2P1-6in1

語の頻度情報による概念・単語選択の改善 ―統合物語生成システムにおける利用―

Revising Concepts and Words Selection based on Word Frequency Information: The Use in an Integrated Narrative Generation System

吉田和樹[*1] Kazuki Yoshida
小野淳平[*2] Jumpei Ono
小方 孝[*1] Takashi Ogata

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

An automatic narrative generation system called INGS (Integrated Narrative Generation System), which we have been developing, selects linguistic concepts using the conceptual dictionaries for generating the organized structure of events in the story generation phase. For controlling the comprehensibility of concepts and the corresponding words, we have presented a method that uses appearance frequency information of words. In this paper, we revise the method by using many linguistic data more than previous attempts. In a result, we confirm that the number of frequency and the degree of easiness of reading is basically proportional.

1. まえがき

物語生成というタスクは階層的ないし多重的な特徴を持つ．それは，問題解決や推論といった知的行為，概念や意味の理解，言語やその他の記号イメージの生成といった種々のサブタスクを含み，重要度において何れの比重がより高いといったことは一般的にはない．物語生成システムはそれらすべての要素を体系的・有機的に取り扱う必要がある．小方のグループが研究開発を進めている統合物語生成システム (Integrated Narrative Generation System, 以下 INGS) （全体構想は[小方 10], 現状のシステム詳細は[Akimoto 14, Ogata 16]）は，上記のような物語生成に関与する多様な機能を有機的に統合することを目指すシステムである．問題解決や推論をはじめとする知能的処理の側面はそのストーリー生成機構や物語言説機構に含まれ，言語等の生成の処理は物語表現機構に含まれる．さらに概念や意味という側面に関しては，ある特殊な主題やジャンルを想定した物語生成における利用を超え，より一般的で広い利用可能性を考慮して，従来の物語生成システム研究におけるそれと比較するとかなり規模の大きな概念辞書の開発を上記 INGS の一機構として進めて来た[Ogata 15]．なお言語辞書研究の領域では，[Takeuchi 16]が日本語の言語シソーラスを物語生成に使用するための具体的方法について考察している．

INGS の文脈の中では，この概念辞書の主要な機能は，物語の基本単位である個々の事象を構成するための概念的・意味的知識を提供することであり，すべて階層的に構成されている．具体的には，概念辞書の一種である動詞概念辞書が，個々の事象の基本構造である格構造の形式を提供する．名詞概念辞書の終端概念が，この格構造を構成する人・物・場所その他の値を提供する（実際はそれをもとに固有対象を指示するインスタンスが生成される）．これらの概念辞書の規模が小さく，例えば特定の主題やジャンルの物語生成向きに収集された概念要素や言語要素に限定されている場合，概念やそれに伴う言語の選択・決定は容易であるが，本研究のように，規模が大きく一般的な概念辞書を使用する場合，様々な理解容易度の概念が混在してしまうことが問題になる．

なお，INGS では，概念も言語によって表現する．例えば，「食べる」のような概念記述を利用する．しかし同時に，「食べる 1」「食べる 2」等と番号でその意味的な違いを表現する．何れも自然言語では基本的に概念と同じ「食べる」という表記を取るが，これに文字表記の改変を加えて（概念辞書のすべての概念に対応する言語表記辞書を用意[小方 15]），「たべる」「タベル」「taberu」等とも表記可能なようにしている．

