2名詞による連体修飾語の換言可能性に関する考察

宮木 衛[†] 増山 繁[†] 山本 和英[‡]
miyaki@smlab.tutkie.tut.ac.jp, masuyama@tutkie.tut.ac.jp, kazuhide.yamamoto@atr.co.jp
豊橋技術科学大学 知識情報工学系[†], ATR音声言語コミュニケーション研究所[‡]

1 はじめに

一般に，換言とは「ある表現を意味内容を保ったまま別の表現に変換する」ことである [1]．これまでに換言処理に関する研究は機械翻訳や主にテキスト自動要約のために行われてきている [2]．例えばテキスト自動要約においては，若尾ら [3]，山崎ら [4] は，文末の丁寧な表現の簡潔な表現への言い換えや，文字数の多い語句の省略形など同意の語句への言い換えを，人手で作成した語句の対応表を用いて実現している．片岡ら [5] は連体修飾表現を名詞型，動詞型，形容詞型と3分類し，これらの相互変換の一つとして動詞型から名詞型への換言処理手法を提案している．

本研究では，“N_1 の N_2” と “N_1N_2” 間の換言可能性について調べる．併せて，両者の使い分けについて考察する．二つの名詞 N_1 と N_2 が存在し，N_1 が N_2 を直接連体修飾している場合として，日本語では二つの表現 “N_1 の N_2” と “N_1N_2” がある．これに関しては，以下の可能性が考えられる．

1. “N_1 の N_2”(助詞型) との表現が一般的
2. “N_1N_2”(直結型) との表現が一般的
3. どちらの表現も一般的で意味が同じ
4. どちらの表現も一般的であるが両者の意味が違う

一般的に2名詞による連体修飾語がある場合，必ず助詞型あるいは直結型といった表現が使われる．しかし，どちらかの表現を使用する語句もあれば，場合によっては両者が使い分けられ使用されている語句も存在する．本研究では，まず両者間の換言可能性について考察する．新聞記事中の助詞型連体修飾語，直結型連体修飾語を全て抽出することによって，1. 2. 3. 4. のそれぞれの場合において仮説を立て，検証を行うことにする．

助詞型，直結型の両者間で換言可能であるとき，どちらの表現を採用するかが問題となってくる．これには頻度を用いる方法が考えられる．2つの表現が新聞記事中にどのくらい使用されているかを調べ，高頻度の表現を採用する方法である．しかし，この方法では常に高頻度の表現を使うことになってしまい，使い分けにはなっていない．そこで本研究では，使い分けに関していくつかの仮説を立て検証を行っていく．

この換言処理は，機械翻訳に有効であると考えている．例えば日英の翻訳を考えた場合，“N_1 の N_2” から “N_1N_2” に換言可能だとわかっていれば，“N_1N_2” という語が対訳辞書にあるかどうかを調べることができる．もし，辞書に “N_1N_2” があれば「の」の訳を考えなくてすむ．また逆に，“N_1N_2” が “N_1 の N_2” に換言可能だとわかっていれば，“N_1N_2” という語が対訳辞書にない場合，“N_1 の N_2” に換言してから翻訳させることができる．英日の翻訳の場合では，英語の “N_1N_2” または “N_2 of N_1” の日本語訳決定の際に，より自然な日本語訳を生成することができると考えられる．また同様の理由で，情報検索や他の自然言語処理に関連した問題にも有効であると考えている．

2 換言可能性の判定

助詞型から直結型への換言あるいは直結型から助詞型への換言が可能であるかの判断を，本手法ではテキストコーパスを用いて行う．テキストコーパスには毎日新聞記事を使用する．これらを考えたとき，次の仮説が考えられる．

仮説1
対象となる名詞句の助詞型，直結型の両方が存在するとき，両者間は常に換言可能である．
例1 使途の報告 ⇔ 使途報告 (換言可能？)

