重文複文文型パターン辞書の統語的意味的分類に関連する調査

坂田純　村上仁一
鳥取大学大学院工学研究科
{d112004,murakami}@eecs.tottori-u.ac.jp

1　はじめに

非線形言語モデルに基づく重文複文文型パターン辞書が開発されている [1]．この辞書の文型パターンを用いた日英機械翻訳では，文構造の持つ意味を保存することで，精度の高い翻訳文を得ると期待されている．また各文型パターンには，文構造の持つ意味が統語的意味的分類情報として抽出，付与されており，文構造の持つ意味を明示的に抽出し提示することで，言語分析と機械翻訳に大いに役立つと期待されている．

統語的意味的分類情報の付与は，現在，統語的分類や節間意味分類のように，複数の項目に分けて記述されている．この分類情報を用いた，文型パターンの選択可能性に関する調査は行われているが [4]，現在の記述形式の妥当性やその有効性の調査は，その困難さによりまだ行われていない．ところで文型パターンを用いた翻訳方式がすでに実装されており，文型パターンが適合した場合は翻訳精度が高いことが明らかになっている．そこで本論文では，統語的意味的分類情報の記述形式の妥当性を検証するため，文型パターン翻訳の翻訳結果を利用して調査を行う．

2　重文複文文型パターン辞書

非線形言語モデルに基づき，重文複文文型パターン辞書が作成されている．その目的を二つに大別すると，一つは複雑な文構造を持つ重文と複文に対し，文構造の持つ意味を明示的に抽出し，言語分析や機械翻訳に利用可能とすることである．文構造の持つ意味は統語的意味的分類情報として記述されている．もう一つは，文構造の持つ意味を保存する文型パターンを作成し，この文型パターンを用いて機械翻訳を行うことである．

重文複文文型パターン辞書では，述部を 2 つか 3 つ持つ日本語文とその英訳文を対象に約 12 万文対が収集され，文型パターン化と統語的意味的分類情報の付与が施されている．文型パターン辞書では，他の要素に置き換えても文構造の持つ意味が変化しない要素を「線形要素」，置き換えにより変化してしまう要素を「非線形要素」と定義している [2]．文型パターンでは線形要素は変数に変換されており，変数化の範囲に合わせ，単語レベル，句レベル，節レベルの文型パターンが作成されている．本論文では単語レベル文型パターンを翻訳に使用する [3]．以下，文構造の持つ意味と文型パターンおよび統語的意味的分類について説明する．

2.1　文構造の持つ意味

文構造の持つ意味を「富士山は大山より高い。」と「Mt. Fuji is higher than Mt. Daisen.」の例で説明する．この対訳文対の文構造の持つ意味として，「二者比較」が考えられる．線形要素を変数に置き換えることで，この対訳文対から，「< 名詞 1> は < 名詞 2> より < 形容詞 3>。」と「< 名詞 1> is < 形容詞 3 比較級 > than < 名詞 2>.」の日英文型パターンが得られる．日本語文型パターンでは，名詞と形容詞が助詞の “は” と “より” によって，この語順で連接されることにより，「二者比較」の意味が保存されている．英語文型パターンにおいても，名詞と形容詞が be 動詞 “is” と接続詞 “than” によってこの語順で連接されることにより，「二者比較」の意味が保存されている．なお変数にはより細分された意味属性を付与しており，翻訳の際に，文型パターン適合のための制約として利用する．

文型パターン辞書では，日本語文構造の持つ意味を，統語的意味的分類情報として付与している．文構造の持つ意味は，抽象化された文構造により可視化されるため，文型パターンの作成と同時に統語的意味的分類情報の抽出を行っている．重文と複文は複数の節を持つため，一文の文構造の持つ意味を，例えば「二者比較」のような単純な表現で表すことが難しい．そのため統語的意味的分類情報は，節の意味属性や節間意味分類のように，複数の項目に分けて記述している．

