結束性を考慮した連体修飾節の言い換え

野上優 *1　乾健太郎 *2 *3

*1 九州工業大学大学院情報工学研究科

*2 九州工業大学情報工学部知能情報工学科

*3 科学技術振興事業団さきがけ研究21「情報と知」領域

{m_nogami,inui}@pluto.ai.kyutech.ac.jp

1 言い換えを利用した結束性評価基準の研究

テキストの翻訳，要約，簡単化など，表層のテキストを出力とするタスクでは，出力テキストの結束性を考慮する必要がある．質の高いテキストには，各文が節レベルで適格であるだけでなく，文どうしの繋がり（結束性）が良いことが要求される．文レベルの処理の技術が成熟しつつある今日，結束性の良さを考慮した翻訳や要約の研究に取り組む試みもいくつか見られるようになってきた．たとえば，結束性を考慮した訳文修正処理を機械翻訳の後処理に組み込む試み [1] や，重要文抽出によって得られた要約テキストを結束性の観点から修正するモデル [8, 12] などがすでに報告されている．

これらの処理に共通して必要となるのは，出力テキストの結束性の良さを評価する技術である．テキストの結束性については，テキスト生成の主要課題として1980年代から精力的に研究されてきた [4, 16]．しかしながら，結束性の評価基準を誰もが利用できるような明示的な形で（たとえば規則の集合として）提示した例は極めて少ない．主な理由としては，(a) テキスト生成研究ではテキストの内容を表す何らかの中間表現を入力と仮定するので，多様な内容のテキストを生成する大規模な実験が実現困難であった，(b) 既存のテキスト生成モデルの多くが生成過程を構成する個々の選択の中に結束性の考慮を埋め込んでいるため，知識の再利用が困難であった，といったことが挙げられる．一方，機械翻訳においても結束性を考慮した研究は極めて少ない [13]．これは，(a) 訳語選択や構造変換など節レベルの問題だけでもハードルが高かった，(b) 頑健な文脈解析が困難だった，などの理由による．また，冒頭で述べたような試みも，必ずしも結束性の評価方法の問題に踏み込んでいるとは言えない．本研究の目的は，与えられたテキストの結束性の良さを評価するための規則集合を構築し，再利用可能な形で提示することである．

テキストの結束性は，what-to-say レベルの結束性（首尾一貫性，coreherence）と how-to-say レベルの結束性（結束構造，cohesion）の2つのレベルに分けて捉えることができる．本稿では，このうち結束構造の良さを評価するタスクを考える．すなわち，適当な修辞構造を持つ節の集合が与えられたとき，それに対応する表層のテキストを次のような観点から評価するタスクである:

- 節の順序の適切さ
- 主題陳述構造の適切さ
- 節間の修辞関係を表す接続表現の適切さ
- 照応表現の適切さ

上述のようなタスクを実現するためには，同じ内容（修辞構造）に対して様々な結束構造を持つテキストを複数生成し，それらを人間が横並びで比較したような言語データを用意するのが望ましい．それには，以下の理由から，言い換えを利用するのが良い．

- 言い換えによって変更される部分は文全体のごく一部なので，節レベルの適格性を確保するのが比較的容易．
- 言い換えの選択肢を組み合わせることによって，結束構造の良いテキスト（正例）と悪いテキスト（負例）をシステマティックに生成することが可能．
- 一般のテキスト生成と異なり，表層のテキストを入力とするので，入力内容の多様性を確保するのが容易．

結束構造を大きく変化させる言い換えの代表的なものに，連体修飾節と主節の分割がある（本稿ではこれを「連体節主節化」と呼ぶ）．たとえば，次の元テキスト (前) は，(後1) のように言い換えても結束性は保存されるが，(後2) のように言い換えると後続文脈との繋がりが悪い．

(前) スウェーデンの首都ストックホルムから南西部に位置するスモーランド地方は別名「ガラスの王国」とも呼ばれている．(以下後続文脈) この地方にある二つの大きな町カルマルからベクショーにかけて，十六ものガラス工場が点在しているからだ．

