概要

日英機械翻訳において, 日本語を線形的に翻訳することは効果的な手段の一つである. 今回は, 日本語名詞句と対応する英語表現との比較検証した. 具体的には, 句レベルの日英文型パターン集 (約 8.8 万件) からその格助詞で構成されている名詞句を抜きだし, 日本語表現と対訳の英語表現とを比較し, 日本語表現 1 つに対する英語表現のパターンを分類した. 名詞句を構成する格助詞 (相当語を含む) は 47 個である. また, 統語構造だけでは, 随伴格を伴うためにパターン抽出が難しい格助詞「と」に関しては, 名詞句を構成する格助詞として扱っていない.

句レベルパターン集には, 名詞句を構成する格助詞は計 29 種類使われていた. また, 先行研究が十分になされている「A の B」型の名詞句は, 意味属性の共起関係や特定の字面を考慮することで明らかにされている. 従って, 今回は「の」のみで構成される名詞句は対象外としている.

得られた日本語構造に対する英語パターンは, 複数個の格助詞で構成される名詞句においてパターン対を支える事例数が少ないために, パターン対を作成し, 汎用的な規則を作成することは困難であった. 格助詞が 1 つで構成される名詞句について, 別の対訳標本でも同様の頻度であるかを調べるために日本語基礎文型 8,010 文に対して, 英語構造の違いを検証した. 同一の英語構造は全体で 160/549 個 (39%) であった.

目 次

1　はじめに

等価的類推思考の原理に基づく機械翻訳の手法 [1] では, 入力文を文型パターンを用いて解析し, 非線形要素と線形要素との区別を行う. 線形要素として取り出された名詞句は, その名詞句の構成要素から英訳が生成されなければならない.

そこで, 本研究では, 線形要素として取り出された名詞句の構造について, 英訳構造の選択の観点から分析することを目的とする. 先行研究において,「A の B」型名詞句の英訳構造の選択は, 名詞 A と名詞 B の意味属性を制約条件に用いた訳出規則で実現できることが示された [2]. 本研究では, より多くの格助詞を含む名詞句の場合に対処するため, 前段階として日本語構造と英語構造とのパターン対を比較する. 3. 自己組織化マップ
組織化マップは、W の異なった組織として生成、または元々の接続のわずかな修正から生成するのもある。

2　名詞句の収集

[1] の実現に向け, 日英文型パターン辞書が構築されている. 句レベルのパターン集には, 線形要素となる日本語名詞句が変数化されて記載されている.

以下に対訳の例を示す.

AC000004-00 1 110
彼のお母さんがああ若いとは思わなかった
I never expected his mother to be so young.
LJ098578:彼のお母さんがああ若いとは思わなかった.
LE045139:I never expected his mother to be so young.
WJ027379:<N1 は>/N2 の/N3 が/ああ/AJ4 とは/V5.hitei.kako.
WE027892:(N1|I) never V5.past N2.poss N3 to be so AJ4.
PJ013130:<N1 は>/NP2 が/ああ/AJ3 とは/V4.hitei.kako.
PE021860:(N1|I) never V4.past NP2 to be so AJ3.
LJ=原文
LE=対訳
WJ=日本語の単語レベル
WE=英語の単語レベル
PJ=日本語の句レベル
PJ=英語の句レベル

2.1 格助詞の個数ごとの名詞句と分布

そこでパターンとその原文より, 線形要素となる日本語名詞句を抽出した. 格助詞を含むものは,19,399 個であった. 構成する格助詞の個数ごとに名詞句を分類すると表 1 になった.

表 1: 格助詞の個数ごとの名詞句と数

格助詞の個数		1 個	2 個	3 個	4 個
名詞句の個数		17583	1645	157	14
内訳	「の」のみ	16528	1190	71	2
	それ以外	1055	455	86	12

- **格助詞が 1 個の例**
  名詞句　来月 で80 歳　　: 年 に 1 回
  対訳　　80 next month　: once a year

- **格助詞が 3 個の例**
  名詞句　日本 での 貴社 の 製品 の 販売
  対訳　　sale of your products in Japan

2.2 格助詞について

名詞句を構成することが出来る格助詞は計 47 個 (相当語を含む). そのうち, 句レベルパターン集には, 計 29 種類の格助詞が使われていた. なお, 手法 [1] において「と」は, 随伴格で用いられる事もあり, 統語構造だけでは判別が困難なため名詞句を構成する格助詞とはみなしていない.

格助詞「の」は,1 個の格助詞で構成される名詞句全体の 93.9%を占めた. 残りは, 「に」(14.7%), 「への」(10.4%), 「や」(9.67%), 「での」(8.72%), 「からの」(7.88%), 「に関する」(6.45%), 「に対する」(6.21%), 「で」(4.66%), 「か」(3.34%), 「における」(0.24%) などで構成されていた.

そこで本研究では,1 個の格助詞で構成される名詞句については,29 種類の格助詞の中から, 頻度の高い上位 10 種類 (計 633 個) について分析する. また, 複数個の格助詞で構成される場合は, 全ての種類の格助詞の組み合わせを調べる. なお, 先行研究が十分になされている「の」のみを含む名詞句は, 格助詞の数に関係なく対象外とする.