2.2　文型パターンおよび統語的意味的分類の例

文型パターンは変数，記号，関数，字面で記述されており，このうち変数が線形要素，関数と字面が非線形要素にあたる．変数化する要素は，変数化可能な要素の品詞をあらかじめ決定しておいてから，各対訳文対ごとに人手で決定した．ただし日英で直接対応する要素のみを変数化したため，日英で個別に文型パターンを作成する場合に比べ，変数化した要素の数が減少している．

単語レベル文型パターンの具体例を表 1 に示す．表 1 の「統語的分類」から「節間キーワード」までの 1 セットが，この文型パターンの統語的意味的分類情報である．統語的意味的分類情報の各項目の説明は表 2 に示す．

表 1 単語レベル文型パターンの記述例

日本語文型パターン原文	彼のお母さんがああ若いとは思わなかった。
英語文型パターン原文	I never expected his mother to be so young.
日本語文型パターン	</*N*1(人間) は >/*N*2(他称単数男, 男) の/*N*3(母, 女) が/ああ/*AJ*4(属性) とは/*V*5(思考動作, 感情動作).hitei.kako。
英語文型パターン	<I|*N*1> never *V*5^past *N*2^poss *N*3 to be so *AJ*4.
統語的分類	第一節が従属節の述部二つの重文
第一節の意味属性	性状規定の表現/人の性状規定/なし/なし
第二節の意味属性	知覚と情緒の表現/個人的感情/なし/なし，その他複数
節間意味分類	補語相当節/引用節/間接引用/トハ型
節間キーワード	とは

表 2 統語的意味的分類情報の説明

項目	説明
統語的分類	節の数および重文と複文の区別を 5 種類で分類.
第一節の意味属性	節の用言による意味分類．4 階層で構成．“/” で階層を区切り，抽象度は左が高く右が低い.
第二節の意味属性	同上
節間意味分類	節間の意味的関係．4 階層で構成.
節間キーワード	節間意味分類を決定する日本語表現

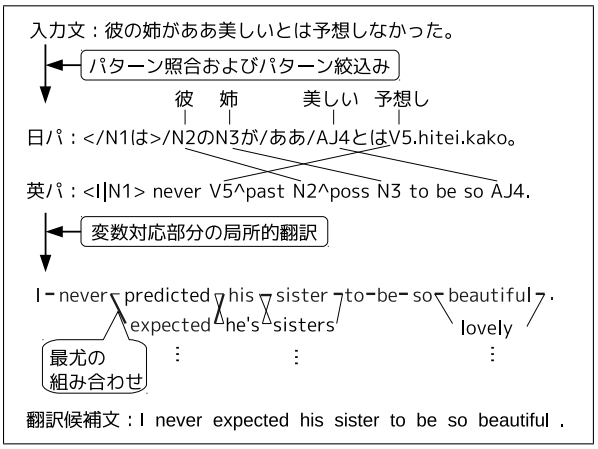

図 1 文型パターン翻訳例

2.3 単語レベル文型パターン翻訳方式

本研究で調査に利用する，単語レベル文型パターン翻訳方式について説明する．翻訳手順とその概要を以下に示す.

1. 日本語パターン検索システムを用いてパターン照合を行い，入力文に適合する日英文型パターンとその変数情報を得る.
2. 意味属性を用いて，日本語文型パターンの絞込みを行う.
3. 英文生成システムにより，日英文型パターン，単語辞書を用いて英語文を生成する．訳語選択では，tri-gram 値と単語翻訳確率を用いて，最尤の単語組み合わせを選択する.

翻訳例を図 1 に示す.

3 調査項目

本論文では文型パターン翻訳の翻訳結果を利用して，統語的意味的分類情報の記述精度や記述形式の妥当性を調査する．まず翻訳結果に対し人手評価を行い，その評価結果と統語的意味的分類情報との間に関連性があるか調べる．最初に数値的な分析を行い，その結果を元に具体例を用いて考察を行う．調査項目を以下にまとめる.

