

統合物語生成システムにおける概念選択／語彙表記選択及びその制御 Concept Selection/Language Representation Selection and the Control in an Integrated Narrative Generation System

小野 淳平*1
Jumpei Ono

小方 孝*2
Takashi Ogata

*1 岩手県立大学大学院
Graduate School of Iwate Prefectural University

*2 岩手県立大学
Iwate Prefectural University

In our Integrated Narrative Generation System (INGS), in generating of the conceptual structures (stories and discourses), a mechanism selects each noun concepts in an event according to the constraint in the noun conceptual dictionary and another mechanism gives a notation to each of the selected noun concepts. This paper presents a method to set frequency information to each noun concepts in the noun conceptual dictionary using co-occurrence relations among words in addition to the frequency of words. And to narrative conceptual structures generated using the above method, we apply a revised technique for deciding language notation.

1. はじめに

本稿は、開発中の統合物語生成システム(Integrated Narrative Generation System: INGS)[小方 2010; Akimoto 2014]における、ストーリー生成処理で生成される事象を構成する名詞概念の選択及びその言語表記に関連する。INGS は物語の概念構造の拡張・変形によりストーリー生成を行うが、この概念構造は、人・物・場所の静的情報を示す状態、ある状態を別の状態に推移させる動的情報を示す事象、事象どうしを意味的に結び付ける関係を含む木構造を成す。新しい事象はストーリー技法及び具体的なコンテンツ知識を用いて生成される。個々の事象は一つの動詞概念と複数の名詞概念(人・物・場所等)のフレーム(格構造)として記述される。動詞的概念はストーリー技法において指定される一方、名詞的概念は名詞概念辞書[Ogata 2015]中から以下のようにして選択される。動詞概念辞書[Ogata 2015]中の個々の動詞概念の定義中にその格構造が含まれそれを構成する名詞概念の制約条件が規定される。この制約条件により上記名詞概念辞書における中間概念が示され、具体的な名詞概念がその中に含まれる名詞概念候補の中からランダムに一つ選択される(引き続きそのインスタンスが生成され、それが事象の値となる)。現状ではある中間概念に含まれる名詞概念は均質ではないため、無作為な選択によって、様々なタイプの名詞概念が選択されてしまうという問題があった。

解決の一方方法は名詞概念辞書の構成自体の見直しであるが、それ以前の方策として[小野 2014]は名詞概念辞書中の各名詞概念の選択に頻度情報を利用する方法を提案した。名詞概念辞書における最下層の終端名詞概念は名詞の語に対応する。そこで小説を中心とする文学作品の文章データ[『青空文庫』]から名詞の語の出現頻度を計算し、これを名詞概念辞書中の名詞概念の頻度情報として概念選択に用いた。頻度情報が上位の概念選択を行った生成文が、ランダム選択による生成文よりも読みやすさが向上するという結果を得た。

しかし小説データの不足から頻度情報が獲得できない語が多数存在した。データの増量という方向はあるが、ここでは頻度情報 0 のデータからでも頻度情報を近似的に獲得する方法として語の共起情報を用いる方法を提案する。ある作家の使用語彙の模倣の試みにおいてデータ不足から頻度情報を獲得できない語が多数存在する場合でも対処が可能となる。INGS では選択された概念に対して表層的な語彙表記を言語表記辞書[鎌田 2013]を用いる。本稿では、上記方法による概念選択結果を実際の表記に変換する部分も含め結果を示すため、言語表記の従来の方法における問題点とその改訂も触れる。

連絡先: 小野淳平, 大学院ソフトウェア情報学研究所, 岩手県滝沢市菓子 152-52, g236m001@s.iwate-pu.ac.jp

まず前提となる事項を概説する。まず名詞概念の選択方法について述べる。名詞概念辞書の構造を図 1 に示す。概念の意味分類に相当する中間概念と個々の具体的な要素を意味する終端概念から構成される。さらに中間概念及び図 2 に示すような終端概念により構成される動詞概念辞書を持つ。個々の動詞概念には、格構造(図 2 の”case-frame”内の情報)と格に挿入される概念の制約条件(図 2 の”constraint”内の情報)が記述される。