(後1) スモーランド地方はスウェーデンの首都ストックホルムから南西部に位置する．別名「ガラスの王国」とも呼ばれている．(以下後続文脈) この地方にある二つの大きな町カルマルから ... ガラス工場が点在しているからだ．

(後2) ＊スモーランド地方は別名「ガラスの王国」とも呼ばれている．スウェーデンの首都ストックホルムから南西部に位置する．(以下後続文脈) この地方にある二つの大きな町カルマルからベクショーにかけて，十六ものガラス工場が点在しているからだ．

そこで，本研究では，上例のような連体主節化の言い換えを取り上げ，言い換え後のテキストの結束構造（以下，単に「結束性」と呼ぶ）を評価するタスクに取り組んだ．本稿では，連体節主節化を構成する選択点を示したあと，これまでの事例分析から得られた結束性評価基準を網羅的に述べ，予備実験の結果を報告する．

2 言い換え事例の収集と分析

連体修飾節は限定節，非限定節，内容節に大別できる [9]．このうち，主節化が最も自然で容易に行えるのは非限定節である．そこで，京大コーパス [7] の一部（文化・読書・芸能・特集面全部，及び総合面の一部，1,840文）から非限定連体節[1]を含む文を網羅的に収集したところ，275件の非限定的連体節の事例を得ることができた．

つぎに，収集した各事例について，先行・後続文脈を無視して，可能な言い換えを（可能ならば複数）作成した．ただし，一文に複数の非限定的修飾節が存在する場合は，それぞれの言い換えを別々の事例として扱った．また，一つの連体節について複数の言い換えが可能な場合も異なる言い換え事例と見なした．

[1] 連体節の定義は文献 [9] に従った．

事例の分析は以下の手順で行った．まず，収集した言い換え事例について，原文と主節文・連体節文を構文的に比較し，連体節主節化を構成する選択点と選択肢をあらい出した．詳細は 3 節で述べるが，たとえば主節文と連体節文の順序の選択などの選択点が明らかになった．

つぎに，275 箇所の非限定的修飾節のうち 195 箇所について，任意の選択肢の組み合せるを適用することにより，1,343 件の言い換え事例を作成した．得られた言い換え事例の各々について先行・後続文脈との「つながりの良さ（結束性）」を評価したところ，結束性に関して深刻な問題のない事例（正例）が 449 件，結束性の低くテキストとして受理できない事例（負例）が 841 件，正例・負例の判断ができない事例が 53 件あった．また，正例については，同じ文を言い換え元とする正例の言い換えが複数あったときに限り，それらの間の相対的な結束性の高さを評価し，（半順序に）ランクづけした．なお，言い換え可否の判断や結束性の評価はすべて作業者の主観で行った．

最後に，正例と負例を比較・分析することにより，結束性を評価するための制約・選好に関する仮説を立てた．ここで，制約は正例と負例を弁別するための規則を指し，制約を満たさない事例は負例と評価する．一方，選好は，複数の正例が存在する場合に，それらをランクづけするための規則である．制約・選好の詳細は4 節で述べる．

3 連体節主節化における選択点

連体節主節化を構成する選択点には，先行・後続文脈との結束性に影響を及ぼさない（逆に，文脈から影響を受けない）結束性独立な選択点と，文脈から影響を受ける結束性依存な選択点がある．

3.1 結束性独立な選択点

結束性独立な選択点では，主として，一文内の統語的・意味的制約を満足するように主節，連体節を変形する．詳細は文献 [15] を参照されたい．

(1) 主節文と連体節文の間の接続表現 主節文・連体節文のうち後続する方の文頭に接続詞（接続表現）を挿入した方が良い場合がある．

(2) 連体節文のテンス・アスペクト表現の修正 原文における連体節と言い換え後の連体節文とでは，テンス・アスペクト表現が必ずしも一致しない．この修正は，「ル形／タ形」の選択と，アスペクト表現（「テイル／テアル」など）の有無の選択の組み合わせからなる．

