科学技術振興機構(JST)戦略的基礎研究推進事業(CREST)
研究領域「高度メディア社会の生活情報技術」
セマンティック・タイポロジーによる言語の等価変換と生成
平成 15 年度研究状況報告会(30 Mar. 2004)

重文･複文の基本文型に対する文型パターン辞書のカバー率

徳久雅人（鳥取大学）
tokuhisa@ike.tottori-u.ac.jp

1. はじめに

等価的類推思考の原理に基づく機械翻訳プロジェクトでは，日本語の重文・複文の表現を対象とした文型パターン辞書の構築を行っている．現在，日英 15 万文対を収録した文対応の対訳コーパス（SEM コーパス[*1] と呼ぶ）を対象として文型パターン辞書の構築を行ったところ，3 レベルの階層的なパターン辞書が構築された．その規模は 22 万パターン対になっている．

次のステップでは，特定の重要なパターンに対するブラッシュアップや，意味類型化に備え，SEM コーパスに含まれる重文・複文構造について種類や頻度を把握しておく必要がある．

そこで，本稿では，基礎日本語文法（第 IV 部 複文）[1]に示された重文・複文の構造（90 種）を「益岡田窪分類」と呼び，益岡田窪分類と SEM コーパスの対応関係を以下の観点から分析する．

(1)一般性：コーパス中の文が, 基本的な従属節で構成されている割合

(2)網羅性：基本的な従属節がコーパスに現れる割合

(3)出現頻度：コーパスの中で使われる基本的な従属節の分布

これらを求めるために，まず，益岡田窪分類の従属節の構造を「従属節パターン」として記述する．次に, コーパスとパターン照合して従属節を抽出し，上述の点について調査する．

2. 重文･複文の基本的な構造

2.1. 基本的な構造の定義

益岡田窪分類[1]では，重文・複文の構造を主節との関係から４つに大別している．

１．補足節： 述語を補う働きをする節で，補足語と同様に格助詞または引用の形式を伴う

２．副詞節： 述語を修飾したり，文全体を修飾したりし，主節と従属節の関係が注目される

３．名詞修飾節： 名詞を修飾し，被修飾名詞と名詞修飾節との関係が注目される

４．並列節： 主節に対して対等に並ぶ関係で結びつく節

これらの分類はさらに下位分類がなされ，合計でのべ 90 個の解釈が存在する．

2.2. 従属節のパターン化

重文・複文の各解釈についての表現の特徴は，主として従属節にある．たとえば次の例文は《同時》の解釈ができる副詞節の文である．

（例文） 私が１６だった時、彼女は７つだった。

この例文での従属節の特徴は次のようにパターンで記述できる．

（従属節パターン） */CL1*(時｜際)[に]/cl/

従属節パターンの定義において，次の点を考慮することで正確に従属節が抽出できる．

(a) 節変数 *CL* の定義に様相表現を含める

(b) 節変数 *CL* の定義にダ文判定詞を適切に含める

(c) 従属節に後続する表現を型付き離散記号（または，ダミー変数）で記述する

こうして，従属節パターンは，90 個の解釈に対して 97 個を定義した[2]．

関連研究として，[3]は，あえて構文情報を使わない手法として，形態素情報に基づく正規表現によるルールで節境界を判別する方法を提案している．本手法は，文全体の構文情報は使わずに，局所的な構文情報を用いることで従属節の抽出精度の向上を

*1Semantically Equivalent Mapping

狙っている．

3. 従属節の抽出実験

3.1. 実験方法

実験の対象は，SEM コーパスにおける日本語文 126,203 文である．

抽出の方法は，2.章で作成した 97 個の従属節パターンを上述の文と照合することによる．照合には [4]の文型パターンパーサを用いた．従属節パターンが文に適合すると，適合内容，および，パターンに付随する解釈を出力する．

なお，ダ文における判定詞「の」,「に」,「で」は，誤った適合が多いため, 本稿では抽出対象にしない．

3.2. 抽出の様子

図１に実験過程でマッチした事例の一部を示す．（文１）に対して２つの従属節パターンが適合している．（適合 1-1）は,《条件》と解釈されている．（適合 1-2）は,《引用》として解釈されている．この場合,（適合 1-2）が正しい解釈である．適合内容のうち正しい解釈が含まれるので,（文１）からの従属節の抽出は成功とする．（文２）および（文３）も同様に成功した事例である．