動詞概念が持つ格に対する制約条件は名詞概念辞書の中間概念で記述され、事象生成における名詞概念は、制約条件に記述された中間概念に含まれる終端概念の集合からランダムに選択される。制約条件には複数の中間概念が記述されている場合が多く、名詞概念選択において、一つの間概念をランダムに選びその名詞概念を選択対象とする方法と、すべての中間概念の下位の名詞概念を選択対象とする方法がある。

この方法に対して[小野 2014]は、『青空文庫』の収録テキストからすべての語の出現頻度を計算し、ある語と表記が一致する名詞概念に対してその出現頻度をその名詞概念における出現頻度とした。『青空文庫』中に現れるが対応する名詞概念が存在しない語が総計 7997 個あった。これを名詞概念辞書に追加するかどうか検討中である。名詞概念に付属した頻度を用いた生成実験では、5 つの基準【最多, 中間, 最少, 0, ランダム(旧来の方法)】を用意し、【最多】頻度に基づく生成により最も違和感の少ない生成ができることを検証した。さらにより細かく名詞概念の選択制御が可能となるように、概念の選択候補を 1 ~ 100 の値に当て嵌め数値の指定により選択可能にしている(図 3)。

以上のようにして生成される、事象を最も重要な基本要素の一つとする構造を物語の概念構造と呼ぶ。ストーリー生成機構及び物語言語機構で生成乃至処理される構造である。それが表層表現機構の入力となり言語や音楽や映像の生成に架橋される。本稿に關係するのは言語生成の部分、特に概念構造中の事象における個々の名詞概念に特定の表記を付与する部分である。なお、概念構造中の事象における名詞概念や動詞概念の記述は概念辞書における記述に従う。その記述は通常の語彙の記述によっているが、しかしそれはあくま

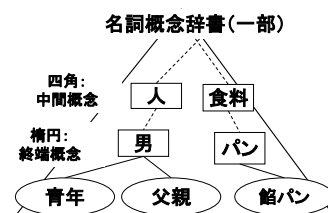


図 1 名詞概念辞書の構造

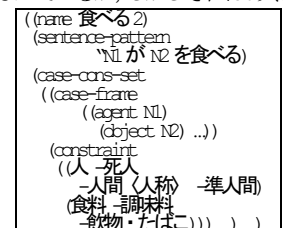


図 2 動詞概念の記述例

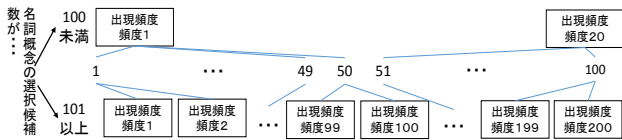


図3 名詞概念の出現頻度を百分割する

で概念のラベルであり、実際の表記は文生成機構のタスクとなる。但し、概念辞書におけるこのラベルとしての表記を、その概念が単語化された際の代表的(規範的)な表記として扱っている。

言語表記処理の理解のためにここでまず、文生成における言語表記の選択に関して述べる。図4に INGS における文の形式的構造のモデルを示す。文は一つ以上の単語で構成され、単語は一つ以上の文字で構成される。ある単語は特定の品詞属性を持ち、ある文字は特定の表記(文字表記もしくは文字種類と呼ぶ)と結び付く。現状の INGS は、品詞として名詞・動詞・助詞を扱い、文字表記として漢字・ひらがな・カタカナ・ローマ字を用いる。これらに加えて、生成される文では、鍵括弧(「」,「」), 句点(。), 読点(、)も使用される。単語を構成する個々の文字ごとに文字表記を付与する方式も考えられるが、現状では単語単位で文字表記を割り当てている。

以上のように、INGS は生成された物語の概念構造から言語表記辞書を用いて文表現を生成する[鎌田 2013]。言語表記辞書は、ある概念を表層的に表現するための複数の表記の可能性を記述・格納する辞書である(図5)。現状、漢字・ひらがな・カタカナ・ローマ字の四種類の表記を用意する。送り仮名を含む漢字の表記以外は同一の文字種のみで構成される。