2.2.1 出現頻度の高い格助詞

47 種類の格助詞のうち, 出現頻度の高い格助詞を表 2 示す.

表 2: 1 つの格助詞で構成される名詞句の分布表

順位	格助詞	数
1	A の B(対象外)	16715
2	A に B	123
3	A への B	87
4	A や B	81
5	A での B	73
6	A からの B	66
7	A に関する B	54
8	A に対する B	54
9	A で B	39
10	A か B	31
11	A における B	27
12	A は B	23
13	A のような B	21
14	A かの B	18
15	A から B	18
16	A としての B	16
17	A にとって B	16
18	A などの B	13
19	A までの B	12
20	A ほどの B	6
21	A ほど B	5
22	A のように B	3
23	A まで B	3
24	A よりの B	3
25	A のみの B	3
26	A にわたる B	2
27	A もの B	2
28	A をめぐる B	2
29	A だけ B	1
30	A くらいの B	0
31	A ぐらいの B	0
32	A だか B	0
33	A だの B	0
34	A という B	0
35	A といった B	0
36	A として B	0
37	A についての B	0
38	A につき B	0
39	A にとっての B	0
40	A に於ける B	0
41	A だけの B	0
42	A みたいな B	0
43	A やの B	0
44	A ようという B	0
45	A など B	0
46	A によゆ B	0
47	A へ B	0

3　日英のパターン対の作成

日本語名詞句に対応する英語構造を明確にするため, 日英で対応する単語を変数 A,B に置き換え,「名詞句のパターン」を作成する.
以下に具体例を示す.

日本文:彼女からの手紙を首を長くして待っている.
英訳:I am anxiously expecting a letter from her.
字面対　　彼女 からの 手紙 ↔ a letter from her
パターン対　A からの B　　↔ B from A

3.1 名詞句パターン対の対応付けの種類

日本語名詞句を, 構成する格助詞が 1 個の場合と複数個の場合とに分類する.

3.1.1 格助詞が 1 つ日本語パターンに対する英語パターン

使用頻度の高い 10 個の格助詞に対する対応付けの結果を表 3 に示す. 対応する英語構造には出現頻度の高い上位 3 つを示す.

表 3: 日本語構造に対応する英語構造の種類と割合

日本語構造	英語構造			
	1 位	2 位	3 位	その他
A に B (123 個)	B to A (13%)	AB (12.2%)	B in A (11.4%)	その他 (64.2%)
A への B (87 個)	B to A (51.7%)	B for A (17.2%)	B in A (5.75%)	その他 (25.3%)
A や B (81 個)	A and B (82.7%)	A or B (13.6%)	A,B (2.5%)	その他 (1.2%)
A での B (73 個)	B in A (41.4%)	AB (30.1%)	B at A (12.3%)	その他 (16.4%)
A からの B (66 個)	B from A (65.2%)	AB (16.7%)	B of A (3.03%)	その他 (9.68%)
A に関する B (54 個)	B of A (27.8%)	B about A (24.1%)	B on A (22.2%)	その他 (25.9%)
A に対する B (52 個)	B for A (26.8%)	B to A (23.2%)	AB (21.4%)	その他 (28.6%)
A で B (39 個)	B in A (26.7%)	B of A (16%)	B on A (16%)	その他 (41.4%)
何か B (31 個)	some(any) B (58.3%)	some(any)-thing B (37.5%)	冠詞 B (4.17%)	なし
A における B (27 個)	B in A (75%)	B at A (10%)	B on A (5%)	その他 (10%)

3.2 複数個の格助詞を含む場合

格助詞が複数個の場合の詳細を示す.

3.2.1 格助詞が 2 つの日本語パターンに対する英語パターン

格助詞が 2 個の名詞句は 455 個で, 日本語パターンは「89 パターン+慣用表現」の合わせて 90 パターンが存在した. また, 英語パターンに関しては 235 パターンであった.

日本語構造と出現頻度, 日本語構造に対応する英語パターン数を表 4 に示す.

表 4: 格助詞 2 つの分布表

日本語構造	個数	%	英語パターン数
A における B の C	36	7.6	19
A や B の C	34	7.2	23
A への B の C	25	5.3	16
A の B への C	22	4.6	14
A での B の C	18	3.8	12
A に B の C	18	3.8	18
A の B での C	16	3.4	15
A の B における C	14	3.0	10
A に対する B の C	14	3.0	11
A で B の C	14	3.0	8
A についての B の C	13	2.7	10
A から B の C	12	2.5	10
A の B に対する C	12	2.5	11
A の B に C	11	2.3	10
A から B までの C	10	2.1	8
A からの B の C	10	2.1	6
A に関する B の C	10	2.1	9
A としての B の C	8	1.7	4
A の B についての C	8	1.7	7
A の B である C	7	1.5	7
A の B による C	7	1.5	7
A の B からの C	6	1.3	6
A の B で C	6	1.3	6
A の B という C	6	1.3	5
A か B の C	5	1.1	5
A の B としての C	5	1.1	4
A の B や C	5	1.1	3
慣用	3	0.6	-
その他	101	21	-

英語パターンを表 5 に示す.
A's は A の所有格を,
A(adj) は日本語では名詞だが, 英語表現では形容詞を,
AB(w) は日本語の「名詞 A+名詞 B」の意味を, 英語表現で 1 単語で表していることを意味している.