INGS のストーリー生成機構における事象生成の際の概念選択は，主に次の二種類の方式で行われる．一つは予め（目的に沿って）かなり限定された概念候補の中からランダムその他の方法によって概念を選択する方式であり，もう一つはあまり絞り込まれていない多数の概念群の中から同じく概念を選択する方式である．ここで問題となるのは後者である．具体的には，事象が生成される時，その格構造の値を事前に絞っておける場合は前者の方式が採用されるが，そうでない場合はより広い範囲からの選択となる．言い換えれば，概念値設定の制約が狭い場合と広い場合の違いであり，問題となるのは広い場合であり，様々な性格を持った概念が混在する可能性がある．この制約は具体的には，格構造における個々の格要素ごとに，概念辞書の階層における範囲指定として定められており，制約がより緩いとはこの範囲がより広いことを意味する．上述のように，INGS では概念辞書の要素は直ちに表層的な言語表現に結び付くので，この問題は概念辞書の使用の段階で解決しておく必要がある．なおこれは，INGS の概念辞書開発に使用された言語資源が時代的にも使用頻度の点においても非常に広範囲であることにも由来する．

事象における制約の緩い概念群の中から特定のものを選択するためには，まずそもそもの概念辞書の構成がより整備されていれば良い．すなわち，概念階層が様々な観点からより精緻に分類され，それぞれの分類の中に同類の概念のみが格納されていれば良い．しかしその作業は非常に手間がかかり，さらに，同一の観点からあまりに分類を細分化してしまうと，逆に使用における柔軟性を阻害する．そこで本研究では，物語生成の際の事象における概念選択をより意識的に行う方式の一つとして，語彙の頻度情報や共起情報等の統計的データを利用する

連絡先：小方孝，岩手県立大学ソフトウェア情報学部，岩手県滝沢市巣子 152-52，t-ogata@iwate-pu.ac.jp

ことにして，これまで幾つかの研究を行って来た．共起データの利用に関しては，[照井 16]に整理する．頻度情報の利用に関しては，主に小説テキストにおける単語の出現頻度情報を INGS における名詞概念の選択に利用することを試みた[小野 14a, 小野 14b, Ogata 16]．これらの研究では，明治から昭和にかけての著作権の切れた小説を中心とする文学作品のデータベースである「青空文庫」のすべての作品から，名詞及び動詞の出現頻度を計量し，INGS における同一表記の終端概念にその値を付与し，概念選択に利用する．約半数の終端概念に上記頻度情報を付与することができた．またこの方法で頻度情報を付与できなかった概念については，共起情報を利用してその値を推定することを試みている[小野 15, Ono 16]（2.3 節参照）．その上で，出現頻度の高低に基づく名詞概念選択によるストーリー生成実験を行い，高出現頻度の名詞概念を優先的に選択することにより，生成結果の語彙をユーザが受容した際の違和感（理解の難しさ）が軽減できることを確認した（但し，本研究の目的は必ずしも理解の難しい語彙の使用を排除することではなく，それを制御可能とすることであり，目的によっては故意に違和感のある語彙や理解を困難にするような語彙を使用することも想定される）．しかしながら，受容側の理解のしやすさや違和感に関するデータはこれまでの研究では実際の人間への調査を通じて行ったので，その数が少なく，網羅的なチェック・確認を行うには至らなかった．そこで本稿では，語彙の理解容易度・困難度を計量する既存の機械的方式を利用して，より包括的な評価の可能性を模索する．

2. 既存研究—語彙頻度情報に基づく概念選択と事象生成—

前述のように，これまで，『青空文庫』から得られた名詞及び動詞の語彙頻度情報を，INGS における名詞概念辞書及び動詞概念辞書に格納された名詞概念及び動詞概念に割り付け，これを利用してストーリー生成機構が事象を構成する動詞概念及び名詞概念を利用して事象生成を行う機構を開発した．従来の名詞概念辞書及び動詞概念辞書の中に得られた頻度情報を追加し，さらにストーリー生成における事象生成機構の中にこれを利用して概念選択を行う機構を追加したことに相当する．

2.1 名詞・動詞概念辞書の構成と事象生成時の利用

INGS は概念辞書として名詞概念辞書，動詞概念辞書，修飾概念辞書（形容詞概念，副詞概念，形容動詞概念）を含むが，頻度情報の処理を行っているのは現在名詞・動詞概念辞書のみである．