[鎌田 2013]による機構では、ユーザが、名詞・動詞・助詞ごとに、使用するべき文字種類(漢字・ひらがな・カタカナ・ローマ字)の割合(0%~100%)を指定する。その値は品詞ごとに合計が 100 となる必要がある。これに基づき表記選択を次の方法で行う—①生成された物語の概念構造(ストーリーもしくは言説)を文生成機構が基本文列に変換したもの(基本文における動詞や名詞はこの段階ではそれぞれの概念辞書の記述で表記されている)を対象として、その中に含まれる名詞・動詞・助詞それぞれの単語の総数を数える、②品詞ごとに指定された文字種類の割合を単語の総数に掛け、各々の文字種類で表記すべき単語数を決める、③この値に沿って各品詞の単語の言語表記をランダムに決定する。例えば名詞が 10 個現れる物語で、名詞の文字種類割合が漢字 60%、ひらがな 40%と指定された場合、6 つの名詞が漢字、4 つの名詞がひらがなで表記される。但し計算の結果品詞の数が小数になる場合、端数は四捨五入する。その結果各品詞における文字種類ごとの単語の数の合計が、品詞ごとの単語の総数を超える場合(例:10 個の名詞のうち 5.5 個の名詞を漢字、4.5 個の名詞をひらがなで表記する場合、どちらも繰り上がり 11 個の名詞を対象とすることになる)、数が少ない文字種類から端数切り捨てで調整する。

2. 出現頻度 0 の概念に対する共起情報の利用

前節の頻度情報に基づく名詞概念選択の方法はすべての概念の出現頻度が用意されていることを前提とする。ある名詞概念の候補すべてが頻度 0 であるような極端な場合、頻度に基づく概念選択は機能せず従来のランダム選択と同じになる。これを回避する一方法は参照データの量を増やすことであるが、ここでは概念どうしの共起情報を利用する方法を提案する。

2.1 出現頻度計算における共起情報の利用方法

出現頻度 0 の名詞概念に関して、「名詞概念 A と共起関係を持つ概念 B の出現頻度と概念 A の出現頻度は比例する」という仮定に基づき、その名詞概念の共起情報を求め、その名詞概念における頻度情報を利用して頻度 0 の名詞概念に頻度情報を設定する方法を提案する。手続きは次の通りである—①KH Coder[樋口 2004]を利用して名詞概念 A と共起関係を持つ名詞概念を共起関係の強さの順に最大で 5 つ抽出する、②抽出した各名詞概念の出現頻度の平均値を計算し、これを名詞概念 A の出現頻度とする。(頻度 0 の概念の数

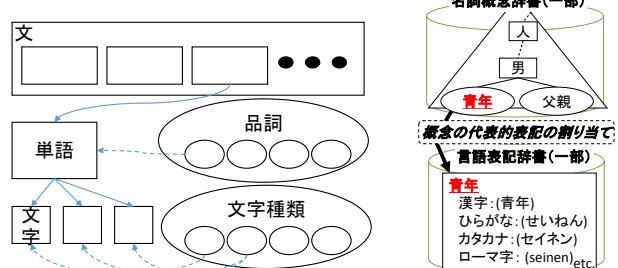


図4 文の形式的構造に関するモデル 図5 名詞概念と言語表記の関係(①-②を繰り返す。)なお KH Coder では、五種類の共起情報計算式を提供しており、どれを用いるかによって計算結果は異なる。

2.2 出現頻度が存在する名詞概念への適用

以下の 2 節でこの方法を実験する。2.2 節では、出現頻度が存在する名詞概念群に対して適用し実際のデータと比較する実験を行い、続く 2.3 節で出現頻度 0 の概念に対する適用を行う。なお、その結果に基づく実際の生成実験については、言語表記の処理の説明(3 節)に引き続き、4 節で述べる。

対象として、敢えて頻度 0 の名詞概念が多くなるように『青空文庫』から横光利一の小説 30 作品を選んだ。これらから出現頻度を得た名詞に対応する名詞概念辞書における終端概念の数は 16695 であり、全終端概念 115765 のうち残り 99070 が頻度 0 となる。中間概念のレベルでは 1193 の中間概念中のすべての名詞概念が 0 である。共起情報の計算対象の方は『青空文庫』に収録された 4960 作品とする。

