Selecting Words and Notation Using Literary Data in the Integrated Narrative Generation System

Jumpei Ono[†]
Graduate School of Software and Information Science, Iwate Prefectural University
152-52 Sugo, Takizawa, Iwate, Japan

Takashi Ogata[‡]
Faculty of Software and Information Science, Iwate Prefectural University
152-52 Sugo, Takizawa, Iwate, Japan
E-mail: [†]g236m001@s.iwate-pu.ac.jp, [‡]t-ogata@iwate-pu.ac.jp

Abstract

This paper presents a way for selecting noun concepts based on the analysis of appearance frequency of noun words in the data of modern Japanese novels ("Aozora Bunko"). The proposed method is incorporated into some mechanisms in our integrated narrative generation system (INGS). We also show an overview of INGS, in particular the mechanism relating the proposed method in this paper. Additionally, we introduce a preliminary experiment to confirm and discuss the effectiveness.

Keywords: Integrated Narrative Generation System, Word Frequency, Conceptual Dictionary, Noun Concepts, Word Notation, Aozora Bunko.

1. Introduction

Language generation controlled by a story or as a narrative is an important communication ability in robotics and artificial life. We have been developing a narrative generation system called "Integrated Narrative Generation System: INGS" [1,2]. In intelligent informatics such as natural language processing and artificial intelligence, narrative generation or story generation is an important and challenging topic in the following scientific and applicable points: an interdisciplinary approach of informatics and literary theories including narratology, generating consistent discourse structures, creating advanced contents technologies, etc. INGS is a synthesis of our various previous researches of narrative generation.

INGS generates narrative events and the sentences using the generation techniques and knowledge elements including conceptual dictionaries [3] and language notation dictionary [4]. Though we will give the mechanism in the next section, overview, INGS selects the noun concepts (terminal nodes in the hierarchical structure) in an event to be generated in the limited range (the lower range of one or more intermediate noun concepts) in the noun conceptual dictionary. But, the selection is randomly processed. As the lower range of an intermediate noun concept includes various types of noun concepts, for example difficult/easy, new/old, etc., various types noun concepts are mixed in the generated events in many cases. In the mechanism of INGS, noun concepts are basically defined by the level of words and noun concepts are transformed to words with each actual letter representation (notation) using a language notation dictionary. So selecting the concepts directly has an influence on the quality of generated sentences. Our

Published by Atlantis Press

goal to solve it is implementing more strategic mechanisms for concepts (and notations) selection and this paper will present a simple method based on the frequency analysis of words' appearance as the first step.

In particular, we automatically analyze frequency information of noun words from novels stored in "Aozora Bunko", which is a digital library that includes a variety of texts of modern Japanese novels mainly, to select noun concepts in events to be generated according to the acquired frequency information. For example, if we use noun concepts according to high-frequency noun words, the output text will be more readable.

There are studies of word familiarity relating to the theme of this paper. [5] considered the readability of a sentence based on the relationship between word familiarity and word frequency and sorted sentences based on readability according to the idea that high-frequency words are high-familiarity words.

In the large framework, our goal is not necessarily only readable story generation. Readable story generation based on high appearance frequency, namely concepts and words selection easy to read, is one of the strategic choices. In contrast, we will be able to make more difficult or strange sentences intentionally using low-frequency words and concepts. Additionally, for the processing relating to the final words notation in INGS, for instance, we will be able to concentrate on analyzing specific authors to simulate or imitate words representation in each style of authors. Though it is not the theme of this paper, we will mention the attempt.

2. An Overview of INGS and Noun Concept Selection Mechanism

This section will give an overview of INGS architecture to show the noun concepts selection mechanism. The architecture of INGS has two types of macro level parts (Fig. 1.): "generation mechanism" and "knowledge mechanism". The former part has three main elements of "story generation mechanism", "discourse mechanism", and "surface generation mechanism (including language, music, and image)". The main elements of the latter "knowledge mechanism" are "conceptual dictionaries (for noun concepts and verb concepts chiefly)", "language notation (letter notation) dictionary", "state-event transformation knowledge base", and so on.