現状の動詞概念辞書は,11951 個の動詞概念とその意味分類 36 個から成る階層構造である．この終端概念に対応する個々の動詞概念は，格構造・制約条件・文型パターンにより定義される．このうち格構造は事象の格構造の規定であり，制約条件とは格構造の構成要素である名詞概念の値としての名詞概念辞書の範囲の定義である．文型パターンはこの格構造を表層文にする際の基本的な文型を定義する．

名詞概念辞書も中間概念と終端概念による階層構造を成し，図 1 に示す通り，動詞概念における上記制約条件は名詞概念辞書における中間概念によって記述される．INGS におけるストーリー生成機構による事象概念の生成に当たっては，その制約条件の範囲内で選択された名詞終端概念を使って，具体的な登場人物・物・場所（各インスタンス）が生成される．図 2 は生成される事象概念の記述例であり，#はインスタンスを示し，その前の「子供」「家屋」「ご飯」が選択された名詞概念の記述に相当する．なおこの例では表層的な記述は名詞概念辞書中の記述がそのまま使用されているが，言語表記辞書を使って「子供」が「こども」さらには「コドモ」のようになる場合もある）．

動詞概念に関しては（図 2 の場合「食べる 2」），ストーリー技法と呼ばれるストーリー生成機構中の機能により決定される．ストーリー技法の最も基本的なカテゴリーは，ストーリーコンテンツ知識（ストーリーコンテンツ知識ベースに格納される）と呼ばれる具体的な事象拡張知識を使って，入力となる事象もしくは複数の事象を含む構造を拡張してより多数の事象を含む構造を作り出すことであり，この処理の中で使用される特定の動詞概念もしくは動詞概念辞書のある範囲が指定される．無論選択が必要となるのは後者の場合であり，この範囲が広いと選択は難しくなる．

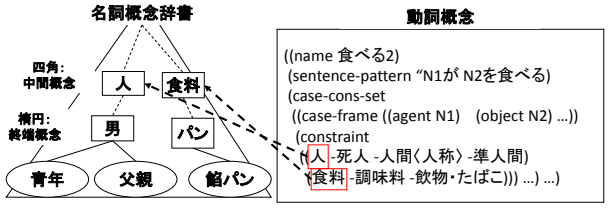

図 1 名詞概念辞書の構造と動詞概念との関係

```
(event 食べる 2 (type action) (ID 1) (time nil)
  (agent age%子供#1) (counter-agent nil) (location loc%家屋#1)
  (object obj%ご飯#1) ...)
```

図 2 事象概念の記述例

2.2 頻度情報を利用した動詞概念選択

[Ogata 16]は，『青空文庫』の収録テクストを対象に，KH Coder[Higuchi 04]を利用してすべての語彙の出現頻度を計算し，ある語彙の出現頻度をそれと表記が一致する動詞概念（但し動詞概念「食べる 2」なら「食べる」の部分）の出現頻度とし（全動詞概念 11951 個中 4886 の動詞概念に頻度情報を付与），この結果を用いた INGS のストーリー生成機構による生成実験を行い，違和感の値を予めユーザが付与した 37 個の動詞概念については，出現頻度の大きさと受け手が感じる違和感の大きさがほぼ反比例していることを確認した．ここからの類推により，動詞語彙の頻度情報と INGS における動詞概念の違和感の大きさとを機械的に関連付けて，主にストーリー生成機構における動詞概念選択を行うという基本方針を定めた．この選択結果は，表層的な文章生成にも直接的に反映される．

2.3 頻度情報を利用した名詞概念選択

[小野 14]は，上記動詞概念と同様の手続きとデータにより，全名詞概念 115765 個中 44332 個に頻度情報を割り付け，ユーザが予め付与した 115 個の名詞概念を用いた実験により，語彙の出現頻度と受け手が感じる違和感の間にはほぼ反比例の関係が成り立つことを確認した．ここからの類推により，名詞語彙の頻度情報と INGS における名詞概念の違和感の大きさとを機械的に関連付けて，主にストーリー生成機構における名詞概念選択を行うという基本方針を定めた．この選択結果は，表層的な文章生成にも直接的に反映される．