(3) 連体節文中の格助詞「の」の交替 連体節文中にガ格を表す格助詞「の」がある場合，これを格助詞「が」に置き換える．

(前) 「女ぶり」は大女優・山田五十鈴との共演で，新人育成に定評のある平岩弓枝の作・演出．

(後) 「女ぶり」は大女優・山田五十鈴との共演で，平岩弓枝の作・演出．平岩は新人育成に定評がある．

3.2 結束性依存な選択点

(1) から (4) の詳細については文献 [15] を参照されたい．

(1) 主節文と連体節文の順序の選択 主節文と連体節文のどちらを先行させるか．

(2) 連体節文におけるギャップの復元 被修飾名詞を連体節文のギャップの位置に復元したあと，それを主題化するか否かの選択がある．

(3) 連体節文の名詞述語化 連体節文が被修飾名詞に対して属性的情報を付加している場合，それを明示するために，「（〈被修飾名詞〉は...する）○○である」といった名詞述語表現を補った方が良い場合がある．

(4) 照応表現の選択 主節文内の被修飾名詞と連体節文に挿入された被修飾名詞の共参照関係を明示するために照応表現の選択が必要になる．指示連体詞の挿入やゼロ代名詞化などが主な選択肢である．

(5) 後続文脈の主題の復元 テキストの結束性を保持するために，省略されている後続文脈の主題[2] を復元した方が良い場合がある．

(前) 電子新聞に積極的なサンノゼ・マーキュリーは，全世界的なネットワークに急成長しているインターネットからもアクセスできるサービスを開始する．(以下後続文脈) 日本企業の広告マーケットへの進出も模索している．

(後) 電子新聞に積極的なサンノゼ・マーキュリーは，インターネットからもアクセスできるサービスを開始する．というのは，インターネットが全世界的なネットワークに急成長しているからだ．(以下後続文脈) サンノゼ・マーキュリーは日本企業の...

(6) 同一主題の削除 連続する二文（主節文と連体節文，または主節文・連体節文のうち後続する方と後続文脈）の主題が一致する場合，同じ主題が続いて冗長なため，後続する文の主題を省略した方が良い場合がある．

(前) 一七四二年に創立されたコスタは，スウェーデン最古の工場だ．

(後) コスタは，スウェーデン最古の工場だ．一七四二年に創立された．

4 結束性を評価するための制約・選好

現在までに得られている制約・選好の仮説を網羅的に列挙する．

4.1 節の順序に関する制約・選好

主節文と連体節文の順序関係は，南による提案としてよく知られる「階層構造に基づく節間の包含関係の制約」[11] に従うと考えられる．南によると，従属節はスコープの包含関係の広さに基づいて 3 種類に分類でき，スコープの広い従属節は狭い従属節を包含できるが，逆にスコープの狭い従属節が広い従属節を包含できない．白井らはこの制約を利用し，従属節の係り受け解析で成果を上げている [17]．この制約を我々の問題に適用するには，次の 2 点の拡張が必要である:

- 南，白井らが扱った対象は，いずれも文内の節間の包含関係に限定されていた．我々の問題においては，まず，従属節の分類を修辞的関係の分類と見なす必要がある．さらに，同様の制約が文を越えた談話セグメントにおいても成り立つかどうか，また主部が従属部に先行する場合にも成り立つか，を調査する必要がある．
- 主節文と連体節文の間の修辞的関係は「対比・逆接」「原因・理由」「継起」「付帯状況」「情報付加」のいずれかであるが [10]．このうち「情報付加」は南，白井らの分類には存在しない．そこで新たに，「情報付加」を既存の「修辞的関係の分類」のどこに分類すべきかを調査する必要がある．

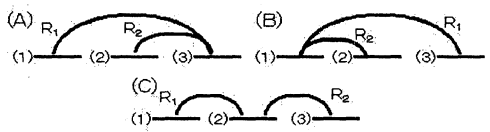
以上の観点から事例分析を行った結果，現在のところ次の仮説が得られている．

(制約 1-1) 先行文脈，主節文，連体節文，後続文脈が下図のようなパタンの修辞構造を形成するとき（図の (1),

[2] 主題の定義は文献 [14] に従った

(2), (3) はそれぞれ先行文脈，主節文，連体節文，後続文脈のいずれか），以下の制約が成り立つ．

- 「情報付加」は，白井らの分類 [17] の「B 類（「原因」など）」と「C 類（「対比」など独立の関係）」の間のクラスに分類される．
- パタン (A) または (B) の場合，修辞的関係 R_1 は R_2 より上位の階層の関係でなければならない．ただし，R_1，R_2 の主部，従属部の順序は問わない．
- R_2 が「情報付加」のとき，パタン (A) は不可．