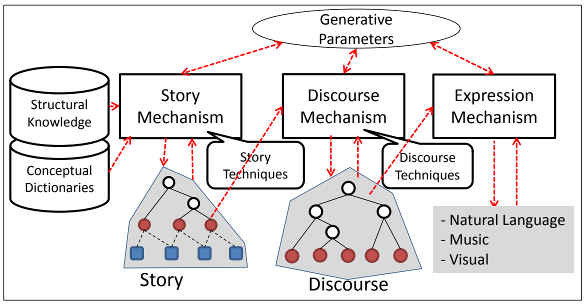


Fig. 1. The architecture of INGS

2.1. *Generation Mechanisms*

In story generation, "story" means the content of a narrative or a temporal sequence of events. In narrative generation process, a story is described as a tree structure in which the basic elements are an event that is described by conceptual representation and a relation for combining among events or sub-structures. Additionally, an event is principally combined with a preceding state and a subsequent state. The major functions are holding the knowledge about a story world and managing the coherency of the flow of events. A story structure including events and relations is generated using one or more story techniques and states associating with the events are made according to another mechanism.

In the next discourse mechanism, "discourse" means how to form a narrative, which is a sequence or a structure of events to be narrated actually. A discourse is corresponded to the structure as the transformation of a story's structure and various discourse techniques process the structural transformation. We use the narrative discourse theory by Genette [6] for the categorization of discourse techniques. In the expression mechanism, "expression" means the part for representing the above conceptual representations with surface media including mainly natural language, such image or visual media as animated movie, and music.

The first two phases are conceptual generation parts in a process. Both of story and discourse structures are commonly represented by each structure to be operated respectively using various types of techniques. As stated above, we call the structural operation techniques "story techniques" and "discourse techniques".

2.2. *Conceptual Dictionaries*

An event is constructed with the instantiation of conceptual materials stored in conceptual dictionaries for noun concepts and verb ones. Each dictionary has a hierarchical structure from higher concepts to lower ones. The noun dictionary currently contains 115765 terminal concepts and 5809 intermediate ones. Each intermediate concept has (1) a list of hyponymy concepts, (2) the number of depth in the hierarchy, (3) the serial number of the super-ordinate concepts, and (4) the range of serial numbers of the hyponymy concepts (Fig. 2.). Fig. 3 shows the description of a noun concept.

```
(<intermediate concept>
   (hierarchy
    (depth <number>)
    (hype <intermediate concept>)
    (hypo <intermediate concept>)
    (terminal <terminal concept>)))
```

Fig. 2. Description form of an intermediate concept

```
(男[Man]
   (hierarchy
    (depth 8)
    (hype 人間〈男女〉[Human<Gender>])
    (hypo 男[接辞] [Man[Affix]])
    (terminal 男[Man] 父[Father] 少年[Boy]…)))
```

Fig. 3. The description of the noun concepts, “Man”

On the other hand, the verb concepts hierarchy has 11951 terminal concepts and 36 intermediate categories. Each terminal verb concept describes the following three elements: (1) a “sentence-pattern” to show a pattern for a sentence including the verb, (2) one or more “case-frame(s)” to show the types of required noun cases, and (3) one or more “constraint(s)” to define the range in the noun conceptual dictionary in which each noun concept in the above case-frame(s) requires. For example, Fig. 4. is the description of the verb concept, “eat”. When INGS materializes the framework of a case-frame, it selects concepts in the noun conceptual dictionary. The objective of the paper is revising the selecting method.

```
((name 食べる2 [eat])
 (sentence-pattern “N1が N2を食べる”[N1 eat N2])
 (case-cons-set
  ((case-frame ((agent N1) (object N2) …))
   (constraint
N1{     ((人[human] –死人[corpse]
        –人間〈人称〉[human<person>]–準人間[semi-human])
N2{     (食料[food] –調味料[flavoring]
        –飲物[drink]・たばこ[cigarette]))) …) …)
```

Fig. 4. The description of a verb concept

2.3. *Concept Selection in Story Generation*

First, we show a story structure to be generated in the story generation mechanism in INGS to explain a story generation process (Fig. 5.). As indicated in the above figure, the hierarchical structure of a story is constructed with three elements: (1) an “event” consists of a verb concept and some noun concepts (as stated above), (2) a “state” is the attribute information to position before and behind of the events (an event changes a state to the next state), and (3) a “relation” connects the lower structure semantically.