2.4 共起情報を利用した頻度情報の推定

頻度情報が 0 の場合，分析テクストを増やすことが一つの方法となるが，単純にテクストを増やすのではなく，共起情報をもとにその頻度情報を推定する方法について考案した[小野 15b, Ono 16]．具体的には，0 頻度の概念と共起関係にある概念の頻度情報の平均値をその概念の頻度情報と仮に見做す．実験的に，小説 30 作品から頻度情報を獲得し，0 頻度の概念つい

て，上記方法で頻度情報の推定を行った．今後はこの方法によって得られた推定頻度情報を前述の方法によって得られた方法と統合する予定である．

3. 語彙難易度判定研究を用いた頻度情報との関係の再調査

以上のように，動詞概念及び名詞概念の頻度情報を用いた概念選択の従来研究では，概念（語彙）が受け手に与える違和感を人間が人手で付与したが，その個数は非常に少なかった．そこで，既存の語彙難易度判定研究を用いて，より多くの語彙を対象とした再検証を行う．なおここで違和感とは，理解の困難さと不自然さの感覚等を含めた総合的な「感じ」を意味するが，下記の語彙の（理解の）難易度と最も関連が深いと考える．

3.1 方法

頻度情報を持つ動詞概念 4886 個，名詞概念 44332 を対象に，「チュウ太のレベルチェッカー」[Kawamura 13]の 7 種の判断基準を用いて，語彙出現頻度と語彙難易度の関係を調査する．表 1 にこのツールが提供する七種の判断基準及びそれぞれに含まれた語彙が上記の動詞・名詞概念をカバーする個数を示す．INGS における動詞概念辞書及び名詞概念辞書に格納される概念（のうちこのツールがカバーしているもの）をこの基準によって分類した．

表 1 判定基準と難易度判定可能な概念の数

判定基準	動詞概念	名詞概念
朝日新聞記事における頻度情報に基づく分類（10 段階表示）	2845	2155
朝日新聞記事における頻度情報に基づく分類（6 段階表示）	2581	3421
旧日本語能力試験出題基準に基づく分類（6 段階表示）	3066	6529
経済用語のレベルに基づく分類（7 段階表示）	1914	2866
筑波大学との共同研究による新基準に基づく分類（6 段階表示）	3940	8040
介護用語のレベルに基づく分類（4(5)段階表示）	2733	5105
単語親密度に基づく分類（4 段階表示）	4337	17509

3.2 結果と考察

図 3 及び図 4 に，最も INGS の各概念辞書における概念をカバーする範囲が広い基準である「単語親密度に基づく分類」に基づく結果を示す．それぞれの図において，左縦軸は概念の理解容易度を示し，数値が大きい程理解が容易であることを示す．右縦軸は概念の頻度情報を示す．なお頻度情報に関しては，最大値と最少値の差が大きくグラフが見づらいため常用対数によって表示している．横軸には，理解が容易な概念から

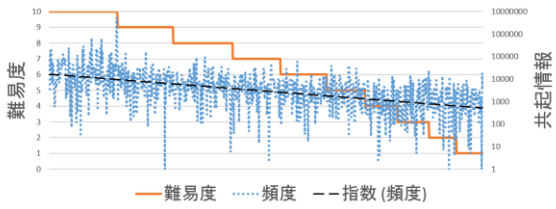

図 3 動詞概念における頻度情報と理解の難しさのグラフ
(単語親密度に基づく分類)