1. 英語文型パターンの適切さと翻訳精度の関係
2. 入力文と文型パターン間の，節の意味属性の一致率
3. 入力文における変数適合部分の割合
4. 具体例を用いた分析から記述形式の問題点を抽出

4 翻訳実験

文型パターン辞書の日英文型パターン約 12 万から抽出した，10 万の日英文型パターンを翻訳実験に使用する．この 10 万文型パターン対から抽出した，変数化された要素の対を用いて，単語辞書の作成と単語翻訳確率の学習を行う．単語連鎖確率（単語 tri-gram 値）の学習は，この英語文型パターン原文 10 万文を用いて行う．学習に使用しない残りの日本語文型パターン原文約 2 万文から，5,000 文を抽出しテスト文に用いる.

4.1 評価結果

適合する文型パターンが得られた入力文から，ランダムに 100 文抽出し人手評価を行った．評価基準は，入力文の意味が十分に読み取れる場合を “〇”，そうでない場合を “×” とした．評価結果は〇が 48 文で×が 52 文であった.

5 分析

5.1 英語文型パターンの適切な割合

まず翻訳に使用された英語文型パターンが，その入力文の翻訳に対し適切であったか調査する．適切な英訳文を生成可能と判断できるときを “適切”，そうでないときを “不適切” とした．適切な場合は，日英間で文構造の変換が成功しているとみなせる.

人手評価を行った 100 文中，適切な文は 75 文，不適切な文は 25 文であった．この 25 文中，形態素解析エラーや文型パターン記述ミス等による文は 18 文であっ

た．これら 18 文を除き，残り 82 文を調査の対象にする．英語文型パターンの適切さと翻訳精度の関係を，表 3 に示す．

英語文型パターンが適切で翻訳精度の低い 27 文は，主に未知語と不適切な訳語選択による．しかし未知語と訳語選択の問題を，文型パターンの問題と明確に区別することは難しい．次節からは，以下の 3 通りに分類して調査を行う．

(a) 文型パターンが適切かつ翻訳精度が高い:48 文
(b) 文型パターンは適切だが翻訳精度が低い:27 文
(c) 文型パターンが不適切なため翻訳精度が低い:7 文

表 3 英語文型パターンの適切さと翻訳精度の関係

	翻訳○	翻訳×
適切	48	27
不適切	0	7

5.2 節の意味属性の一致率

文型パターンが適合した時点で，統語的意味的分類情報のうち統語的分類，節間意味分類，節間キーワードの 3 つは，入力文と文型パターンの間でほぼ一致する．そこで節の意味属性に対し，入力文と文型パターンの間で一致する割合を調べた．表 4 に調査結果を示す．表 4 では，全一致は全ての節の意味分類が一致した文の数，部分一致は少なくとも一つの節が一致した文数，全不一致は一致した節がない文数，一致率は全一致節数/全節数を示している．なお統語的分類に記述ミスがあった文が 4 文あり，これらは以後分析から除く．

表 4 において，(c) ではほとんどが全不一致であり，逆に (a) では全不一致は少数しかみられない．したがって節の意味属性が入力文と文型パターン間で一致するほど翻訳精度が高くなることを示している．しかし (a) においても一致率は約 60% であり，高い翻訳精度を得るのに，節の意味属性の一致が必須ではないことを示している．

表 4 節の意味属性の一致率

	全一致	部分一致	全不一致	一致率
(a)	17	21	8	0.60(58/97)
(b)	12	8	6	0.62(33/53)
(c)	0	1	5	0.08(1/12)

5.3 変数適合部分の割合

文型パターンの長所は，抽象化された文構造でありながら元の文構造の持つ意味を保存できることである．しかし日英で直接対応の取れる要素を変数化したため，日英で個別に変数化した場合に比べ，変数化要素の割合が低下している．変数化要素の割合が低下するほど，文型パターンの保存する文構造の持つ意味は抽象度を失っていき，文構造を抽出することの効果が減退していく．したがって文型パターン翻訳方式は，文型パターンの変数の割合が低下するごとに，翻訳メモリを用いた翻訳方法に近似していくことになる．ただし日英対訳文対間での要素の対応を，自動で適切にとることは難しく，人手で対応を調べて変数化を行った文型パターンの方が翻訳精度は高くなると考えられる．