（文１）僕は子供の頃サンタクロースは本当に北極から来るものだと固く信じていた。	
（適合 1-1） /*CL1* と[、]/cl/	《副詞節・条件》
（適合 1-2） /*CL1* と[、]/	《補足節・間接引用》
（文２）あなたの言うことは正しい。	
（適合 2-1） /*CL1* こと/n/	《補足節・コト型》
（適合 2-2） /*CL1*^rentai(こと|の)/	《名詞修飾・内容》
（文３）車を止めてエンジンを切りなさい。	
（適合 3-1） /*CL1*.te/cl/	《副詞節・因果》
（適合 3-2） /*CL*^genzai.te/cl/	《副・付帯状況》
（適合 3-3） /*CLV*^genzai.te/cl/	《並列節・総記》

図１：従属節パターンのマッチした例

次に，従属節パターンがマッチしなかった事例の一部を図２に示す．（文４）では「そのようなことをするには」という部分が抽出できなければならなかったが，益岡田窪分類には該当する構造が無かったため抽出できなかった．この節を抽出するには，「節＋には」という新しい従属節パターンが必要である．（文５）は形態素解析の想定違いにより，作成した従属節パターンでは抽出できなかった場合である．（文６）は,「ＡとＢが同じ」というダ文の従属節があると判断するか,「ＡとＢが目立つ」という主節があると判断するかによる．ここでは，後者の立場でみることが妥当と判断した．

（文４）そのようなことをするには狡猾さが必要だ。
・「節＋には」のパターンが登録されていない
（文５）運が尽きてからでは遅い。
・「てから」が１つの形態素として解析されていた
（文６）ドイツ人と日本人観光客が同じくらい目立っていた。
・単文である

図２：従属節パターンのマッチしなかった例

3.3. 結果

従属節パターンの適合した文は，122,264 文であった．非適合の文は，3,939 文であった．適合，非適合の事例からそれぞれ 50 個をランダム検査する．

適合事例については，正しい解釈を含むならば正解とみなす．この正解率を「含有正解率」と呼ぶことにする．含有正解率は 100%であった．

非適合事例については，次の３つに大別できた．

- 新しい従属節パターンが必要：　38%
- 従属節パターンのマッチに失敗：　32%
- 単文とみなすほうが妥当：　30%

4. カバー率

4.1. 一般性

一般性を次の式で求める．

$$\frac{〈\text{適合事例数}〉\times〈\text{含有正解率}〉}{〈\text{総文数}〉}=\frac{122,264\times100}{126,203}=96.9(\%)$$

したがって，SEM コーパス，および，文型パターン辞書は，基本的な従属節で構成されていることがわかる．

4.2. 網羅性

網羅性を次の式で求める．

$$\frac{〈\text{適合した従属節パターンの種類数}〉}{〈\text{全ての従属節パターン種類数}〉}\times100=\frac{96}{97}\times100=99.0(\%)$$

したがって，SEM コーパス，および，文型パターン辞書は，基本的な表現を網羅していることがわかる．

4.3. 出現頻度

図３に従属節パターンの出現頻度の分布を示す．縦軸は出現回数，横軸は従属節パターンであり，出現回数の降順で並べている．第 10 位までで全体の 73.4%，第 20 位までで全体の 86.5%を占める．

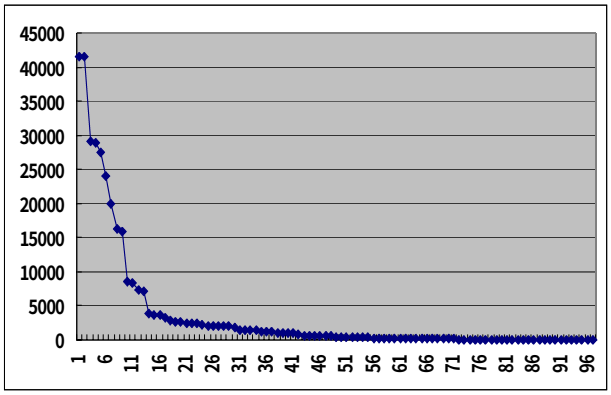

図３：従属節パターンの出現頻度

第 20 位まで，および，下位 10 位の詳細を表１に示す．「連体修飾節」，「テ型節」，「ト型節」，「連用節」，「形式名詞の修飾節」は，重点的にパターンを作り込む必要があるといえる．

また，益岡田窪分類の項目の観点からまとめた，出現頻度を表２に示す．表１のパターン別にみると連体修飾に次いで「テ型節」が重要であると言えたが，大分類で見ると，副詞節が重要であるといえる．これは，副詞節のパターンにバリエーションが多く（61 種），パターン別に見たのでは頻度が下がったためである．