仮説2
対象となる名詞句の助詞型，直結型のどちらかしか存在しないとき，両者間は常に換言不可能である．
例2 石油会社 ⇔ 石油の会社 (換言不可能？)

なお，前節で示した4. の「どちらの表現も一般的であるが両者の意味が違う」場合のほとんどは直結型が固有名詞になる語句である．例としては「日本の大学 ⇔ 日本大学」である．しかし，例外として「人の間 ⇔ 人間」「手の本 ⇔ 手本」といったように直結型が固有名詞にならない例もある．これらのうち助詞型の表現は実際にはほとんど出現しない表現であるので，以後の議論の対象外とする．

次節では，これらの仮説の検証を行う．

2.1 検証

ここでは前節の仮説に基づいて以下の検証を行った．

1. 助詞型(あるいは直結型)のどちらか一方しか存在しない場合，換言可能であるか．

例 3
名詞句：「ミサイル 協議」，助詞型出現数：0，直結型出現数：71
ミサイルの協議 ⇔ ミサイル協議 (換言可能？)

2. 対象となる名詞句の助詞型，直結型の偏り度が低い場合，換言可能であるか．

例 4
名詞句：「各国 市場」，助詞型出現数：3，直結型出現数：2
各国の市場 ⇔ 各国市場 (換言可能？)

3. 対象となる名詞句の助詞型，直結型の偏り度が高い場合，換言可能であるか．

例 5
名詞句：「半島 情勢」，助詞型出現数：1，直結型出現数：40
半島の情勢 ⇔ 半島情勢 (換言可能？)

なお，偏り度とは対象となる名詞句が助詞型，直結型のどちらかにどのくらいの割合で出現しているかを示す指数である．例えば，助詞型の出現数が 10，直結型の出現数が 100 とした場合，偏り度は次式により 10 と算出される．

定義 1
偏り度 = 大きい方の出現度 / 小さい方の出現度

偏り度が小さいということは，名詞句の出現度が同じくらいの割合にあることを示す．逆に偏り度が大きいということは，名詞句の出現度が助詞型，直結型のどちらかに大きく偏っていることを示す．なお，どちらか一方しか出現しない場合は偏り度は算出できないものとする．

また，ここで換言可能であるかの判断は，読みやすさ，意味の変化の度合を基に筆者の主観で行った．対象となる名詞句の抽出は形態素解析結果のみで行った．形態素解析器には JUMAN を使用した．本研究では，N_1，N_2 は各々JUMAN の形態素解析で決定された名詞を指す．なお，本研究で扱う名詞としては時相名詞，形式名詞は扱わないものとする．検証対象語句は 1998 年毎日新聞記事 7 月から 12 月分から，無作為に抽出した．また一つの語句に極端に多い記事が含まれている場合，検証対象記事は無作為に抽出した 5 記事と限定した．

なお，検証対象語句数と記事数を表 1 に示す．

2.2 検証結果

検証結果を表 2 に示す．

2.3 考察

検証結果から，最初に示した仮説はすべて成立しないことがわかる．2 つの名詞句の両方が存在するとき，両者間は「常に換言可能である」とは検証結果からは言えないことがわかる．また，2 つの名詞句のどちらかしか存在しない場合でも両者間が「常に換言不可能である」とも言えない．ただ，“N_1N_2” から “N_1 の N_2” の換言の場合，専門用語である場合やその表現自体が聞き慣れている，また括弧の中の表現のために換言できないという語句もあることがわかっている．さらに，同じ語句でも，周りの文脈次第では換言可能となったり不可能となったりすることもわかっている．

表 1: 検証対象語句数と記事数

		語句数	記事数
1.	“N_1 の N_2” ⇒ “N_1N_2”	2	16
	“N_1N_2” ⇒ “N_1 の N_2”	3	23
2.	“N_1 の N_2” ⇒ “N_1N_2”	5	26
	“N_1N_2” ⇒ “N_1 の N_2”	5	29
3.	“N_1 の N_2” ⇒ “N_1N_2”	5	34
	“N_1N_2” ⇒ “N_1 の N_2”	5	25

