

和語動詞に対する日英対訳用例文の収集について

池原 悟* 白井 諭* 相沢 弘**

* : NTTコミュニケーション科学研究所

** : 株式会社スパルインターナショナル

1. まえがき

日英機械翻訳において用言の意味を適切な英語に対応づける方法として、結合価パターンの方が期待され、この方法を実現するため、パターン収集に関する様々な研究が行われてきた⁽¹⁻³⁾。しかし、市販の人間用の辞書(和英辞書等)の語義と用例から人手でパターンを作成する方法では、十分な数の用例が収集できないことが問題であり、対訳コーパスからパターン対を自動学習する方法では、必要とされる対訳用例の収集が困難であることなどにより、パターン対を網羅的に収集することができなかった。

これに対して、我々は、用言の用例作成では、人間用の辞書の情報よりも、人間の持つ知識の方がより柔軟で多彩であると考えて、人手によるパターン作成用の例文の作成を試みた。その結果、日英機械翻訳で特に問題となる和語系動詞の用例が多数収集でき、それにより必要なパターン対がほぼ網羅的に作成できる見通しを得たので報告する。

2. ALTのパターン対と収集の問題

(1) パターン対記述の枠組み

日英機械翻訳システムALT-J/E⁽⁴⁾における意味辞書の構成を図1に示す。この中で、用言に対する結合価パターン対は、構文意味辞書として準備されるもので、3種の意味属性体系にまとめられた言葉(意味属性名)を使用し、以下に示す基準で作成される。パターン対の記述例を図2に示す。

①日本語と英語の結合価パターンを対にして記述する。両者は、用言、格要素(主名詞+助詞)、副詞素、様相情報から構成される。

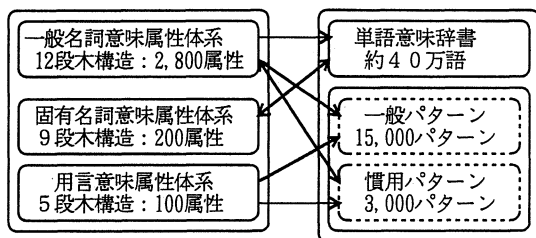


図1. ALTにおけるパターン対記述の枠組み

②主名詞は、日英の動詞が訳し分けられる最小限の深さの意味属性を用いて記述する。

③主名詞を意味属性で代表できないときは、名詞そのものを記載する。(このような格要素を持つパターンを慣用パターンと呼ぶ。)

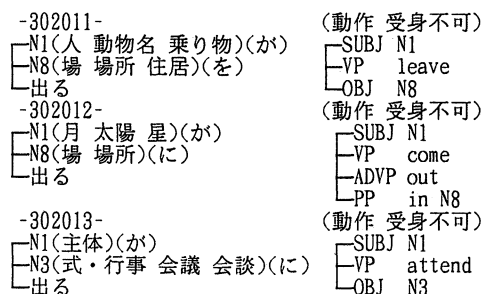


図2. パターン対記述の例

(2) パターン対収集の問題点

従来、パターン対は、和英辞書に記載された日英の用言の意味的対応関係とその例文を元に、人手で作成してきたが、翻訳実験では、不足するパターン対が続出することが問題であった。また、最近、パターン対の作成を支援するため、学習などの技術を応用して、日英対訳コーパスから半自動的に収集する方法が提案されているが、この方法でパターン対を網羅的に収集するには、単純な構文の対訳用例が大量(ある検討⁽¹⁾では、約1,000万文)に必要であることが問題であった。

これらの状況から、パターン対収集の問題を整理すると、以下の3点に集約される。

- ①日英翻訳で必要なパターン対の数が不明な点
- ②それを網羅的に収集する方法が不明な点
- ③収集作業に多大なコストが予想される点

ここで、必要なパターン対の数は(①)は、適切に選んだ用言の範囲で収集したパターン対による実験で推定できる。また、③は、網羅性のある収集方法であることが前提となる。従って、当面、②の問題の解決が急務と言える。

3. パターン対作成の方針と方法

(1) 本論文の着眼点とパターン対作成の方法

前章では、パターン対を網羅的に収集できる方法を見極めることが急務であることを述べた。

ところで、辞書に登録された動詞の用法と実際の日本文に現れた動詞の用法の関係をみると、実際の日本文には、辞書に記載されていない用法が多数現れることに気がつく。通常、人が母国語の文章を執筆する際、その意味を確認するために国語辞書を参照するのはまれであるから、人は、記述内容の思考過程や文脈などに応じて適切な言葉が自然に思い浮かべ、それを使用していることになる。この点から見て、言葉の意味的用法に関する知識は、通常の辞書よりも母国語筆者の頭の中にある方がむしろ豊富であり、辞書にはその一部が記載されているに過ぎないと推測される。すなわち、適切な刺激を与えて自分の知識を内省すれば、言葉の様々な用法が類推され、辞書に記載されている以上の種類の用例が思い浮かべられると思われるから、それを元にパターン対を作成すれば、網羅性に高いパターン対が収集できると期待される。