まず提案方法の妥当性の程度の検証のために、出現頻度が存在する名詞概念群に対しこれを適用して実際のデータと比較する。その際 KH Coder が提供する五種類の共起情報計算式(表 1)も比較する。計算式ごとに、既に正確な出現頻度を持つすべての名詞概念について共起情報を使った提案手法により推測の頻度情報を求めた。元の正確な頻度と提案手法による計算結果の差の平均(μ)、標準偏差(σ)及び元の頻度と計算結果の差(16695 個)が平均一標準偏差($\mu - \sigma$)以上かつ平均+標準偏差($\mu + \sigma$)以下の範囲に含まれている割合を表 1 に示す。この表より確率比が、他との比較において、差の平均が小さく、標準偏差の値も小さいため、次の実験ではこれを使用する。

KH Coder における確率比の計算式を図 6 に示す。Sa は検索対象 A が含まれる文の総数、Sb は比較対象 B が含まれる文の総数、Sa ∩ Sb は A と B の両方を含む文の総数、Sall はサンプルのテキストに含まれる文の総数である。Sa, Sb は不定であるが、例えば「亜鉛」を検索対象とし比較対象を「軟膏」とした場合、Sa は 47、Sb は 2 となる。Sall は 4960 作品に含まれる文の数(1526940)である。

2.3 出現頻度が 0 の概念への頻度情報設定

上のデータを使って頻度 0 の名詞概念の頻度設定を確率比を用いて行った結果、頻度情報が設定された終端概念は 104808 となり、処理前より 88113 増加した。図 7 に示すように全般的な変化が見られた。表 2 は実際の結果の例である。「亜鉛」の場合、共起関係を持つ概念には「国家 8」「鴉 3」「海港 28」「肋材 0」「陽極 0」等(数字は頻度)があり、共起関係の強い 5 つの概念の平均から頻度情報を 7.8 と設定した。全体として、設定された頻度情報が 1 未満の概念は合計 13070 あり、そのうち 1759 がカタカナ語(名詞概念辞書における記述がすべてカタカナである名詞概念)であった。また頻度情報が 1000 以上の名詞概念は 197 あり、大きい順に 5 つ挙げると「熱泉」「盗泉」「炭酸泉」「単純泉」「アコーディオン」となった。

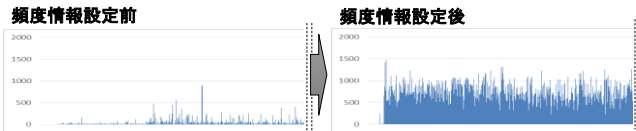
次に、頻度設定ができなかった名詞概念の例を表 3 に示す。タイプ A は対象の概念と共起する名詞概念が『青空文庫』に存在しな

表 1 共起情報計算式とそれぞれの結果

手法名	平均 (μ)	標準偏差 (σ)	" $\mu \pm \sigma$ " の範囲に含まれる割合
共起	46.59	139.18	94%
確率差	33.75	95.94	94%
確率比	10.13	32.38	96%
Jaccard	34.62	98.56	94%
Ochiai	46.59	139.18	94%

$$f = \frac{|S_a \cap S_b|}{|S_a|} \bigg/ \frac{|S_b|}{|S_{all}|}$$

図 6 確率比の計算



横軸:終端概念, 縦軸:頻度情報

図7 確率比に基づく計算による出現頻度設定の変化(一部)

った場合であり, タイプ B は対象の概念と共起する名詞概念が『青空文庫』に存在するが名詞概念辞書には存在しなかった場合であり, タイプ C は対象の概念と共起する名詞概念のすべての頻度が 0 だった場合である。A は合計 168, B は合計 9, C は合計 10819 存在した。

タイプ A に対処するためには対象データの量を増やす必要がある。タイプ B には対応する名詞概念が存在しない語を概念辞書に追加する必要がある。タイプ C のためには, 提案手法の反復による頻度情報獲得手法を考案した。例えば図 8(円は特定の名詞概念を, それらを結ぶ線は共起関係を示す)で, 状態①は共起情報を利用しない方式により得られた頻度の状態を示し, 状態②は共起情報を利用した方式による頻度設定を示す。この時, 概念 E は共起概念のすべてが頻度 0 のため, 頻度が 0 のままである。ここでさらに共起情報を用いて頻度を設定した場合が状態③に当たる。概念 A と F の平均を取ることで, 状態③における概念 E の頻度情報は 1 に設定される。