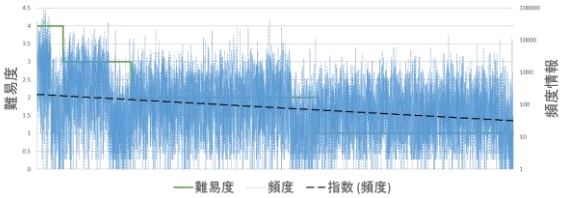

図 4 名詞概念における頻度情報と理解の難しさのグラフ
(単語親密度に基づく分類)

順に，図 3 では 4337 個，図 4 では 17509 個の概念が並ぶ．グラフを横断するように伸びている破線は，指数近似による頻度情報の推移の近似曲線である．以上のグラフを通じて，その近似曲線と難易度の比例関係から，おおよそ 3.1 節で述べた仮定を満たす傾向を持つと判断できる．しかし特徴的な現象も見られたので，以下に考察する．

「朝日新聞記事における頻度情報に基づく分類」は 10 段階及び 6 段階のどちらも，各難易度の段階における概念の総数は，他の基準に比べて均等に判定されている（図 5）．すなわち，一つの段階に含まれる概念の数の標準偏差の値が小さいため，各段階に含まれる概念の数のばらつきが少ない．（これらの基準と比べ，他の基準は標準偏差が 10 倍以上の値となる．）この基準を利用した場合，例えばある広い選択範囲で選択される概念の種類が一種類だけになってしまうようなことを少なくできると考えられる．

次に，仮定と異なる結果を示した概念について考察する．ここで仮定と異なる結果とは，「出現頻度が少ないにも拘らず理解容易度が高い（理解が容易）」あるいは「出現頻度が多いにも拘らず理解容易度が低い（理解が困難）」ことを意味する．表 2 に各判定基準において該当する動詞概念及び名詞概念をそれぞれ 10 ずつ示す．まず，「出現頻度が少ないにも拘らず理解が容易である」ものについては，カタカナ語が特に多く含まれることが分かる．これについては，本研究で出現頻度を計量したテクストが『青空文庫』であるところから，比較的古い語彙が多いことに起因すると推測される．本来，これらのツールの作成に当たって使用されたコーパスの年代を考慮して頻度調査対象のテクストを選ばなければ正確な結果を見ることは難しい．逆に「出現

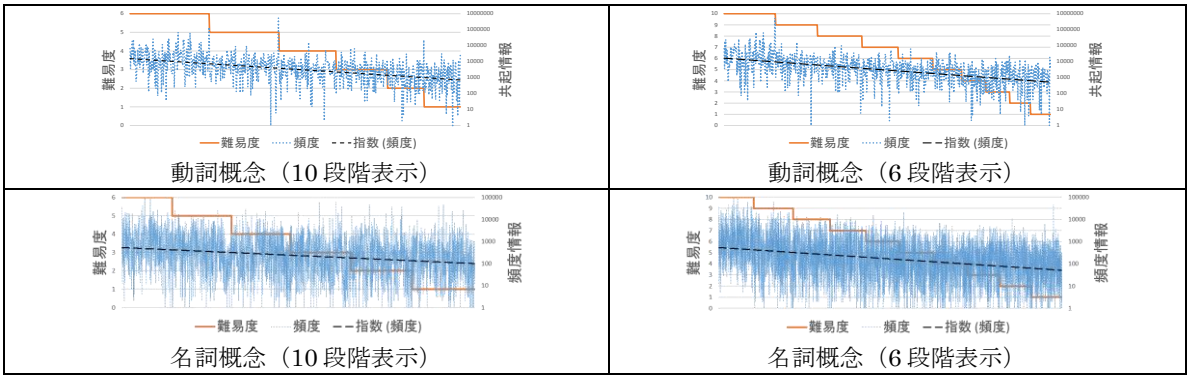

図 5 頻度情報と理解の難しさのグラフ(朝日新聞記事における頻度情報に基づく分類)