表 5 に，入力文における変数に適合した形態素の割合を示す．表 5 中の数は数値 % 以上 ~ 数値 % 未満の文数を示している．表 5 において，(a) と (b) の間で大きな差は見られない．また変数化要素の極端に少ない文型パターン (~20%) を使用したのはわずか 5 文であり，文構造の持つ意味を保存することの翻訳への効果があると考えられる．

表 5 入力文における変数適合部分の割合とその内訳

%	~20	20~30	30~40	40~50	50~
(a)	4	3	13	24	1
(b)	1	3	9	9	4
(c)	0	1	2	1	3

5.4 具体例を用いた分析

本節では具体例を用いて特徴や問題点を抽出する．表 6 に，(a) における，入力文と使用文型パターン間で節が全不一致かつ変数適合部分の割合が 40% 以上の例を示す[*1]．

まず “入力文の各種情報” からわかる事項を述べる．第二節の意味属性のうち，“地域社会生活の行為” と “移動行為の表現” は日本語表現 “ため息を漏らす” の意味属性として不適切である．むしろ “知覚と情緒の表現” と “日常生活の行為” を付与すべきである．節の意味属性は用言の意味属性を元にして付与している．節の意味属性の過剰な付与または不足は，調査を行った入力文と日パターン原文の約半数 (85/156) においてみられたため，節の意味属性を付与し直す必要がある．また日パターンの動詞変数 *V*3 の意味属性のうち，“属性”，“結果”，“生成”，“感情動作”，“思考動作” は，この日本語文では不適切である．変数の意味属性は自動で付与しており，用言の変数の意味属性は多義性を考慮した付与方法を用いるべきと考えられる．

次に “使用文型パターンの各種情報” を見ると，節間意味分類の “副詞的引用節/因果関係/原因/時間的原因” が不適切である．節間意味分類は節間キーワードを元にし

[*1] 節の意味属性は 4 階層のうち抽象度の最も高い 1 階層目のみ記述している．日パターンは日本語文型パターンで英パターンは英語文型パターンを示す．入力文の各種情報における日パターン原文と英パターン原文は，入力文と参照文と同じため省略．

て付与されており，節間意味分類の過剰な付与または不足は，約 20%(33/156) においてみられた．節の意味属性に比べて少ないが，こちらも付与し直す必要がある．以上より現在の付与精度では，統語的意味的分類の翻訳への利用は困難と考えられる．また複数項目への分割や 4 階層の記述により，統語的意味的分類を用いて言語分析を行うには熟練を要する可能性が高い．言語分析への利用には，複数の項目をまとめる新たな記述項目の追加が必要と思われる．

最後に，入力文と使用文型パターンの各種情報を比較し，考察を行う．入力文と日パターン原文では，言明されている事象が明らかに異なっている．そこで表 6 の節の意味属性と統語的意味的分類を用いて，人手で文全体の文構造の持つ意味を抽出すると，一例として表 7 の記述が得られる．表 7 における節の意味属性と節間意味分類は，表 6 から適切と判断できる記述を選択して用いている．表 7 より，入力文と日パターン原文共に時間的経過を表現する文構造を持ち (“一節の行為の後に二節の行為”)，そのためこの文型パターンにより高い翻訳精度の文が得られている．以上より，適切な節の意味属性と節間意味分類を付与できれば，現在の統語的意味的分類の記述形式のまま，文型パターンの選択等に利用可能と考えられる．また表 7 の例は，統語的意味的分類の新たな記述項目の表現形式を示唆するが，詳細な検討が必要なため，稿を改めて記述する．

6 今後の調査予定

今後調査すべき項目を以下にまとめる．

1. 節の意味属性と文型パターンの節構造の持つ意味との関係
2. 複文の具体例を用いた分析
3. 日本語表現と英語表現との関係

上記 1 は文型パターンが保存する文構造の持つ意味の調査に必要となる．表 6 の考察は重文に対して行われており，複文に対しても分析が必要である (上記 2)．本論文での分析と考察は，日本語の統語的意味的分類のみを対象としており，翻訳への効果を検証するために，3 の調査が必要である．