たとえば，次の (後 1) は，パタン (A) で R_1 = 対比・逆接，R_2 = 情報付加 である．一方 (後 2) は，R_1 と R_2 が逆転しているので，不適格である．

(前) 価格破壊の主役となった流通業界は「増収」と「減収」で見方が二つに割れ，「金融」「運輸・サービス」では「影響なし」が多数を占めた．

(後 1) 流通業界は「増収」と「減収」で見方が二つに割れた．流通業界は価格破壊の主役となった業界である．一方，「金融」「運輸・サービス」では「影響なし」が多数を占めた．

(後 2) * 流通業界は「増収」と「減収」で見方が二つに割れた．一方，「金融」「運輸・サービス」では「影響なし」が多数を占めた．流通業界は価格破壊の主役となった業界である．

この他，連体節主節化に依存した形の規則がいくつか得られている．これらは今後，さらに整理し，一般化する必要がある．

(制約 1-2) 主節文と連体節文に共参照関係がある場合，連体節文は先行できない．

(前) オードリー・ヘプバーンが亡くなって丸二年たつが，WOWOWでは彼女が最後に出演した「世界の庭園」を八日から放送する．

(後) オードリー・ヘプバーンが亡くなって丸二年たつが，WOWOWでは「世界の庭園」を八日から放送する．「世界の庭園」は彼女が最後に出演した番組である．

(制約 1-3) 先行文脈があり，新情報の被修飾名詞を情報付加している場合，連体節文は先行できない．

(先行文脈) 以下はアメリカンファミリー生命保険会社の協力を得て実施した全国世論調査結果の詳報である．

(前) 調査は，層別多段無作為抽出法で選んだ全国の 20 歳以上の男女 4000 人を対象に留め置き法で実施した．

(後) 調査は，全国の 20 歳以上の男女 4000 人を対象に留め置き法で実施した．この 4000 人は層別多段無作為抽出法で選んだ．

(制約 1-4) 原文の先頭に接続表現がある場合，連体節文は先行できない．

(前) 一方，すっかり市民権を得た「価格破壊」は 4 社だった．

(後) 一方，「価格破壊」は 4 社だった．すっかり市民権を得た言葉である．

(選好 1-1) 主節文と先行文脈に共参照関係がある場合，主節文先行の方が良い．

(先行文脈) 中村屋 ... だが，パン屋だった中村屋がライスカレーを食べさせるようになったのは昭和初年だった．

(前) そのきっかけは中村屋がインド独立運動の闘士であったビハリ・ボースをかくまうようになったことにある．

(後) そのきっかけは中村屋がビハリ・ボースをかくまうようになったことにある．ビハリ・ボースはインド独立運動の闘士であった．

(選好 1-2) 時間関係のある文は時間順に並べた方が良い．

(前) ギングリッチ議長が大統領夫妻をどう評価していたか質問したチャンさんに，キャサリーンさんは最初「言えません」と拒否した．(以下後続文脈) チャンさんが「ここだけの話でこっそり話してくれませんか」と重ねて頼むと「彼女はあばずれだ，と言っていたわ」と語った．

(後) チャンさんはギングリッチ議長が大統領夫妻をどう評価していたか質問した．キャサリーンさんは最初「言えません」と拒否した．(以下後続文脈) チャンさんが「ここだけの話でこっそり話して ...