4.4. 同形異義

同形異義パターンは，主なものとして５つがある．

- **P1 連体節＋名詞**：S1《補足語修飾節》，S2《内容節》
- **P2 節＋て**：S1《原因》，S2《総記》，S3《付帯状況》
- **P3 節＋と**：S1《引用》，S2《条件》
- **P4 仮定節＋ば**：S1《条件》，S2《累加》
- **P5 連用節＋ながら**：S1《付帯状況》，S2《逆接》

各パターンの適合部分を 20 個のサンプル検査により，どちらの解釈が多いのか調べる．

他に分類したものは，別の解釈のほうが望ましい場合，別のパターンがテストのパターンを含んでいる場合，従属節の判定が誤っている場合，など解釈のカテゴリに含まれない状況を全て含む．

表１：従属節パターンの出現頻度の詳細（一部）

順位	パターン概形	簡易解釈	頻度
1	*CL*^rentai *N*	修飾	41,554
2	*CL*^rentai *N*	内容	41,554
3	*CL* て～	原因	29,194
4	*CL*^genzai て～	総記	28,989
5	*CLV*^genzai て～	付帯状況	27,474
6	*CL*^rentai (こと|の)	内容	23,995
7	*CL* と	引用	19,936
8	*CL* と～	条件	16,235
9	*CL*^renyou	総記	15,812
10	*CL* の *J*	の型	8,480
11	*CL* こと *J*	こと型	8,274
12	*CL* よう	引用	7,286
13	*CL* が～	逆接	7,154
14	*CL* ので～	導出	3,811
15	*CL* ば～	条件	3,590
16	*CL*^katei ば～	累加	3,568
17	*CL* ように	引用	3,202
18	*CL* ても～	譲歩	2,761
19	もし *CL* ば *CL* ところだ	反事実	2,730
20	もし *CL* ば *CL* のに	反事実	2,730
87	*CL* だけに～	一般導出	27
88	*CL*^genzai た上で～	前提動作	19
89	*CL*^genzai た程～	程度	19
90	*CL*^genzai た(とすれば|とし たら|とすると)～	仮想的	14
91	*CL* の(は|が)*NP J* だ	強調	11
92	*CL*(一方|反面)～	対比	9
93	*CL* くせに～	非難	9
94	*CL*^genzai 割に～	程度違い	6
95	*CL*^genzai たくらい～	例示	2
96	*CL* かというの *J*	内容節	1
97	*CL*^genzai た割に～	程度違い	0

表２：重文・複文分類の大分類ごとの出現頻度

大分類名	出現割合	出現回数
補足節	15.6%	(53,744)
副詞節	35.4%	(122,216)
名詞修飾節	32.4%	(111,635)
並列節	16.6%	(57,386)

表３：従属節パターンにおける多義の比率

	P1	P2	P3	P4	P5
S1	5	14	6	15	18
S2	9	1	3	3	0
S3	-	2	-	-	-
他	6	3	11	2	2
計	20	20	20	20	20

従属節の解釈における多義解消には，従属節と主節の用言意味属性でカバーできる問題なのか，従属節の事象の結果と主節の事象の前提条件の整合性な

ど世界知識を要する問題なのか, 検討が必要である.

5. おわりに

文型パターン辞書の重文・複文に対するカバー率を決めるものは，パターン作成の原文のコーパスである．そこで，SEM コーパスにおいて基礎日本語文法に示された重文・複文の構造（益岡田窪分類）がどれだけ対応しているかについて調査した．基本的な従属節で構成されている文の割合（一般性），基本的な従属節がコーパスに現れる割合（網羅性），および，基本的な従属節の種類ごとの出現頻度を求めたところ，SEM コーパスは基本的な重文・複文をカバーしていることが確認できた．これにより，「連体修飾節」,「テ型節」,「ト型節」について重点的なパターンの精度向上が必要であるといえる．また副詞節全般にも注意を払う必要があるといえる．

今後の課題は，従属節の解釈多義の選択である．

参考文献

[1] 益岡隆志，田窪行則：基礎日本語文法，くろしお出版，1989.

[2] 徳久雅人：基礎的重文複文構造の含有率についてのトップダウン調査---中間報告，知識ベース班研究会議資料，2004.

[3] 丸山岳彦，柏岡秀紀，熊野正，田中英輝：節境界自動検出ルールの作成と評価，言語処理学会第９回年次大会発表論文集，pp.517-520，2003.

[4] 徳久雅人，村上仁一，池原悟：文型パターンパーサの試作，言語処理学会第１０回年次大会発表論文集，pp.608-611，2004.