

## 物語生成のためのショートショート作品の文体比較

English title

大塚 裕子<sup>\*1</sup>

Hiroko OTSUKA

藤倉 仁<sup>\*1</sup>

Hitoshi FUJIKURA

\*1 公立はこだて未来大学  
Future University Hakodate

The purpose of this study is to make of knowledge for short-short stories by automatic narrative generation. We think that the knowledge for short-short stories by automatic narrative generation is the distinctive words and text features, such as a ratio of a part of speech or a conversation, the number of a sentence, a conversation sentence. In this study, in order to clarify the knowledge that generates short-short story automatically, we analyzed the texts of the five typical writers of short-short story. As a result of analysis, as for the length of one sentence, it was revealed that there was not a difference by a writer. In addition, the number and the length of the conversation are understood a different thing by a writer.

## 1. はじめに

本研究は「きまぐれ人工知能プロジェクト『作家ですよ』」において、ショートショート作品の作家ごとの作風や文体の違いを、表現の種類や統計量、あるいはタグ付けした文間関係などによる計算可能な要因から求めることを目的としている。そのため、本稿では複数のショートショート作家の作品の基礎的な分析結果を報告する。

物語生成研究を大きく捉えた場合、人工知能分野では俳句や和歌、川柳の自動生成研究が行われている[土佐 2008]。近年、物語生成研究の次段階として、ある程度制限された文字数において物語が完結するショートショートの自動生成が注目されている[松原 2013]。作品数が多く、物語の特徴が明確であることから星新一の作品が研究対象として注目され、作品の物語構造に関する研究[村井 2013]や、特徴的な語彙に関する研究[村井 2011]も行われている。本研究では、星新一「らしさ」や、その他の作家「らしさ」に寄与する文体および表現の特徴を明らかにするため、作家ごとに文字長・会話長や、会話の割合、品詞割合等を分析する。この分析に先立ち、テキストの OCR による電子化や、段落・文・会話文のタグが付与されているタグ付きコーパス作成も行う。

## 2. ショートショートテキストコーパスの作成

文体比較に先立ち、分析の対象となる基本データを作成する。

## 2.1 対象としたショートショート作品

本研究では表1に示したとおり、星新一と他 4 作家のショートショート作品の文体特徴についての比較・考察を行う。対象とする作家は、阿刀田高、フレドリック・ブラウン、ロアルド・ダール、星新一、筒井康隆である。選択理由は、(1)星新一とならばショートショート作家として、阿刀田高、筒井康隆の名前が挙げられること、(2)着目したフレドリック・ブラウン作品は星新一による翻訳であること、(3)フレドリック・ブラウンとならば海外のショートショート作家としてロアルド・ダールの名前が見られること、による。分析データは、星ライブラリおよび新潮文庫から提供されている星新一の 433 作品のテキストデータから無作為に抽出した 20

テキストと、その他 4 作家の 86 作品を電子化した計 106 テキストを対象とする。

表1 対象としてショートショート作品テキスト

著者・訳者	書籍(文庫)名	対象数	出版年	出版社
阿刀田高	ナポレオン狂	13	2012	講談社文庫
フレドリック・ブラウン著 星新一訳	フレドリック・ブラウン傑作集	28	1982	サンリオ SF 文語
ロアルド・ダール著 田口俊樹訳	あなたに似た人	11	2013	ハヤカワ文庫
星新一		20*		
筒井康隆	笑うな	34	2012	新潮社文庫

\*433のテキストファイルを対象に無作為に抽出した。

## 2.2 BCCWJ コーパス基準に基づくタグ付け

本研究では文体特徴の分析のため、「現代日本語書き言葉均衡コーパス」(Balanced Corpus of Contemporary Written Japanese, 以後、「BCCWJ」と示す)における電子化フォーマット[山口 2010]のタグの認定基準に基づき、テキストデータへの段落・文・会話文のタグ付けを行う。電子化フォーマットの仕様としては、文章記述言語が XML、文字コードが JISX0213:2004 と指定されている。

BCCWJ の電子化フォーマットに示された各タグの意味を以下に示す。

[paragraph]: 文頭が半角空白によって一字下げが行われ、文末に改行が行われている場合

[sentence]: 文末が「.」「!」「?」、または( ) [ ] { } < > 《 》 「 」 『 』 【 】 のいずれかの括弧によって括られている場合

[quote]: カギ括弧(「」)で括られている場合

## 2.3 コーパスの作成および分析手順

コーパス作成は次の手順で行った。作家ごとに電子化テキストの総文字数を計算し表2の結果を得た。

表2 作家ごとの作品総文字数

	総文字数
阿刀田	200714
ブラウン	259171
ダール	158684
星	61160
筒井	105208

① OCR ソフトによるテキストの電子化

② コーパス作成

2-1 タグ付けプログラムの作成

2-2 タグ付きコーパスの作成

③ 分析・考察

3-1 タグ付きコーパスを対象とした

文長・会話長等の分析

表3 作家ごとの統計平均値と標準偏差

	平均総文字数	標準偏差	平均文数	標準偏差	平均文長	標準偏差	平均会話文数	標準偏差	平均会話文長	標準偏差
阿刀田高	15439.5	8181.5	221.8	78.1	32.0	18.9	93.9	40.9	23.9	27.6
フレドリック・ブラウン	17873.9	9726.0	235.2	255.1	26.8	14.4	60.4	80.6	48.0	38.9
ロアルド・ダール	14425.8	6926.2	243.1	75.3	28.8	18.3	123.4	63.1	28.9	26.2
星新一	3058.0	1372.9	66.4	35.8	26.1	11.5	27.3	16.4	45.5	34.2
筒井康隆	3094.4	2258.4	69.4	63.0	25.1	14.5	33.1	30.2	30.3	27.4