そこで、以下では、特に意味的用法が多彩とされる和語系動詞を対象に、人間の知識を内省することによって用例を書き出して英語訳を付与し、動詞毎の対訳用例集を作成する。その結果に基づいてパターン対を作成することとする。

(2) 用例作成の方針

動詞用例作成の方針を以下に示す。

- ①日本語中心に動詞の用法を思い浮かべ、ニュアンスの異なっているものを例文として書き出す。

<備考>実際に英語訳を想起するのは困難であるから、英語に訳すとニュアンスが違いそうだと思う範囲で考える。結果的に、訳語が同じ英語動詞になる例文が含まれても良い。

- ②一般的な表現を集めることとし、慣用表現と意識されるものは集めない。

<備考>但し、結果的に慣用表現が作成される用例となってもかまわない。

(3) 用例作成の方法

具体的には、以下の方法で用例を作成する。

- ①IPAL⁽⁶⁾の和語系動詞806語(ひらがな表記の同じ語を縮退)を対象に、IPALの用例と各種辞書を参照して、意味的用法を類推し、用例を作成する。

但し、IPALの用例と同種と判断できる用例は対象外

とする(IPAL用例が使える)。結果的に、同じパターン対が作成されるものであってもかまわない。

- ②作成する用例文は、可能な限り、「単純な名詞を格要素に持つ単文」とする。

- ③連想が可能な限り用例を書き出すが、収集する用例数の目標はおおむね、IPAL語義の2~3倍とする。

4. 対訳用例の作成収集の結果

(1) 収集目標と実績の関係

言語学に興味のある筆者の一人が担当し、前章で述べた方法によって動詞用例を作成した。収集された用例の数を表1に示す。

表1. 収集できた用例数

項目	動詞数	IPAL辞書		新規に作成した用例			用例の合計
		語義数	用例数	目標	結果	割合	
合計	806	3,375	5,243	6,750	10,497	1.56	15,740
平均	1語当り	4.2	6.5	8.4	13.0	1.56	19.5

表より、動詞当たり約13.0件の用例が作成でき、動詞当たり平均して、IPALの語義数の3倍(IPAL用例の2倍)の用例が新たに作成された。この結果、パターン対作成用として、和語動詞806語全体では、IPAL用例と合わせて1.57万件(動詞当たり約20件)の用例が準備できた。

(2) 用言の数と集まった用例の数

適当にサンプリングして抽出した23の動詞について、IPAL辞書の語義・用例と新たに作成した用例の数を比較して表2に示す。この表から以下のことが分かる。

- ①IPAL辞書では、語義数の1.5倍前後の用例が収集されているのに対して、新規作成を加えた用例は、5倍前後に増大している。

②動詞によるばらつきは大きい、この割合は、概ね、語義数の多いものほど大きい傾向がある。このことから、日本語内で語義を考えたときに比べて、英語を意識して動詞の意味を考えると、分類がより詳細になる傾向が見られること、また、その傾向は、一般に語義の多い動詞ほど強いと思われる。

(3) 用例作成の生産性

用例作成にかかった工数は約1人年であったから、1時間当たりの用例作成数は、約5件となる。

ここで、用例作成の困難さについてみるため、IPAL用例とは別に、IPAL語義の数(N件)だけの用例を新たに思い浮かべる時間をTとすると、2N件の用例を作成する時間は、おおよそ2Tであったが、3

N件の用例作成では、さらに2T(合計2T+2T=4T)の時間が必要であり、4N件の用例作成するには、さらに3~4T(合計では4T+(3~4)T=(7~8)T)の時間が必要であった。従って、最初のN件の用例の生産性は、 $5 \times (7 \sim 8) = 35 \sim 40$ 例文/時間となる。

作業全体から見た時間当たりの生産性(5例文/時間)は、決して大きな値ではないが、対象とした動詞が、和語系動詞の中でも重要な語をほぼ網羅していること、後に述べるように、辞書から収集できない意味の用法が多数網羅されていることを考えれば、全体で約1人年の工数は、結果に見合う工数だと言える。

5. パターン対作成の状況

表2で示した23の動詞を対象に、IPALの用例と新たに得られた用例を参照して、日英翻訳用の結合価パターン対を作成した。その結果を同じ表2の右の欄に示す。

(1) 対訳用例と作成されたパターン対の関係

表2から、動詞によるばらつきが大きい。概ね、2~3の用例から1つのパターンが作成されていることが分かる。また、動詞毎に見た用例数とその用例

から得られたパターン対の数の関係を見ると、作成されたパターン対の多い動詞のグループでは、パターン数の割により多くの用例が集められている傾向がある。このことは、意味的用法の多彩な動詞では、パターン対を網羅的に収集するには、より多くの用例を収集しなければならないことを窺わせる。