7 おわりに

本論文では，重文複文文型パターン辞書の統語的意味的分類に対し，文型パターン翻訳の翻訳結果を用いて記述形式の検討を行った．分析結果から，統語的意味的分類の不適切な付与が多数存在することが明らかになった．考察から適切な統語的意味的分類を付与すれば，現在の記述形式で文型パターン翻訳に利用可能であることが示唆された．しかし言語分析への利用では，新たな記述項目の追加が必要と示唆された．今後は残された調査課題と共に，統語的意味的分類の新たな記述項目の調査を行う予定である．

表 6 具体例

入力文	母は家計簿を見てため息を漏らした。
参照文	As my mother looked over the family finances, she sighed.
出力文	Mother looked household account book and breathed a sigh .
人手評価	○
入力文の各種情報	
日パターン	/$1^{/*N*1(女, 母) は }/*N*2(文書類, 帳) を/*V*3(属性, 結果, 身体動作, 生成, 知覚動作, 感情動作, 思考動作)(て—で)$1/*N*4(感嘆, 呼吸) を/漏らした。
英パターン	As *N*1 *V*3ˆpast over *N*2, *N*1ˆpron *V*4ˆpast.
統語的分類	第一節が従属節の述部二つの重文
一節意味属性	知覚と情緒の表現, 知的な行為の表現, 日常生活の行為
二節意味属性	知的な行為の表現, 地域社会生活の行為, 現象事象の表現, 移動行為の表現
節間意味分類	副詞的引用節/因果関係/原因/時間的原因, 並列節/順接的並列/総記の並列/なし
節間キーワード	て
使用文型パターンの各種情報	
日パターン	/</*N*1(人間) は >/*N*2(作物, 野菜) を/*V*3(属性変化, 身体動作)(て | で)</*N*4 を >*V*5(状態, 身体動作)#6(.genzai|.kako)。
英パターン	<You|N1> *V*3 *N*2 and *V*5#6(ˆpresent|ˆpast) <them|*N*4ˆobj>.
変数情報	*N*1=母,*N*2=家 計 簿,*V*3=見 (見 る),*N*4=た め 息,*V*5=漏らし (漏らす)
日パターン原文	野菜を煮て食べる。
英パターン原文	You boil the vegetables and eat them.
統語的分類	第一節が従属節の述部二つの重文
一節意味属性	日常生活の行為
二節意味属性	日常生活の行為
節間意味分類	入力文と同じ
節間キーワード	て

表 7 文構造の持つ意味の抽出例

入力文の統語的意味的分類	
一節意味属性	日常生活の行為
二節意味属性	日常生活の行為，知覚と情緒の表現
節間意味分類 1	因果関係
文構造の意味 1	一節の行為 (原因) により二節の行為および表現 (結果)
節間意味分類 2	順接的並列
文構造の意味 2	一節の行為の後に二節の行為
日パターン原文の統語的意味的分類	
一節意味属性	日常生活の行為
二節意味属性	日常生活の行為
節間意味分類	順接的並列
文構造の意味	一節の行為の後に二節の行為

参考文献

[1] 池原悟. 鳥バンク (Tori-Bank). http://unicorn/toribank). 2007.

[2] 池原悟, 阿部さつき, 徳久雅人, 村上仁一. 非線形な表現構造に着目した重文と復文の日英文型パターン化. 自然言語処理, Vol.11, No.3, pp.69-95, 2004.

[3] 坂田純, 徳久雅人, 村上仁一, 村田真樹. 意味的等価変換方式に基づく単語レベルパターン翻訳方式の評価. 言語処理学会第 20 回年次大会発表論文集, pp.298-301, 2014.

[4] 中村聡, 村上仁一, 徳久雅人, 池原悟. 「意味」による文型パターン検索方式の最適化. 言語処理学会第 12 回年次大会発表論文集, pp.632-635, 2016.