(選好 1-3) 主節文と連体節文に対比・逆接の関係がある場合，連体節文先行の方が良い

(前) 一九九六年米大統領選の共和党有力候補の一人とみられていたディック・チェイニー元国防長官は三日，出馬を断念する考えを明らかにした．

(後) ディック・チェイニー元国防長官は一九九六年米大統領選の共和党有力候補の一人とみられていた．しかし三日，出馬を断念する考えを明らかにした．

4.2 主題化に関する制約・選好

(制約 2-1) 連体節文におけるギャップの格がガ格のときは，常にこれを主題化する．ただし，連体節文が理由を表す「〜からだ」を持つ場合，あるいは連体節文先行で先行文脈が存在しない場合を除く．

(前) プロディジィでサービスを行っているアクセス・アトランタの基本料金は月額約七ドル．

(後) アクセス・アトランタの基本料金は月額約七ドル．アクセス・アトランタはプロディジィでサービスを行っている．

(選好 2-1) ギャップの格がガ格で，理由を表す「〜からだ」の文では，主題化より格補完の方が良い [14]

(前) 逆に，輸出にほとんど依存しない竹中工務店，アサヒビール，サントリーは一〇〇円を適正水準と回答した．

(後) 逆に，竹中工務店，アサヒビール，サントリーは一〇〇円を適正水準と回答した．これら三社が輸出にほとんど依存しないからだ．

(選好 2-2) ギャップの格がガ格以外の場合，格補完より主題化の方が良い．

(前) ＣＤに入っているのはシュラルメンの「ウィーンはウィーン」やソプラノのメラニー・ホリデイが歌う「公爵さま」など六曲．

(後) ＣＤに入っているのはシュラルメンの「ウィーンはウィーン」や「公爵さま」など六曲．「公爵さま」はソプラノのメラニー・ホリデイが歌っている．

4.3 接続表現に関する制約・選好

(制約 3-1) 主節文と連体節文に原因・理由の関係がある場合，主節文先行ならば，連体節に文末表現「〜からだ」を付加する必要がある．(例文は (選好 2-1) を参照)

(制約 3-2) 主節文と連体節文に逆接の関係があり，連体節文先行の場合，その関係を明示する接続表現が必要．（例文は (選好 1-3) を参照）

4.4 照応・省略に関する制約・選好

今回はあまり深くは分析しなかった．これについては，他の研究成果 [3] を取り入れる予定である．

(制約 4-1) 省略されている後続文脈の主題と直前の文の主題が一致しない場合，後続文脈の主題を復元しなければならない．(例文は3.1の (5) を参照)

(制約 4-2) 連続する二文の主題が一致しない場合，主題を削除してはいけない．

(前) 大統領の「風邪宣言」は 調停の最後のヤマ場になるとみられていた 四日のベルルスコーニ氏との会談 の直前に発表された．

(後) 大統領の「風邪宣言」は四日のベルルスコーニ氏との会談の直前に発表された．この会談は調停の最後のヤマ場になるとみられていた．

(選好 4-1) 連続する二文の主題が一致する場合，後続する文の主題は省略した方が良い [6]．(例文は3.2の (6) を参照)

5 実験

2 節で述べた言い換え事例のうち（結束性の評価ができなかった事例は除く），連体節の長さが 3 文節以上である事例 100 箇所（言い換え事例 360 件）を無作為に選び，前節の制約・選好を適用した結果と人手による結束性の評価結果を比較するクローズドテストを行なった．

5.1 実験環境

実験には，我々のグループで開発中の言い換えエンジン [2] を用いた．入力には，京大コーパス [7] から抽出した形態素・構文情報に，その他の構文・意味・談話情報（ギャップの格，修辞的関係，主題（省略されているものも含む），従属節の分類，被修飾名詞の新・旧情報）を人手で付加したものを与えた．

言い換えテキスト生成部は入力を受け取ると，結束性独立な選択点については最適な選択肢を一つだけ選択し，結束性依存な選択点については任意の選択肢の組み合せを試み，個々の組み合せについて言い換えテキスト候補を一つ出力する（ただし，照応表現の選択と名詞述語化は未実装である）．したがって，出力は節レベルでは適格であるが，談話レベルの結束性は保証されない．結束性評価部は，各言い換え候補に対して制約・選好に関する規則を適用し，各規則に応じたスコア付けを行なう．

5.2 実験結果

まず，制約の有効性を検証するために，制約による正例・負例の弁別結果と人手による弁別結果がどの程度一致するかを調べた．結果を表 1 に示す．負例については，不正解の 39 事例のうち，未実装の名詞述語化および照応表現が原因のものが 16 事例あった．それらの事例を除くと，実質的な正解率は 89.1% である．クローズドテストではあるものの，多様な事例の結束性評価を少数の制約である程度説明できたと言える．