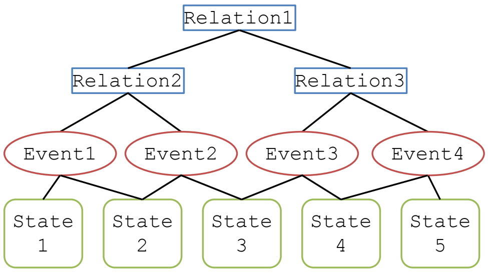


Fig. 5. The form of a story structure

A story generation process is equal to expanding or transforming a story structure. In particular, under some parameters (“macro-structure”, “length”, etc.) given in the first step, a story technique makes a new event or a sub-structure including one or more new events using a variety of story contents knowledge to integrate them into the original structure using various relations. According to a generated new event, new states are also generated. When an event is generated, the above story contents knowledge gives the basic form based on the description of a case-frame in Fig. 4. Each constraint indicates the range in the noun conceptual dictionary to decide a noun concept. Actually, each noun concept in an event is transformed into an instance using attribute information if necessary. Finally, each generated event is transformed into a natural language sentence using specific word representation. Fig. 6 shows the process.

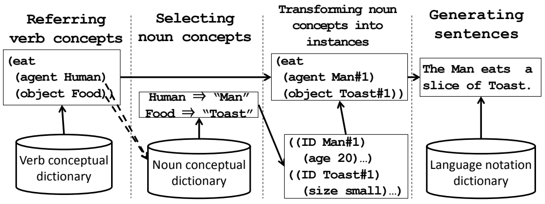


Fig. 6. The process of generating an event

As mentioned above, in applying constraints to choose noun concepts, terminal concepts under an intermediate node in the noun conceptual dictionary are not fully organized. In the following sections, we will present a method using words' appearance frequency instead of revising the organization or structure.

3. Concept Selection Based on Word Frequency

We will describe the method to analyze word frequency and show some generation examples using the result.

3.1. *The Method of Word Frequency Analysis*

We have analyzed word frequency in 4980 texts (mainly novels) in Aozora Bunko from 1872 through 1963 to use it for noun concept selection in story generation. The image of the processing is shown in Fig. 7. KH Coder [7] is used in the analysis in the first step to account the word frequency. And the mechanism links acquired frequency information to the corresponded noun concepts in the noun conceptual dictionary.

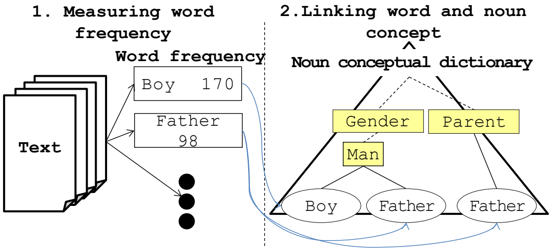

1. Measuring word frequency
2.Linking word and noun concept
Word frequency
Noun conceptual dictionary
Text
Boy 170
Father 98
Gender
Parent
Man
Boy
Father
Father

Fig. 7. An image to use frequency

We describe the method in detail. Firstly, in the texts (corpus) to be used, we have accounted the word frequency for "general nouns", "nominalized adjectives", and "nouns as stems of adjectival verbs".

On the other hand, in the noun conceptual dictionary to give frequency, for such 3549 noun concepts that become a concept by the combination with another concepts as "～学校[～school]", we do not give frequency. Additionally, terminal noun concepts that have a same name sometimes exist under different intermediate concepts. For example, two intermediate concepts, "君主[王](monarch)" and "貴人[君主](nobleman)", respectively include a terminal concept, "王(king)". The number of this type of terminal concepts was 28421. When such overlapped terminal noun concepts appear in analyzing a text, the mechanism adds frequency 1 to all the concepts. We count the overlapped noun concepts as one concept, so the above 28421 are equal to the number of sets of overlapped concepts and 36187 concepts that the 28421 are eliminated from all the overlapped concepts are not included in the range of the noun conceptual dictionary to add frequency. Considering the above conditions, the number of terminal noun concepts to which appearance frequency should be added is 76029 (for 115765 terminal concepts in the noun conceptual dictionary).

According to the conditions, the mechanism analyzes the word appearance frequency in the target texts by morphological analysis. As stated above, the mechanism accounts the frequency of "general nouns", "nominalized adjectives", and "nouns as stems of adjectival verbs". Next, the proposed mechanism relates the above nouns with frequency to terminal noun concepts in the noun conceptual dictionary by simple matching. Though each description of noun concepts is theoretically a kind of label and the linguistic notation is proposed by the language notation dictionary, a general notation is actually used for each of the description.