表 5: 格助詞 2 つの英語表現

英語	個数	%
C of B in A	20	4.4
B's C in A	12	2.6
B's C to A	11	2.4
C of A and B	11	2.4
A's B(adj) C	8	1.6
C of A B	8	1.6
A's C to B	7	1.5
B's C of A	7	1.5
AB(w) C	6	1.3
B(adj) C for A	6	1.3
B(adj) C of A	5	1.1
C from A to B	5	1.1
C in A B	5	1.1
C in B of A	5	1.1
C of A's B	5	1.1
C of B to A	5	1.1

3.2.2 格助詞が 3 個, 4 個の日本語パターンに対する英語パターン

格助詞が 3 個の場合は名詞句が 86 個存在した. 日本語パターンは「55 パターン」, 英語パターンは「72 パターン」存在した. 1 つの日本語パターンに対する英語パターンは複数存在したが, パターン対を支える事例数が少ないため, パターン対は汎用的な規則ではない. また, 格助詞が 4 個の場合も同様である. 格助詞 3 つの場合の日本語表現の頻度を表 6 に示す.

表 6: 格助詞 3 つの場合の日本語表現の頻度

日本語	個数	%	英語パターン数
A の B や C の D	5	6	5
A における B の C の D	4	4.7	4
A の B に関する C の D	4	4.7	4
A の B への C の D	4	4.7	4
A の B で C の D	3	3.4	3
A の B に対する C の D	3	3.4	2
その他	70	81	70

4　異なる対訳標本に対する検討

1 つの格助詞で構成された名詞句について, 作成した名詞句パターン対が, 別の母集団においても同様の頻度で出現するかを調べる. 本調査は, 異なる母集団を日本語基本文型 [3] における 8,010 文とした.

日本語パターンが適合したのは 549 個であった. 表 2 の英語構造が使われていた頻度を表 7 に示す.

表 7: 日本語基本文型例文における英語構造の違い

同一英語構造の種類	数	%
1 位の英語パターンと同一の英語構造	113	21
2 位の英語パターンと同一の英語構造	13	2.4
3 位の英語パターンと同一の英語構造	34	6.2
その他のパターンと同一の英語構造	54	9.8
単語パターンと同一の英御語構造が存在せず	389	71

4.0.3　結果

表 2 における上位 3 つの英語パターンを同一の英語構造である名詞句は 160/549 個 (29%) であった. 個別の英語パターンで見ると,「A に B」,「A で B」以外の格助詞は,44/249 個 (19%),41/198 個 (21%) で極端に低かった.「A に B」,「A で B」以外の格助詞を合わせた値は,75/102 個 (73.5%) であった.
結果の詳細を表 8 に示す.

表 8: 個別で見た英語構造の違い

格助詞	同一英語構造の種類	数	%
A に B	1 位の英語パターンと同一の英語構造	12	4.8
	2 位の英語パターンと同一の英語構造	4	1.6
	3 位の英語パターンと同一の英語構造	28	11
	その他のパターンと同一の英語構造	28	11
	英語パターンと同一の英語構造が存在せず	177	71
A で B	1 位の英語パターンと同一の英語構造	39	20
	2 位の英語パターンと同一の英語構造	0	0
	3 位の英語パターンと同一の英語構造	2	1.1
	その他の英語パターンと同一の英語構造	21	11
	英語パターンと同一の英語構造が存在せず	136	69
その他	1 位の英語パターンと同一の英語構造	62	64
	2 位の英語パターンと同一の英語構造	9	9.3
	3 位の英語パターンと同一の英語構造	4	4.1
	その他の英語パターンと同一の英語構造	3	3.1
	英語パターンと同一の英語構造が存在せず	19	20

4.0.4　異なる対訳標本から見た翻訳規則の充足性

「A に B」,「A で B」の場合は, 名詞 B がサ変名詞となる事があり, 低い値を示した.「A に B」,「A で B」以外の格助詞を合わせた値は高い.

よって,「A に B」,「A で B」以外の格助詞で構成される日本語構造は, 別の母集団においても英語パターンの頻度が高いことから, 機械翻訳の可能性がある.

5　今後の課題

本研究では, 文型に対して線形要素となる名詞句について構造の解析を行った.
今後, 意味属性を用いて名詞句のパターン対を作成していく.

参考文献

[1] 池原ほか:等価的類推思考の原理による機械翻訳方式, 信学技報告 TL2002-34 pp.7-12 (2002)

[2] 徳久ほか:意味属性の共起による「A の B」型名詞句の翻訳規則, 情報科学技術フォーラム (FIT-2003) 情報技術レターズ, Vol.2, pp.87-88 (2003)

[3] 橋本ほか:日本語基本文型, 情報処理振興事業協会 (1997)