つぎに，選好の有効性を検証するために前述の 100 箇所の連体節のうち，制約を満たす言い換え候補が 2 個以上が存在した 37 箇所について，選好による正例のランクづけと人手によるランクづけがどの程度一致するかを調べた．結果を表 2 に示す．「全部評価」の「一致」は候補集合全体のランクづけが人手のものと無矛盾であった事例数，「1 位評価」の「一致」は 1 位の候補が人手の評価と一致した事例数を指す．制約の場合と同様，小規模ではあるが，少数の選好によって人手によるランクづけをある程度説明できている．

表 1: 正例・負例の弁別能力

	弁別成功	弁別失敗	総数	正解率
正例	116	17	133	87.2%
負例	188	39	227	82.8%

表 2: 正例間のランクづけの能力

	一致	不一致	総数	ランクづけ正答率
全部評価	26	11	37	70.3%
1 位評価	35	2	37	94.6%

6 おわりに

システマティックに生成された言い換え事例を分析することにより結束性評価基準を構築する試みについて報告した．得られた制約・選好は仮説にすぎず，洗練・拡張が必要なことは言うまでもない．とくに，連体節主節化に特化した部分の一般化が必要である．これには，連用節の主節化 [5] など，他の種類の言い換えに対象を広げることが有効であると考えられる．また，オープンテストによる評価も不可欠である．

謝辞

本研究を進めるにあたり，実験環境を提供して下さいました藤田篤氏（九州工業大学）に深く感謝いたします．

参考文献

[1] 江原暉将, 福島孝博, 和田裕二, 白井克彦. 聴覚障害者向け字幕放送のためのニュース文自動短文分割. 情報処理学会自然言語処理研究会, NL-138-3, pp. 17–22, 2000.

[2] 藤田篤, 乾健太郎, 乾裕子. 名詞言い換えコーパスの作成環境. 電子情報通信学会思考と言語研究会, TL2000-32, 2000.

[3] 橋本さち恵. 日本語文生成における照応表現の選択に関する研究. 東京工業大学大学院情報理工学研究科修士論文, 2001.

[4] 乾健太郎. 文章生成. 自然言語処理 - 基礎と応用 -. 電子情報通信学会, 第 4 章, pp. 116–158, 1999.

[5] 神田慎哉. 連用節主節化に関する規則の追試と洗練. 九州工業大学情報工学部知能情報工学科学士論文, 2000.

[6] 久野すすむ. 談話の文法. 大修館書店, 1978.

[7] Kurohashi, S. and Nagao, M. Building a Japanese parsed corpus while improving the parsing system. *NL-PRS*, 1997.

[8] Mani, I., Gates, B., and Bloedorn, E. Improving Summaries by Revising Them. *ACL-99*, 1999.

[9] 益岡隆志, 田窪行則. 基礎日本語文法 (改訂版). くろしお出版, 1994.

[10] 益岡隆志. 複文. くろしお出版, 1997.

[11] 南不二男. 現代日本語の構造. 大修館書店, 1974.

[12] 難波英嗣, 奥村学. 書き換えによる抄録の読みやすさの向上. 自然言語処理研究会報告書, NL-133-8, pp. 53–60, 1999.

[13] 成田一. 英日・日英機械翻訳の実力. 言語処理学会第 6 回年次大会発表論文集, pp. 51–54, 2000.

[14] 野田尚史. 「は」と「が」. くろしお出版, 1996.

[15] 野上優, 藤田篤, 乾健太郎. 文分割による連体修飾節の言い換え. 言語処理学会第 6 回年次大会発表論文集, pp. 215–218, 2000.

[16] Reiter, E. and Dale, R. *Building natural language generation systems*. Cambridge Unversity Press, 1999.

[17] 白井諭, 池原悟, 横尾昭男, 木村淳子. 階層的認識構造に着目した日本語従属節間の係り受け解析の方法とその精度. 情報処理学会論文誌, Vol. 36, No. 10, pp. 2353–2361, 1995.