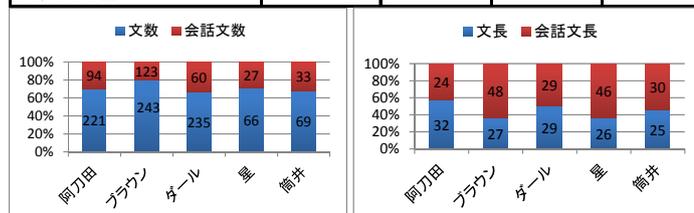


図1 文数・会話文数の割合 図2 文長・会話文長の割合

3-2 ①の電子テキストを対象としたテキストマイニングツールによる分析

阿刀田や筒井のテキストデータとしての網羅性は十分ではないが、一般的に言われているショートショート1作品文字数(約8000字以下)からすると、阿刀田は1作品が長め、星、筒井は短めの傾向が見られる。

### 3. 文体比較

文体比較は、文長・会話文長・会話率の比較(3.1)と、品詞に着目した比較(3.2)を行った。

#### 3.1 文長・会話文長・会話率の比較

表3は各作家の1作品における総文字数、平均文数、平均文長(1文の長さの平均)、平均会話文数、平均会話文長(1回のカギ括弧内の文字の長さ)の平均値と標準偏差である。また、図1は文数と会話文数の割合を、図2は文長と会話文長の割合を示したものである。これらの結果から以下のことが明らかになった。

1文の長さは、どの作家も25~30字と大きな差がないため、上述した作品の長さ(総文字数)の相違は文数から生じているといえる。1作品あたりの平均文数の偏りから、フレドリック・ブラウンの作品は作品によって長さ(総文字数)にばらつきがあるとみられる。

#### 3.2 品詞比較

テキストマイニングツールKHコーダーによる解析結果で見ると、品詞の表れ方については下記のとおり、全体に一定の傾向が見られる。

- 頻度が多いのは動詞、名詞、ひらがなの副詞(副詞B)
- 頻度が少ないのは人名、地名、組織名を含む固有名詞。

しかし、個々の品詞に着目すると、作家ごとの特徴が見られる。例えば、形容詞については「悪い」といった評価形容詞の場合、ロアルド・ダール、星新一、筒井康隆が頻度割合が1位と高く、阿刀田高は20位以降であるなど、表れ方が異なる。「白い」「赤い」「青い」「黒い」といった色彩形容詞、「大きい」「小さい」「高い」「早い」といった属性形容詞にも表れ方に相違がみられる。

形容詞や時間の表現の用法には作品の印象要因になるもの、作品の展開を読者に予測させるものがあるため[土田2003]、これらについても引き続き詳細に分析する予定である。

### 4. おわりに

ショートショート作家の作品ごとの文体や表現の特徴に関する基礎的分析を行った。今後は表現の対象や、表現間の関係など詳細についても分析することが課題である。

謝辞: 星新一氏の作品利用の快諾に関し、星ライブラリおよび新潮社に謝意を表す。

### 参考文献

- [松原 2013] □ 松原仁, 佐藤理史, 赤石美奈, 角薫, 迎山和司, 中島秀之, 瀬戸英明, 村井源, 大塚裕子, コンピュータに星新一のようなショートショートを創作させる試み, 2013.
- [村井 2013] □ 村井源, 物語プロットデータベースのためのデータ構造の検討, 情報知識学会誌, Vol.23, No.2, pp.308-315, 2013.
- [村井 2011] □ 村井源, 松本 斉子, 佐藤 知恵, 往住 彰文 物語構造の計量分析に向けて-星新一のショートショートの物語構造の特徴-情報知識学会誌, Vol. 21, No.1, 2011.
- [土田 2003] 土田知則, 青柳悦子, 伊藤直哉, 『現代文学理論』, 新曜社, 2003.
- [土佐 2008] 土佐尚子, 尾原秀登, 美濃導彦, 松岡正剛, コンピュータによる俳句創作支援システム, 映像情報メディア学会誌, Vol.62, No.2, pp.247-255, 2008.
- [山口 2010] 山口昌也, 高田智和, 北村雅則, 間渕洋子, 大島一, 小林正行, 西部みちる, 『現代 □ 日本語書き言葉均衡コーパス』における電子化フォーマット □ ver.2.2, □ 特定領域研究「日本語コーパス」 □ 平成 22 年度研究成果報告書, 2010.