表 2 仮定と異なる結果となった動詞概念

判定基準	名詞概念		動詞概念	
	頻度多―理解困難	頻度少―理解容易	頻度多―理解困難	頻度少―理解容易
朝日新聞記事における頻度情報に基づく分類(10 段階表示)	眼, 助, 何事, 巡査, 書物, 観念, 舌, 病人, 脚, 悪魔	疑問, 野党, 喪主, 路線, 自治体, 業界, 員, 融資, 減税, メーカー	下る, 生える, 足る, 縫う, 脱ぐ, 震える, ごまかす, 枯れる, 触る, 散らす	増える, 当たる, 果たす, 強める, 回る, 決まる, 受け入れる, 超える, 含める, 設ける
朝日新聞記事における頻度情報に基づく分類(6 段階表示)	手, 心, 下, 口, 夢, 犬, 死, 石, 愛, 全集	党首, 税制, 機器, 同省, 核兵器, 与野党, 当たり, 株価, パソコン, エイズ	切る, しまう, 酔う, 見上げる, 越す, 狂う, 折る, 散る, 吐く, 利く	図る, 広がる, 目指す, 務める, 下がる, 増える, 見込む, 固める, 減らす, 当たる
旧日本語能力試験出題基準に基づく分類(6 段階表示)	ファイル, 死, 勧告, 扉, 世, 傍, おなか, ストーブ, タクシー, 花瓶	手洗い, 朝, ペット, パーティー, カレンダー, シャワー, 交差点, コピー, 御手洗, 色色	及ぶ, 経る, 辿る, 告げる, 呟く, 保つ, 負う, 貫く, 生かす, 凝る	跳ぶ, 被る, 射す, 要る, ほてる, 磨く, 撮る, 終わる, 閉める, 締める
経済用語のレベルに基づく分類(7 段階表示)	姿, 作成, 例, もと, 考え, 存在, 労働, 方法, 一般, 地方	リポート, 消しゴム, ハンバーグ, バレーボール, すばらしさ, 受け付け, ガソリンスタンド, 色色, パソコン, サッカー	示す, 得る, 求める, 守る, 向ける, 加える, 迫る, 及ぶ, 合わせる, 生じる	沸く, 渇く, 留まる, 片付ける, 閉まる, 履く, 点く, 弾ける, 無くす, 引っ越す
筑波大学との共同研究による新基準に基づく分類(6 段階表示)	眼, 世間, 帆, 我, 吉, 小屋, 怪, 縁, 言, 屋敷	手洗い, 朝, ペット, パーティー, カレンダー, シャワー, 交差点, コピー, 御手洗, 色色	感じる, 生じる, 経る, 下す, 辿る, 放つ, 告げる, 呟く, 応じる, 狂う	起きる, 作る, 切る, 寝る, 飲む, 住む, 食べる, 着る, 会う, 待つ
介護用語のレベルに基づく分類(4(5) 段階表示)	ファイル, 死, 勧告, 扉, 世, 傍, ストーブ, タクシー, 花瓶, 食べ物	ギター, 万, 何分, おじいちゃん, ジュース, 朝, ペット, カレンダー, 苹果, サッカー	及ぶ, 経る, 辿る, 告げる, 呟く, 負う, 貫く, 生かす, 凝る, 試みる	跳ぶ, 被る, 射す, 要る, ほてる, 磨く, 撮る, 終わる, 閉める, 締める
単語親密度に基づく分類(4 段階表示)	校正, うち, もと, 勧告, 帆, 我が, 室, 相違, 全集, 地	シャンプー, サッカー, コミュニケーション, コマーシャル, ケチャップ, クラブ, グラタン, ギフト, いぬ, アレルギー	する, 居る, 申す, 貰う, 連れる, 喋る, 湧く, 溢れる, 成る, 下す	燃える, 負ける, 苦しむ, 遊ぶ, 運ぶ, 飛ぶ, 起きる, 乗る, 押す, 売る

頻度が多いが理解が困難」の方であるが, 一語のみで出現する名詞が多く, 動詞については比較的古い表現が多い.