横光利一の 30 作品から得た頻度情報は, 共起情報を用いる二回の処理(図 8 では状態③)で変化しなくなり, 最終的に頻度が 0 のままだった概念は 9 つであった。これらは共起する名詞概念の数が他の概念と比べて少なく, 共起する概念の頻度情報が未更新だった。

3. 言語表記方法と言語表記辞書の改訂

ストーリー生成機構は, ストーリー技法により選択された動詞概念とそこに記述された格構造から個々の事象が生成され, 以上のような方法で事象中の名詞概念が決定される(事象は物語言説機構でも利用される)。名詞概念の内容は個別的なインスタンスでありそれが別の機構によって生成されるが, ある事象が「少年」という名詞概念の「太郎#1」というインスタンスを含む場合, その文表現において「少年」をそのまま使用する場合と「太郎」という名前を使用する場合がある。しかしここでは, 「少年」という名詞概念の名称を使用するものとして議論を進める。(さらに「少年の太郎」「若い少年である太郎」のような名詞句が使用されるだろう。)ここでは, 共起情報を利用した頻度設定の結果を, INGS による実際の生成に適用するが, 前提として事象の言語表記の方法及び言語表記辞書の一部に若干の改訂を加える。

3.1 事象の言語表記方法の改訂

文生成機構は, 生成された概念構造中の事象を逐次的に文に変換する。デフォルトでは概念辞書に記述された表記がそのまま使用されるが, 言語表記辞書を参照してそのヴァリエーションを増す。[鎌田 2013]が示したのはその一方法(1 節の最後), すなわち文字種類の割合の指定に基づき動詞, 名詞, 助詞の表記をランダムで決定する方法であった。これは単語単位で表記決定を行うが, 漢字の場合送り仮名の分だけ平仮名の割合が増え, その調整処理を行っていないので指定された品詞ごとの文字種類割合が正確に結果に反映されない。ここでは, 割合計算を純粋に文字単位で行えるようにこの方法を改訂する。図 9 はその改訂内容を示し, 詳細を次に説明する。

表 2 出現頻度の計算結果の例

概念	計算後の出現頻度
曲鉛	7.8
標準語	109.6
武士	0.8

表 3 頻度設定ができなかった概念の例

タイプ A	タイプ B	タイプ C
薬科大学	トーダンス	老師
占領国	モダンバレー	予知
コミックス	当り屋	名声

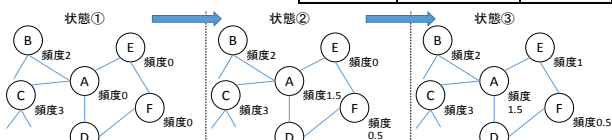


図 8 共起を用いた繰り返し処理による頻度設定のイメージ

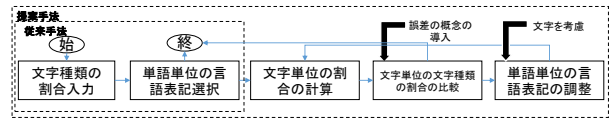


図 9 提案手法における従来手法からの改訂

ユーザによる指定に基づき単語単位で表記を決める処理は[鎌田 2013]と同じである。その結果に対して調整を行い, 指定された割合に近付ける。しかし, 直前の段落で述べた漢字表記における送り仮名や漢字表記を持たないカタカナ語等により, 文字種類の割合の指定を完全に満たせない場合があり, 今回はその差を 5% まで誤差として許容する。この調整は以下の手続きで, 名詞, 動詞, 助詞の順に行う—①生成された文列に含まれる当該の文字種類の割合を文字単位で求め, 指定された割合とここで計算した割合の差が 5% 以内なら③に, そうでなければ②に行く。②5% を上回った文字種類を持つ言語表記のすべての単語からランダムに一つ選択し, 指定された割合を下回る文字種類の文字が含まれた言語表記に変更する。変更候補が複数ある場合は指定された割合を下回る文字種類の割合が減少しない言語表記を優先的に選ぶ。それでも絞り切れない場合ランダムに決定する。①に戻る。③すべての品詞が処理済みなら文列を出力して処理を終了し, 未処理なら次の品詞を設定して①に行く。

