KAWARATANI **

Recently, it has been known that the youth reading (novels, magazines, internet columns, and so on) has decreased. On the other hand, light novels are selling well, and internet novels are also very popular among youth. These facts indicate that analyzing these properties can reveal the tendency for young people to read. In this paper, we propose a new method for margin analysis of various light novels using Python and openCV, and we show the verification results of the method.

keyword image analysis, margin, light novel

1. 緒言

近年、若者の読書量が減少していることが知られている。全国大学生生活共同組合連合会が行った第 54 回学生生活実体調査 [1] では、図 1 に示すとおり大学生のうち 1 日の読書量は 2012 年以降減少の一途を辿っており、全く読書をしない大学生も 6 年間で約 14 % 増加している。同調査では高校までの読書時間分布も調査しているが、その結果からは図 2 に示すとおり小学校卒業後からの読書量の減少が著しいことがわかる。

このように若者の読書量は減少しているが、一方でライトノベルの売行きは好調である。オリコンが発表している 2019 年年間本ランキング [2] によると、文庫カテゴリの売上ランキング上位 25 位のうち 1 位および 10 位がライトノベルであり、全体的に売上部数が文庫より多い BOOK カテゴリにおいても、売上ランキング 31 位および 34 位がライトノベルであった。また、ライトノベルと同様のものとしてインターネット上の小説 (以下ネット小説と呼ぶ) があり、ネット小説プラットフォームの 1 つである「小説家になろう」を始めとしてこちらも活況を呈している。

株式会社カイユウが 2019 年に行った、小説家になろうを運営するヒナプロジェクトへのインタビュー [3] では、同サイトの一ヶ月あたりのアクセス数が 20 億回、ユニークユーザが 1,400 万人程度であり、登録ユーザの約 58 % が 10~20 代であることが示されている。

昔と違い、昨今では身の回りに娯楽が多く存在し、インターネット上にもゲームを始め無料の娯楽が充実しているため、人はどの娯楽で楽しむかをより自由に選択できる。そんな中においてライトノベルおよびネット小説が好調であるということは、これらにそれだけの若者を惹き付ける要因があるということに他ならない。その要因としては以下が考えられる。

- (a) 文章が全体的に平易かつ会話文が多く読みやすい。
- (b) 登場人物の多くがアニメや漫画に出てくるような特徴的なキャラクターであり馴染み深い。
- (c) ページ数は小説に近いが、余白が多いため文章量が少なく気軽に読める。
- (d) 若者受けする絵が表紙や挿絵にふんだんに使われており、手に取りやすい。
- (e) (ネット小説のみ) スマートフォンで手軽に読める。

このうち文章に関係するのは (a) から (c) の 3 項目である。これらの要因を解析することで、若者が読みたくなる文章がどのようなものか分かり、読書量の改善に繋げられるほか、学校教材等への応用も期待できる。

本研究では、上記 (a) から (c) の要因のうち作品著者の技量に依らず、かつ定量的に解析できるものとして (c) の余白に着目し、その形状や量を画像解析によって分析することで、若者が読みたくなる文章がどのようなものかを明らかにすることを最終目的とし、そのための余白解析の方法を検討した。本論文ではその方法を紹介するとともに、今後の解析内容についても述べる。

*近畿大学工業高等専門学校
総合システム工学科 制御情報コース

**近畿大学工業高等専門学校専攻科
生産システム工学専攻

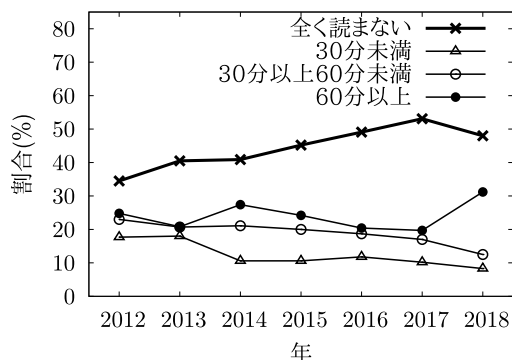


図 1: 大学生の 1 日あたりの読書時間分布。年々読書にかかる時間が減少していることがわかる。(参考: 第 54 回学生生活実体調査の概要報告, 全国大学生生活共同組合連合会)