表 2: 検証結果

		記事数	換言可能数
1.	“N_1 の N_2” ⇒ “N_1N_2”	16	0
	“N_1N_2” ⇒ “N_1 の N_2”	23	15
2.	“N_1 の N_2” ⇒ “N_1N_2”	26	15
	“N_1N_2” ⇒ “N_1 の N_2”	29	26
3.	“N_1 の N_2” ⇒ “N_1N_2”	34	19
	“N_1N_2” ⇒ “N_1 の N_2”	25	16

3 使い分けに関する考察

前節の結果より，対象となる名詞句の両者が同じくらいの頻度で出現している場合，どちらを使用しても 8 割の割合で不自然ではないことがわかった．また，両者の頻度が極端に異なる場合，“N_1 の N_2” と “N_1N_2” のうち高頻度のほうが必ずしも自然であるとも言えない．したがって，低頻度のほうをすべて高頻度のほうに置き換えても問題はないという仮説は成立しない．また，この極端な場合として，一方しか出現しない場合も多くある．このことから，“N_1 の N_2” と “N_1N_2” のうち頻度の高いものを必ずしも選択することはできない．また，両者が出現する場合も，すべて高頻度のものに置き換えても問題はないとは言えない．

では新聞記事中に出現する両者はどのような使い分けを行っているのかが疑問となる．本研究では以下の仮説を立てた．

仮説 3
単なる人間 (新聞記者) の気まぐれで，そもそも使い分け規則などない．

仮説 4
偏り度が低い場合は単なる人間の気まぐれである

可能性が高いが，偏り度が高い場合は，何らかの規則性が存在する．

仮説 5

偏り度に関わらず，使い分けには必ず何らかの規則性が存在する．

本研究では，仮説 4 について検証することにした．その理由として，例えば "N_1 の N_2" が出現する数が 5，"N_1N_2" が出現する数が 150 であった場合，ほとんどの文章中において "N_1N_2" という表現が使われるのに対して，何故 "N_1 の N_2" という表現を使わなければならなかったのかという疑問が生じたからである．これを検証することによって，その他の 2 つの仮説の解決にもつながると考えたからでもある．

3.1 検証

偏り度が高い場合の使い分けの検証を行う．ここで，検証条件について述べる．偏り度が高いとは偏り度が 30 以上のものを指すことにする．また，形態素解析誤りを排除するために，両者の出現数が 5 以上のものに限定した．1998 年毎日新聞記事 7 月から 12 月分から，この条件に当てはまる名詞句を抽出したところ，30 の名詞句が得られた．なお，名詞句の抽出方法は形態素解析の結果のみで行った．形態素解析器には JUMAN を使用した．また，"N_1N_2" の表現が固有名詞となる場合は検証対象から省いた．さらに，一つの名詞句に極端に多い記事が存在する場合，検証対象記事は無作為に抽出した 5 記事に限定した．なお，ここで得られた名詞句はすべてが "N_1N_2" の出現数が極端に多い場合の例である．

これらの表現に基づき以下の仮説を置き，その検証を行う．

仮説 6

"N_1N_2" の前後の語句との関係により使い分けの規則があるのではという仮説を置く．具体的に言うと，N_1 の直前の語句 (N_1 の直前が複合名詞の場合，その前) または N_2 の直後の語句 (N_2 の直後が複合名詞の場合，その後) との関係によって "N_1 の N_2" であるのか "N_1N_2" になるのかが決まるのではないかということである．

実際に名詞句の例を挙げて説明する．

名詞句

「保険制度」 助詞型出現数 : 5, 直結型出現数 : 226, 偏り度 : 45.2

記事内容

… 2000 年 4 月からスタートする公的介護保険の制度についてどの程度知っていますか?