3.2 言語表記辞書の小規模な改訂

なおこの方式を導入するために言語表記辞書を部分的に拡張した。まず漢字の送り仮名として従来ひらがな表記だけを格納していたが, カタカナ表記を追加した。例えば, 「愛する」に加えて「愛スル」を追加・格納する。さらに, カタカナ語は外来語を表す場合が多く, その表記はアルファベットなどで綴られた外国語の音をカタカナで記したものである。そのため, 小文字/大文字の使用, 長音の表した方等で表記にブレがある。これらを統一する方針を示す(以下は未実装)。

(1) カタカナにおける小文字の取り扱い

名詞概念辞書には, 「ウインドウ」や「ウイーク」のように母音がかたかなの大文字で書かれたものと小文字で書かれたものが混在している。この場合, [概念:ウインドウ]→[表記:ウインドウ, ウインドウ], [概念:ウイーク]→[表記:ウイーク, ウイーク]のように 2 つを併記する。また「ロマンティック」のような表現と「ロマンチック」のような表現も混在する。この場合も, [概念:ロマンティック]→[表記:ロマンティック, ロマンチック], [概念:ロマンチック]→[表記:ロマンチック, ロマンティック]のように併記する。現状の変更計画はここまでであるが, 実際の文章中にはさらに「ロマンティーク」のような例もあり, 今後は網羅的に表記の可能性を検討しより体系的に表記辞書改訂を行う。

(2) カタカナにおける長音の取り扱い

現状では「コンピュータ」, 「レディー」のように, 語尾に「ー」を付加するものとしがないものが混在する。そこで付加された名詞概念には付加されていない表記も併記し, 逆に付加されていない名詞概念には付加された表記を併記する。(例:[概念:コンピュータ]→[表記:コンピュータ, コンピューター], [概念:レディー]→[表記:レディー, レデイ])

4. 概念/表記選択の提案手法を統合する生成

2.3 節における頻度設定の結果と上述の表記方式とを適用して, INGS によるストーリー生成のシミュレーションを行う。

4.1 共起情報を利用した方法における頻度設定の向上

共起情報を用いた頻度設定とそれを用いない頻度設定とを比較する。方法であるが, 共起情報を用いない頻度情報と用いた頻度情報を使い, 一つのストーリー構造からそれぞれ 100 個ずつ生成を行う。ここで一つのストーリー構造とは, INGS で生成したストーリーに対して図 10 のように事象中の格を変数とした構造である。今回は, ストーリー生成技法の一種として使用されている, プロップに基づくストーリー

生成機構を用いる[Imabuchi 2012]. 100個ずつの生成では、制約条件と頻度情報に基づき変数を満たすべき名詞概念の選択を行う。

このストーリー構造では、35箇所まで事象における名詞概念選択が行われ、そのすべての箇所まで頻度情報を実際に使った概念選択が行われた場合その割合を100%とする。各名詞概念の制約としての中概念が複数存在する場合、ランダムにその何れか一つが選ばれる。表4に比較結果を示す。「要素数」とは、選択候補となった終端概念の数を示す。共起情報未使用の場合と比べ共起情報を使用した場合は、頻度選択実行の割合は大きく、またより多くの選択候補から実行されていることが分かる。

4.2 生成シミュレーション

共起情報を用いて計算した頻度情報を用いて生成されたストーリーに対して表記を付与する、生成シミュレーションを行う。ストーリーにおける各事象の範囲を制限する中間概念が複数ある場合はランダムにその一つが選ばれ、それに含まれる終端概念の中から最大の頻度を持つ概念が一つ選ばれる。こうして作られたストーリーが文生成機構によって文(単文)の集合に変換され、さらにその中の各単語に表記が付与される。表記付与には次の二つの方法が用いられる——一つは、横光利一の小説『上海』における品詞ごとの文字種別の割合の計算結果を用いる方法であり、もう一つは文の表面上の読みやすさの計算式[建石 1988]に基づく方法である。ここでは、4.1節で行った共起情報を使用した頻度情報設定方法による生成結果の中から一つを選択した。何れの方法の場合もこれを入力とする。