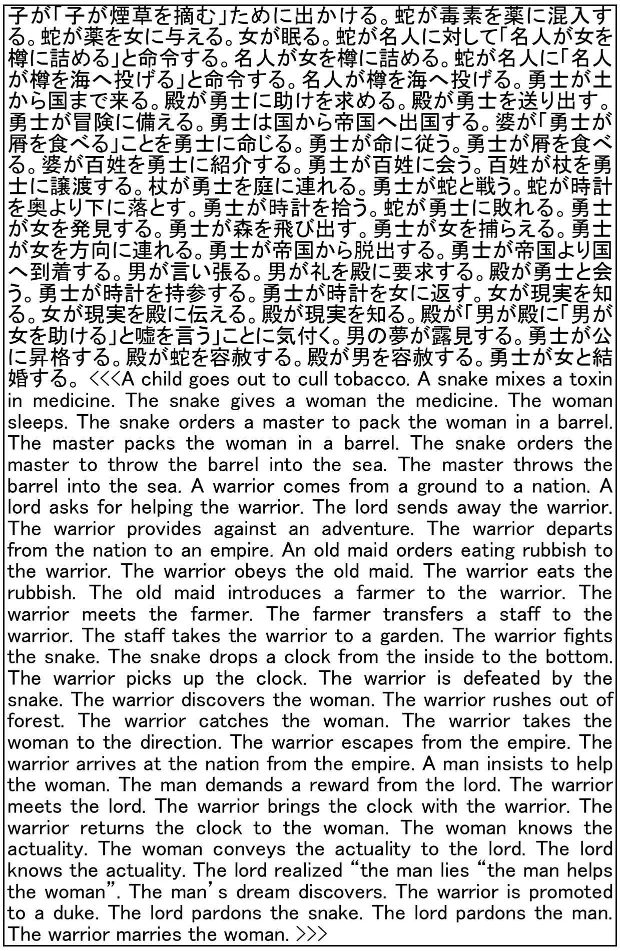
As a result, we have acquired word frequency for 57131 nouns in the target texts and linked the frequency to 44332 terminal noun concepts. 44332 concepts are equal to 58% for all terminal noun concepts in the noun conceptual dictionary in INGS.

3.2. *Concept Selection Based on Frequency*

Fig. 8. is an example of story generation using noun concepts with the highest frequency. Actually, it is the result transformed by "sentence generation mechanism" in INGS. An event in a story is simply transformed into a sentence based on the sentence pattern. Additionally, the future version of INGS is necessary to change some noun representations in each sentence to another nouns using attributes for each noun concept. For instance, "child" will be changed to an actual name such as "Taro". The mechanism will enable to represent a noun concept by a noun phrase, such as "Taro in a forest".

4. Preliminary Evaluation and Discussion

We conducted a survey to find out the effectiveness of selecting noun concepts in stories generated based on the acquired frequency. The subjects were 4 students.



子が「子が煙草を摘む」ために出かける。蛇が毒素を薬に混入する。蛇が薬を女に与える。女が眠る。蛇が名人に対して「名人が女を樽に詰める」と命令する。名人が女を樽に詰める。蛇が名人に「名人が樽を海へ投げる」と命令する。名人が樽を海へ投げる。勇士が土から国まで来る。殿が勇士に助けを求める。殿が勇士を送り出す。勇士が冒険に備える。勇士は国から帝国へ出国する。婆が「勇士が屑を食べる」ことを勇士に命じる。勇士が命に従う。勇士が屑を食べる。婆が百姓を勇士に紹介する。勇士が百姓に会う。百姓が杖を勇士に譲渡する。杖が勇士を庭に連れる。勇士が蛇と戦う。蛇が時計を奥より下に落とす。勇士が時計を拾う。蛇が勇士に敗れる。勇士が女を発見する。勇士が森を飛び出す。勇士が女を捕らえる。勇士が女を方向に連れる。勇士が帝国から脱出する。勇士が帝国より国へ到着する。男が言い張る。男が礼を殿に要求する。殿が勇士と会う。勇士が時計を持参する。勇士が時計を女に返す。女が現実を知る。女が現実を殿に伝える。殿が現実を知る。殿が「男が殿に「男が女を助ける」と嘘を言う」ことに気付く。男の夢が露見する。勇士が公に昇格する。殿が蛇を容赦する。殿が男を容赦する。勇士が女と結婚する。<<<A child goes out to cull tobacco. A snake mixes a toxin in medicine. The snake gives a woman the medicine. The woman sleeps. The snake orders a master to pack the woman in a barrel. The master packs the woman in a barrel. The snake orders the master to throw the barrel into the sea. The master throws the barrel into the sea. A warrior comes from a ground to a nation. A lord asks for helping the warrior. The lord sends away the warrior. The warrior provides against an adventure. The warrior departs from the nation to an empire. An old maid orders eating rubbish to the warrior. The warrior obeys the old maid. The warrior eats the rubbish. The old maid introduces a farmer to the warrior. The warrior meets the farmer. The farmer transfers a staff to the warrior. The staff takes the warrior to a garden. The warrior fights the snake. The snake drops a clock from the inside to the bottom. The warrior picks up the clock. The warrior is defeated by the snake. The warrior discovers the woman. The warrior rushes out of forest. The warrior catches the woman. The warrior takes the woman to the direction. The warrior escapes from the empire. The warrior arrives at the nation from the empire. A man insists to help the woman. The man demands a reward from the lord. The warrior meets the lord. The warrior brings the clock with the warrior. The warrior returns the clock to the woman. The woman knows the actuality. The woman conveys the actuality to the lord. The lord knows the actuality. The lord realized “the man lies “the man helps the woman”. The man’s dream discovers. The warrior is promoted to a duke. The lord pardons the snake. The lord pardons the man. The warrior marries the woman. >>>