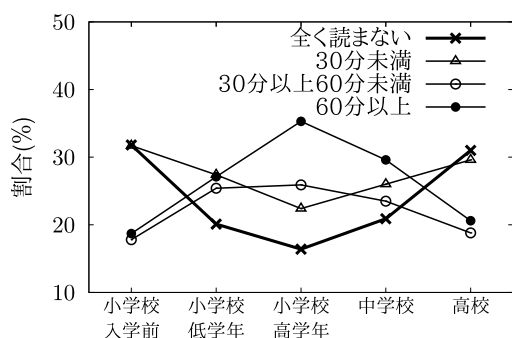


図 2: 高校までの 1 日あたりの読書時間分布。小学校卒業後から読書時間が著しく減少していることがわかる。(参考: 第 54 回学生生活実体調査の概要報告, 全国大学生生活共同組合連合会)

2. 本研究における余白の定義と解析対象

一般的に余白とは書物においては文字のない部分を意味するが、小説、ライトノベル等においてとりわけ余白があるのはページの上下左右の端にあるマージンと、文章を改行することでページ下部にできる領域である。これらのうち、マージンについては、小説であるかライトノベルであるかを問わずある程度の余白があるが、改行によって生まれる下部の余白については小説とライトノベルで大きな違いがある。会話文や比較的短い文章が多いライトノベルでは改行を多用するため、ページ下部に広い余白ができがちであり、作品によってはページの下部半分近くが余白となることがある。

そこで本研究では、解析対象の余白の定義を「改行によって生まれる文字のない領域およびページ下部のマージン」とした。ページ下部のマージンを定義に含むのは、これがほぼ一定であるため相対評価および変化量の評価において影響しない事と、解析の容易さが理由である。

また、ネット小説は横書きであることが多いうえに、Web ページであるため見る環境によって余白の量が変化してしまう。そのため、原則として文庫本として物理的に得られるライトノベルおよび小説を主な解析の対象とした。ただし、小説家になろうではネット小説を縦書きの PDF で出力できる

ため、小説家になろうでも比較的長期に渡ってある程度高いクオリティで書かれている共著者の 1 作品 [4] のみネット小説も対象としている。

3. 余白解析の方法

余白を検出するにあたり、本研究では言語として Python、画像解析ライブラリには OpenCV を用いた。本研究の方法では、余白解析を以下の順序で行う。

1. サンプルをスキャンし、画像データにする。
2. 読み込んだ画像を扱いやすいように処理する。
3. 様々な分割数で余白を検出し、最適な分割数を求める。
4. 求めた分割数で余白を検出する。

解析するサンプルは、文庫型のライトノベルおよび小説から適当なページを選出し、スキャンして画像にする必要がある。スキャン時のずれなどを補正する必要があるが、簡単な Linux の bash スクリプトを作成し自動化した。

サンプル画像はそのままでは扱いづらいので、OpenCV で二値化および白黒反転処理を行う必要がある。当初、画像のノイズ除去のために morphology 処理の 1 つである Opening 処理をする予定であったが、解像度の都合上明朝体の文字の線が細く消えてしまうため、ノイズ処理はしないこととした。

余白の検出と分割数

余白の検出は、画像の最下部から上部へ向けて、白 (文字) が見つかるまで順に 1 ピクセルずつ値を確認し、見つければそのときの高さ (最下部からのピクセル数) を保存する、という処理を左端から右端まで順に繰り返すようにした。横方向の一度の移動量は、横方向の分割数から、

$$\text{移動量} = \frac{\text{画像の横ピクセル数}}{\text{分割数}}$$

として求められる。分割数は明らかに 2 以上かつ画像の横ピクセル数以下の整数であるが、はじめは適切な値は不明である。そこで本研究の方法では、分割数を適当な刻み幅で変化させつつ何度も余白を検出し、得られた余白の面積が収束した時の分割数を採用することとした。このようにして決められた分割数で再度余白を検出することで、余白の量を知ることができる。