表5は、『上海』の文字種別の割合の計算結果である。計算に当たっては、台詞部分を除外し、品詞ごとに割合を求めた。この割合に基づく生成結果(誤差5%以内)を図11に掲げる。

次に、文の表面上の読みやすさに基づいて文字種類の割合を得るための方法について述べる。これには、上述した[建石 1988]による「読みやすさの評価式」(図12)を利用した。ここで、RSは評価、lsは文の平均長、laはアルファベット連の平均長、lhはひらがな連の平均長、lcは漢字連の平均長、lkはカタカナ連の平均長、cpは句点当たりの読点の数を表す。なお連とは、同一文字種の文字の一続きを意味する。但し今回は読点の数は0とする。本研究ではこの式を利用し、次の方法で読みやすい文字種別の割合を求めた(簡略化のため名詞・動詞の区別はせず、助詞はひらがな100%に固定)——①漢字・ひらがな・カタカナ・ローマ字の割合について[100%,0%,0%,0%]の状態から開始し、漢字の割合を5%減らし、右隣の割合を5%増やす処理を、漢字が25%になるまで繰り返す。次に、[25%,75%,0%,0%]の状態から同様にひらがなが25%になるまで繰り返す。他も同様の手続きを行う。さらにこの処理を文字種類の並び順を入れ替えて繰り返す。すなわち、[漢字、ひらがな、カタカナ、ローマ字]→[ひらがな、カタカナ、ローマ

(\$)プロパ民話...
(\$)予備部分
(event 出かける 1...
(agent [A]) (location [B])...)
(event 乱暴する 1...
(agent [C]) (location [B])...)

図10 変数の格を持つストーリーの一部 ([A][B][C]は変数)

表4 共起情報未使用/使用における頻度選択実行割合

要素数	共起情報未使用における割合	共起情報使用における割合
10以上	36%	79%
3つ以上	74%	96%
1つ以上	95%	100%

表5 『上海』における品詞ごとの文字種別の割合

文字の種類	名詞における割合	動詞における割合	助詞における割合
漢字	89%	37%	0%
ひらがな	5%	63%	100%
カタカナ	6%	0%	0%
ローマ字	0%	0%	0%

内大臣が巡歴に出かける。法律家が老人に乱暴する。老人が負傷する。衣料が老人に欠如する。法律家がシューターを殺害する。子爵がノクターンを歌う。老人が雪上から左辺まで来る。老人が青徴を知る。老人が霰に備える。老人が冒険へ出立する。老人が最盛期に向かう。老人が名誉市民と掴み合う。老人が名誉市民に幻視で負ける。弾薬が四季報を老人に教える。老人が弾道を辿る。老人が出札口に向かう。老人が抜け道から南南東へ移動する。法律家が「老人が玉杯を焼く」ことを老人に課する。老人が玉杯を焼く。老人が労働争議を解決する。老人が衣料を入手する。老人が南南東より脱出する。弾薬が南南東より脱出する。老人が左辺に到着する。書記が出目金を国母に要求する。法律家が真実を語る。書記の仮死が露見する。老人がロングスカートを着る。老人が幻像へ昇格する。国母が法律家を容赦する。国母が書記を容赦する。老人が出目金を得る。

図11 表5の文字の割合に基づき生成された文

$$RS = -0.12 \times ls - 1.37 \times la + 7.4 \times lh - 23.18 \times lc - 5.4 \times lk - 4.67 \times cp + 115.79$$

図12 読みやすさの評価式[建石 1988]