Fig. 8. The result of a story generation based on the frequency (Japanese: the original sentences, English: the translation)

For the survey, we prepared five stories including the above example based on the following frequency: [Random] when the mechanism selects a noun concept in an event, it randomly selects a concept from the range of intermediate concepts, [Max] the mechanism selects a concept with the highest frequency, [Middle] it selects a concept with the middle frequency, [Min] it selects a concept with the lowest frequency, [Zero] it selects a concept with no description of frequency. In all cases, if there are two or more candidates, the mechanism randomly selects one concept.

The procedure of the experiment is as follows. The subjects read each story and select one of the evaluation values for pointed nouns: (1) no strange, (2) a few strange, and (3) very strange. Additionally, we set a time limit (5 minutes) for a story. And we prescribed to not change the already described values and not necessary to confirm previous values.

Table 1. shows the result. The values indicates that the nearer 1, the smaller the sense of strange. This result was same with previous survey [8] which had been conducted for 8 subjects in the halfway stage of the analysis of Aozora Bunko. An interesting point was as follows. As we did the experiments using generated sentences directly, the subjects were sometimes had an effect on contextual information within a sentence or between sentences. It indicates that contextual conditions were also considered in selecting concepts. Considering the problem will be a future issue. On the other hand, in the “average” in this experiment, differences between [Max] and the others were relatively small. This indicates that controlling extremely unreadable word representation is difficult in this way. It will be also an important future topic.

Table 1. The evaluation result

Subject	Random	Max	Middle	Min	Zero
A	1.79	1.04	1.48	1.63	1.85
B	1.40	1.00	1.46	1.45	1.23
C	2.39	1.69	2.25	2.77	2.79
D	1.06	1.06	1.11	1.26	1.04
Average	1.66	1.20	1.57	1.78	1.73

The result of the frequency analysis, there were 12799 words which did not exist in the noun conceptual dictionary. Table 2. shows the two types. “No existing words” (7928) will be able to add to the noun conceptual dictionary. This category includes many complex words. And, in many cases, as the parts are stored in the current version of the noun conceptual dictionary, we would like to address the way for using the knowledge to add the concepts.

In the type of “the problem of notation”, a similar noun concept with a word exists in many cases. The number of such words was 4871. In Table 3., we show concepts to be corresponded to such words. We will be able to connect such words to elements in the language notation dictionary as mentioned above. The language notation dictionary is a writing system which stores letter representations for all concepts using Kanji, Hiragana, Katakana, etc. The extracted words from Aozora Bunko which does not match with concepts directly sometimes are found out in the language notation dictionary. We will be able to use the mechanism for such words and concepts.

Table 2. Types of words which did not exist

Types		Example
The problem of notation		ネジ[Screw]
		ボランティア[Volunteer]
		食慾[Appetite]
No existing words	Complex words	駅前通り[Street in front of station]
	Others	カフェ[Cafe]

Table 3. Extracted words and the similar concepts

Words	Corresponding concepts
ネジ[Screw]	螺子[Screw]
ボランティア[Volunteer]	ボランテア[Volunteer]
食慾[Appetite]	食欲[Appetite]

5. Connecting to Word Representation

As described above, we have been developing a language notation dictionary to give some notation or word (letter) representation for each noun concept (and verb concept). In the part of sentence generation in INGS in which transforms concepts to actual word representation, we have been trying to simulate the characteristics in various authors' novels [9]. In particular, for a novel or a part of a novel, we statistically analyzed the percentage of Kanji, Hiragana, Katakana, etc. for main word classes. In Fig. 9. and Fig. 10., we apply the above method to sentences generated by the proposed method in this paper (Fig. 8.).