この記事について「公的介護保険制度」という語句自体は他の記事を探しても出現するパターンである．しかし，「する公的介護保険制度」で調べると存在はしない．一方，「公的介護保険制度について」で調べると 1 件存在する．また，「介護」「保険」と連なっているために「保険」の直後に「の」が入っているのではないかと考え，「公的介護の保険制度」で調べたところ，存在しないことがわかっている．「保険制度」という語句が 226 件も存在するのに対し「公的介護の保険制度」になると 1 件も存在しないことがわかる．

以上をまとめると，「保険制度」という語句が出現した場合，「保険」「制度」の間に「の」を入れるべきであるかの使い分けを判断しなければならない．その判断材料として，前後の語句を確認する．前後の語句が名詞である場合は，更にその前後も見る．例の場合，「公的介護保険制度」という語句になる．「公的介護保険制度」という語句の前後から「する公的介護保険制度」「公的介護保険制度について」という語句を抽出する．その語句がコーパス中に存在しているかを調べる．なし，もしくは，少数である場合，「保険」「制度」の間に「の」を入れるべきであると判定する．しかし，「保険制度」という語句の偏り度があまりにも大きいために，ほんとうに「の」を入れてよいのか迷う．そこで，「保険」という語句の前に「の」を入れ，「公的介護の保険制度」という語句がコーパス中に存在するかを調べる．なし，もしくは，少数の場合，「保険」と「制度」との間に「の」を入れることを認定する．

検証方法は，名詞句がどのような形で出現するかで変わる．そのパターンとパターン別の検証方法を表 3 に示す．ただし，検証対象部分の前後の形態素が読点，濁点等の記号あるいは文頭，文末であった場合は，「(形態素)〜」および「〜(形態素)」の検証は省くこととする．なお，ここで形態素とは形態素解析結果ではなく，筆者が独自に判断したものとする．検証方法は表 3 に存在する語句が毎日新聞記事 1998 年 7 月から 12 月中に存在しているかを調べることである．

また，パターン別の名詞句の形と例を以下に示す．下線は助詞型ではなく，検証対象部分を表す．

パターン 1 … 複合名詞 (N_1の N_2) …

例 6 それでも 産業構造の改革 は早かった。

パターン 2 … 複合名詞 (N_1の N_2) 複合名詞 …

例 7 そのため内容は具体的な手立ての提示よりも、子供に対する時代や社会・生活環境の問題点 の総花的羅列に傾いたうらみがある。

パターン 3 … (N_1の N_2) 複合名詞 …

例 8 今後、団体や個人会員を募り、甲子園球場を核に、全国の高校 球児の古里となる街づくりを目指す。

パターン 4 … (N_1の N_2) …

例 9 勉強しなくなる 日本の学生 とはまるで反対だ。

3.2 検証結果

パターン別の検証結果を表 4 に示す．なお，ここで語句 A がコーパス中に存在していたとき A と示し，逆に $\neg A$ はその語句がコーパス中に存在しなかったことを示す．本研究では使い分けの判断ができた条件とし

表 3: パターン別検証方法

パターン 1 (検証対象記事数 : 51)
複合名詞 "N_1N_2" (= A) (形態素) 複合名詞 "N_1N_2"(形態素) (= B) (複合名詞) の "N_1N_2" (= C)
パターン 2 (検証対象記事数 : 19)
複合名詞 "N_1N_2" 複合名詞 (= A) (形態素) 複合名詞 "N_1N_2" 複合名詞 (形態素) (= B) (複合名詞) の "N_1N_2" 複合名詞 (= C) 複合名詞 "N_1N_2" の (複合名詞) (= C)
パターン 3 (検証対象記事数 : 28)
"N_1N_2" 複合名詞 (= A) (形態素)"N_1N_2" 複合名詞 (形態素) (= B) "N_1N_2" の (複合名詞) (= C)
パターン 4 (検証対象記事数 : 63)
(形態素)"N_1N_2"(形態素) (= B)

ては，パターン 1,2,3 では，$\neg A \wedge \neg C$, $A \wedge \neg B \wedge \neg C$, パターン 4 では $\neg B$ である．