4. 余白解析方法の検証

本研究における余白分析の方法が正しく適用できることを確認するため、実際にそれぞれ異なる 7 作品のライトノベルからランダムなページを見開き 1 ページずつ選択し、ネット小説 1 作品を縦書き PDF 出力したものと合わせて 8 つのサンプルを元に余白の解析を行った。

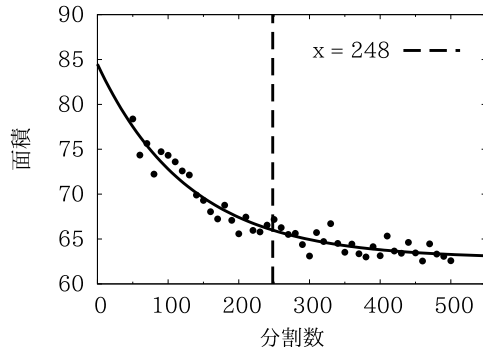


図 3: サンプルページの分割数の最適化。分割数 248 で閾値となることが分かる。

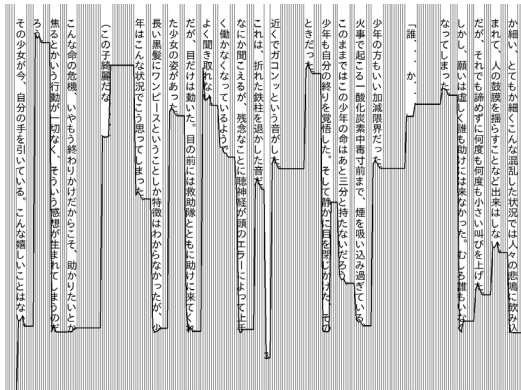


図 4: 余白の検出状況を可視化したもの。空行の余白検出に議論の余地がある。

4.1 分割数の最適化

サンプルの横のピクセルサイズは 2200px 程度であったため、分割数は 10 から 500 までも 10 刻みで変化させて余白の高さを検出し、台形法で余白の面積 S を求めた。何れのサンプルも図 3 のように分割数に対して指数関数的な減少を見せたため、分割数を x として、

$$S = a \exp(-bx) + c$$

でフィッティングして a 、 b 、そして c を決定し、 c の値の 105 % を閾値として、閾値における分割数を求めた。最大は 289 であったため、300 を以降の統一の分割数とした。分割数が大きい分には変化は 5 % 以内であるから問題ない。

4.2 余白の検出

上で得られた分割数 300 を用いて余白の検出を行った。図 4 はその結果を視覚化したものである。文章がない行の余白の定義が十分でないため、うまく余白を検出できていない部分が見受けられるが、本論文の方法によって高い精度で余白を検出できていることがわかる。

5. 結言

本論文では、若者が読みたくなる文章の傾向を把握するための方法として、ライトノベル等の余白傾向の解析方法を紹介した。ライトノベルやネット小説は若者に人気の文字媒体であり、若者の読書量が減少していると言われている現在でも進んで読まれている。これらの文章について研究することで、若者の読書量を増加させることができるほか、教育機関では学生が読みやすい教材を作成できると期待している。今後は本論文の方法を用いて様々なライトノベルやネット小説の余白解析を行い、若者が読みたくなる文章構造の理解を進めたい。

参考文献

- [1] “第 54 回学生生活実体調査の概要報告”, 全国大学生活共同組合連合会, 2019 年.
<https://www.univcoop.or.jp/press/life/report.html> (参照 2019-01-06)
- [2] “2019 年 年間本ランキング”, oricon ME, 2019 年,
<https://www.oricon.co.jp/confidence/special/53961/> (参照 2019-01-06).
- [3] “Vol.1 個人発サイトがエンタメ／出版業界を席卷する理由”, KAI-YOU premium, 2019 年.
<https://premium.kai-you.net/article/53> (参照 2019-01-06)
- [4] “鮮血の剣姫”, 世捨て人, 2014.
<https://ncode.syosetu.com/n3268cg/>