子が「子が煙草を摘む」ために出かける。蛇が毒素を薬に混入する。蛇が薬を女に与える。女が眠る。蛇が名人に「名人が女を樽へ詰める」と命令する。名人が女を樽に詰める。蛇が名人に対して「名人が樽を海に投げる」と命令する。名人が樽を海へ投げる。勇士が土から邦に来る。<The rest is omitted>

Fig. 9. Sentences represented according to the notation analysis of 「蜘蛛の糸 (The Spider's Thread)」

子が「子が煙草を摘む」ことにでかける。蛇が毒素を薬にこんにゅうする。蛇から薬を女へあたえる。女がねぶる。蛇が名人に対して「名人が女を樽へ詰める」こととめいれいする。名人が女を樽へつめる。蛇が名人に「名人が樽を海へ投げる」こととめいれいする。名人が樽を海へなげる。<omission> 勇士ガトケイをヒラウ。ヘビガ勇士にヤブレル。ユウシが女ヲハッケンスル。勇士ガモリをトビダス。ユウシが女ヲトラエル。勇士が女ヲホウコウヘツラネル。ユウシガ帝国よりダッシュツスル。<The rest is omitted>

Fig. 10. Sentences represented according to the notation analysis of 「みずから我が涙をぬぐいたまう日 (The Day He Himself Shall Wipe My Tears Away)」

6. Conclusion

In this paper, we have proposed a way for selecting noun concepts based on the frequency analysis of the data of Japanese modern novels. We have incorporated the proposed method into our integrated narrative generation system (INGS), in particular the story generation sub-system and the noun conceptual dictionary. And, we have confirmed the effectiveness through the preliminary experiments. Our future plan includes the following immediate and technological topics: the processing of complex words, expanding the noun conceptual dictionary using the result of the analysis, increasing the matching using the language notation dictionary. Additionally, the strategic control of the readability (including un-readability) of texts to be generated is more theoretical issues to be addressed.

References

[1] T. Ogata and A. Kanai, *An Introduction to Informatics of Narratology: on thought and technology of narrative generation* (Gakubunsha, Tokyo, 2010).

[2] T. Akimoto and T. Ogata, An Information Design of Narratology: The Use of Three Literary Theories in a Narrative Generation System, *The International Journal of Visual Design* 7(3) (2014) 31-61.

[3] K. Oishi, Y. Kurisawa, M. Kamada, I. Fukuda, T. Akimoto and T. Ogata, Building Conceptual Dictionary for Providing Common Knowledge in the Integrated Narrative Generation System, *Proc. of the 34th Annual Conference of the Cognitive Science Society* (2012) 2126-2131.

[4] S. Kumagai, S. Funakoshi, T. Akimoto and T. Ogata, Development of a Language Dictionary and a Simple Narrative Sentence Generation Mechanism, *Proc. of the 26th Annual Conference of the Japanese Society for Artificial Intelligence* (2012) 1N1-OS-1a-3.

[5] K. Tanaka-Ishii, S. Tezuka and H. Terada, Sorting Texts by Readability, *Computational Linguistics* 36(2) (2010) 203-227.

[6] G. Genette, J. E. Lewin (trans.), *Narrative Discourse: An Essay in Method* (Cornell University Press, NY, 1980). (G. Genette, *Discours du récit, essai de méthode, Figures III* (Seuil, Paris, 1972).)

[7] K. Higuchi, Quantitative Analysis of Textual Data: Differentiation and Coordination of Two Approaches, *Sociological Theory and Methods* 19(1) (2004) 101-115.

[8] J. Ono and T. Ogata, Selecting Noun Concepts Based on Accounting Data: As a Mechanism in the Integrated Narrative Generation System, *IEICE Technical Report* 114(366) (2014) 49-54.

[9] M. Kamada and T. Ogata, Analysis and Imitation of the Japanese Letter Notation in Narrative Texts, *Proc. of the 27th Annual Conference of the Japanese Society for Artificial Intelligence* (2013) 2I4-5in.