内大臣が巡歴に出かける。ほうりつかが老人に乱暴する。ろうじんが負傷する。衣料が老人に欠如する。ほうりつかがシューターを殺害する。ししやくがのくたーんを歌う。老人が雪上よりさへんまで来る。ろうじんが情かひを知る。ろうじんが霰に備える。ろうじんが冒険へ出立する。老人がさいせいきに向かう。老人がめいよしみんとつかみ合う。ろうじんが名誉市民に幻視で負ける。たんやぐが四季報をろうじんに教える。ろうじんが弾道を辿る。老人が出札口に向かう。ろうじんが抜け道から南南東へ移動する。ほうりつかが「ろうじんが玉杯をやく」ことをろうじんに課する。老人がぎよくはいを焼く。老人が労働争議を解決する。老人がいりようを入手する。老人がなんなんとうより脱出する。たんやぐが南南東から脱出する。老人がさへんへ到着する。しよきが出目金を国母に要求する。法律家が真実を語る。しよきの仮死が露見する。老人がろんぐすかーとを着る。ろうじんが幻像に昇格する。国母が法律家を容赦する。こくもが書記を容赦する。ろうじんが出目金を得る。

図13 読みやすさの計算式に基づく生成例

表6 読みやすさに基づく文字の種類別の割合

文字の種類	名詞における割合	動詞における割合	助詞における割合
漢字	40%	40%	0%
ひらがな	60%	60%	100%
カタカナ	0%	0%	0%
ローマ字	0%	0%	0%

字、漢字][カタカナ、ローマ字、漢字、ひらがな][ローマ字、漢字、ひらがな、かたかな]ごとの割合を作成し、重複分を除去する。結果として172種の組み合わせができる。②前の方法で用いたストーリー構造を使用し、すべての組み合わせごとに文字表記を付与し読みやすさの評価式を適用した。その結果、最も評価が高い文列は図13に掲げたもので、その文字種類の組み合わせは表6であった。この試みの詳細な検討は別稿に譲る。

5. おわりに

INGSでは現在、概念構造(ストーリーや物語言説)生成において、事象を構成する名詞概念を名詞概念辞書中の制約範囲の中から選択してインスタンス化し、文生成機構で個々の名詞概念に具体的表記を付与する。名詞概念選択に当たって文学作品の頻度情報を参照する方法は既に提案した[小野 2014]。本稿では、頻度情報が得られなかった名詞概念について、名詞どうしの共起関係情報を利用して仮の頻度情報を設定し、処理をしやすい方式を提案した。さらに、名詞概念に加え動詞概念と助詞も含めて従来の表記付与の方式[鎌田 2013]を改善する手法も提案し、上記の語の出現頻度と語どうしの共起関係を用いた名詞概念選択の方式とこの表記付与の方式をつなげた生成を試みた。提案方式によって頻度情報のみを利用する方式よりも頻度情報設定の割合が大幅に改善された。この結果の妥当性や有効性について検証することが今後の一つの課題である。表記に関しては、前提となる作業として、言語表記辞書自体の構成と内容の見直しと改訂・拡張が一課題となる。

参考文献

[Akimoto 2014] Akimoto, T. and Ogata, T.: An Information Design of Nanatology: The Use of Three Literary Theories in a Narrative Generation System, The International Journal of Visual Design, 7(3), 31-61, 2014.
[樋口 2004] 樋口耕一: テキスト型データの計量的分析—2つのアプローチの峻別と統合—, 理論と方法, 19(1), 101-115, 2004.
[Imabuchi 2012] Imabuchi, S. and Ogata, T.: A Story Generation System Based on Propp Theory: As a Mechanism in an Integrated Narrative Generation System, Lecture Note of Artificial Intelligence, Springer, 7614, 312-321, 2012.
[鎌田 2013] 鎌田まみ, 小方孝: 物語文章における文字表記の分析と模倣, 人工知能学会全国大会予稿集, 214-5in, 2013.
[小方 2010] 小方孝, 金井明人: 物語論の情報学序説—物語生成の思想と技術を巡って—, 学文社, 2010.
[Ogata 2015] Ogata, T.: Building Conceptual Dictionaries for an Integrated Narrative Generation System, Journal of Robotics, Networking and Artificial Life, 1(4), 270-284, 2015.
[小野 2014] 小野淳平, 小方孝: 計量データに基づく名詞概念の選択—統合物語生成システム—における一機構として—, 信学技報, 114(366), 49-54, 2014.
[建石 1988] 建石由佳, 小野芳彦, 山田尚勇: 日本語の読みやすさの評価式, 情報処理学会研究報告ヒューマンコンピュータインタラクション, 1988(25), 1-8, 1988.