4. あとがき

筆者らが開発を進めている INGS では, 事象を構成する動詞及び名詞概念及び対応するそれぞれの語彙の選択のために, テクスト資料における語彙の出現頻度を利用し, 受け手における理解難易度と対応付ける方法を用いている. これまでの筆者らの研究ではこの対応付けのための理解難易度に関するデータが少なかったため, 本稿では既存の理解難易度判定ツールを用いてより多くのデータによる検証を行った. その結果, 動詞概念の場合も名詞概念の場合も, 語彙の出現頻度の大きさと対応する概念の理解容易度とは基本的に比例関係にあることが分かったので, 今後 INGS の中にこの方法を本格的に組み込んで行くことにする. 但し, 比例関係にない概念も存在するので今後その原因を追究する. 寧ろこの種の現象が物語生成の技法として有効な意義を持つ可能性もあるので, 今後検討を進める. これはまた, 選択するテクスト資料の性格(ジャンル, 時代等)と関連することが予想されるので, その検討も今後の課題である. テクスト資料の変更によって異なる結果を得ることができれば, これを生成戦略の中に組み込むことも可能だろう.

参考文献

[Akimoto 14] Akimoto, T. and Ogata, T.: An Information Design of Narratology: The Use of Three Literary Theories in a Narrative Generation System, The International Journal of Visual Design, 7(3), 31-61 (2014)

[Higuchi 04] Higuchi, K.: Quantitative Analysis of Textual Data: Differentiation and Coordination of Two Approaches, Sociological Theory and Methods 19(1), 101-115 (2004)

[Kawamura 13] Kamuwara Y.: The Basic Concept for Multilingualization of the Reading Tutorial Dictionary Tool, The 8th Symposium on Japanese Language Education in Europe (2013)

[小方 10] 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)

[小方 15] 小方孝, 小野淳平:統合物語生成システムにおける言語表記辞書とその利用, 信学技報, 115(70), 25-30 (2015)

[Ogata 15] Ogata, T.: Building Conceptual Dictionaries for an Integrated Narrative Generation System, Journal of Robotics, Networking and Artificial Life, 1(4), 270-284 (2015)

[Ogata 16] Ogata, T.: Introduction: Computational and Cognitive Approaches to Narratology from the Perspective of Narrative Generation, In Ogata, T. & Akimoto, T, Computational and Cognitive Approaches to Narratology, IGI Global (in press)

[Ogata 16] Ogata, T. and Ono, J.: Controlling the use of Semantic Concepts in an Integrated Narrative Generation System: The Use of the Verb Frequency Information, Proceedings of International Symposium on Artificial Life and Robotics, (2016)

[小野 14a] 小野淳平, 小方孝:小説データに基づく統合物語生成システムの概念・語彙選択, 人工知能学会第二種研究会ことば工学研究会(第 47 回)資料, 47-53 (2014)

[小野 14b] 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)

[小野 15] 小野淳平, 小方孝: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, 第 29 回人工知能学会全国大会論文集, 3G4-OS-05a-3, (2015)

[Ono 16] Ono, J. and Ogata, T.: A Way in Verb Concept Selection using Co-occurence Information of Verb Concpets: A Mechanism in an Integrated Narrative Generation, Proceedings of The 4th IIAE International Conference on Industrial Application Engineering 2016, (2006) (in press)

[Takeuchi 16] Takeuchi, K.: Thesaurus with Predicate-Argument Structure to Provide Base Framework to Determine States, Actions, and Change-Of-States, In T. Ogata & T. Akimoto, Computational and Cognitive Approaches to Narratology, IGI Global (in press)

[照井 16] 照井和舎, 小野淳平, 小方孝:語の共起情報による概念・単語選択の改善―統合物語生成システムにおける利用―, 2016 年度人工知能学会全国大会(第 30 回)予稿集 (2016) (印刷中)