表 4: 検証結果

		記事数
パターン 1	$\neg A \wedge \neg C$	26
	$\neg A \wedge C$	3
	$A \wedge \neg B \wedge \neg C$	12
	$A \wedge \neg B \wedge C$	8
パターン 2	$\neg A \wedge \neg C$	19
	$\neg A \wedge C$	0
	$A \wedge \neg B \wedge \neg C$	0
	$A \wedge \neg B \wedge C$	0
パターン 3	$\neg A \wedge \neg C$	16
	$\neg A \wedge C$	0
	$A \wedge \neg B \wedge \neg C$	3
	$A \wedge \neg B \wedge C$	4
パターン 4	$\neg B$	23

3.3 考察

検証結果から判断できることは，文字列照合しようとする語句の文字数が多い程検証パターンに照合しないことがわかる．例えばパターン 2 では「複合名詞＋"N_1N_2" ＋複合名詞」を文字列照合しているため，それに照合しない文字列の出現確率が非常に高く，使い分けができているという判断はできない．これは，パターン 1，パターン 3 でも同様のことが言える．両パターンともだいたい半分の語句で全ての検証パターンに照合しないことが検証結果からわかるが，複合名詞を含めて文字列照合を行っているため，当然の結果と言える．しかし，照合しない記事のうち，6〜7 割の割合で，前後の形態素と照合していないため，全く使い分けが判断できないという状況でもない．

また，パターン 4 では 5 割以下の割合で使い分けの判断ができていないことがわかる．これは，前後の語句の関係による情報が少ないことによるものと考える．現在は前後 1 つの形態素のみで照合を行っており，語句の関係の弱さが伺える．前後の形態素の幅を広げるなどの検証が必要であると考えている．

したがって，本研究の仮説 7 による検証では仮説 4 が証明できたとは言えず，今後更なるあるいは別の検証が必要であると考えている．

使い分けができたと認定された語句と認定されなかった語句の例について表 5 にまとめる．

表 5: 使い分けができたと認定された語句

認定された語句	認定されなかった語句
の自然環境の保全と	自由党党首の会談
生活環境の問題点	は経済の危機打開に
景気の回復傾向	や環境問題に
と人間の環境に	

4 むすび

本研究では，助詞型連体修飾語，直結型連体修飾語間の換言可能性について 2 つの仮説を立て検証した．検証の結果，両方の連体修飾語が存在するとき両者は必ずしも換言可能ではなかった．また，どちらかの連体修飾語しか存在しないとき，両者間は常に換言不可能であった．また，対象となる連体修飾語が互いに換言可能である場合の使い分けについても仮説を立て検証を行った．前後の語句による関係によって使い分けが判断できるという仮説を立て検証を行ったが，本研究の検証方法では使い分けの判断ができたとは言えず，今後に課題を残したと言える．

謝辞 言語データとして，毎日新聞 CD-ROM 版の使用を許可して頂いた毎日新聞社に深謝する．

参考文献

[1] 佐藤理史:論文表題を言い換える, 情報処理学会論文誌, Vol.40, No.7, pp. 2937-2945(1999)

[2] 山本和英 : 換言処理の現状と課題, 言語処理学会第 7 回年次大会ワークショップ論文集, pp. 93-96, (2001)

[3] 若尾孝博, 江原暉将, 白井克彦 : テレビニュース番組の字幕に見られる要約の手法, 情報処理学会研究報告 NL122-13, pp. 83-89(1997)

[4] 山崎邦子, 三上真, 増山繁, 中川聖一 : 聴覚障害者用字幕生成のための言い換えによるニュース文要約, 第 4 回年次大会講演論文集, 言語処理学会, pp. 646-649(1998)

[5] 片岡明, 増山繁, 山本和英 : 動詞型連体修飾表現の "N_1 の N_2" への言い換え, 自然言語処理, Vol.7, No.4, pp. 79-98(2000)