(2) 日本語の意味分類と英語との対応

和英辞書から作成したパターン対の数(慣用パターンを除く)は、和英辞書で定義された語義の数と考えることができる。同様、最終的に得られたパターン対の数は、英語と対比して分類された語義の数と考えることができる。そこで、和語動詞の語義の分類精度についてみると、表3から以下のことが観察される。

- ①人の知識の中にある動詞の意味の分類精度は、和英辞書の意味分類精度に比べて、2~3倍(表では2.6倍)細かい。
- ②日本語内で精密に分類された(IPAL辞書の)意味分類も、英語との対応関係を考えるときは、十分とは言えず、さらに、2倍程度(577/271=2.1倍)、細かく分類する必要がある。

表2. 収集された用例と作成されたパターン対の数

番号	動詞の種別 (IPAL表記(かな))	和英辞書		IPAL辞書		収集された用例			作成された P T対の数	
		登録語義数	PT数	語義数	用例数	新規作成	総数	倍率1		倍率2
1	でる(出る)	29(13)	22	32	49	145	194	6.1	4.0	65(3.0)
2	出す(出す)	16(7)	16	27	53	95	148	5.5	2.8	53(2.8)
3	はいる(入る)	8(6)	7	23	34	105	139	6.0	4.1	49(2.8)
4	たつ(立つ, 発つ, 建つ, 経つ)	27(8)	10	13	24	89	113	8.7	4.7	45(2.5)
5	たてる(立てる, 建てる)	21(4)	9	9	17	74	91	10.1	5.4	45(2.0)
6	いれる(入れる)	13(7)	5	19	30	113	143	7.5	4.8	45(3.2)
7	あがる(上がる)	18(6)	7	18	31	90	121	6.7	3.9	39(3.1)
8	おちる(落ちる)	16(5)	8	11	21	53	74	7.4	3.5	38(1.9)
9	あげる(上げる)	23(11)	8	21	31	98	129	6.1	4.2	37(3.5)
10	おとす(落とす)	16(5)	6	14	19	53	72	5.1	3.8	26(2.8)
11	あく(空く, 開く)	5(3)	9	10	12	25	37	3.7	3.1	23(1.6)
12	あける(空ける, 明ける, 開ける)	10(6)	11	11	17	32	49	4.5	2.9	22(2.2)
13	きまる(決まる)	2(3)	3	8	17	32	49	6.1	2.9	15(3.3)
14	きめる(決める)	1(5)	3	14	20	28	48	3.4	2.4	12(4.0)
15	たつ(断つ, 絶つ)	4(4)	8	4	--	12	12	3.0	--	11(1.1)
16	くずれる(崩れる)	9(3)	5	4	6	13	19	4.8	3.2	11(1.7)
17	くずす(崩す)	5(3)	6	4	4	8	12	3.0	3.0	10(1.2)
18	さく(割く, 裂く)	5(2)	3	5	7	7	14	2.8	2.0	10(1.4)
19	うめる(埋める, うめる, うずめる)	6(3)	3	8	9	9	18	2.3	2.0	5(3.6)
20	うまる(埋まる, うまる, うずまる)	2(2)	2	8	10	5	15	1.9	1.5	5(3.0)
21	さける(避ける)	3(1)	3	6	11	9	20	3.3	1.8	4(5.0)
22	さける(裂ける)	1(1)	1	1	3	4	7	7.0	2.3	4(1.8)
23	さく(咲く)	1(1)	1	1	1	3	4	4.0	4.0	3(1.3)
合計		241(109)	156	271	426	1102	1,528	5.6	3.6	577(2.6)
動詞当たりの平均		10.5(4.7)	6.8	11.8	18.5	47.9	66.4	--	--	25.1(2.6)

<備考>・登録語義数：研究社の和英辞書、但し、()内はライトハウスの和英辞書の場合
 ・P T数：辞書から作成されたパターン対の数(慣用パターンは除く)
 ・倍率1=用例合計/語義数、倍率2=用例合計/IPAL用例数
 ・最終的なP T数の欄の()内は、用例数との比を示す。

(3) 日本語語義と英語パターンの対応例

IPAL辞書の語義分類と最終的に得られたパターン対の語義分類の関係を見るために、「あがる」、「あげる」、「だす」、「でる」の4動詞について、IPAL辞書の語義がパターン対にどのように対応しているかを調べ、その結果を表3に示す。

この表は、日本語の語義のうち、英語の1パターンもしくは複数パターンに対応する語義は63.3%(43.9+19.4)であることを示している。残りの36.7%は、日本語の語義分類自体を組み替える必要がある。従って、英語との対比で語義分類を行うには、日本語内で分類した語義を再分類するだけでは適切でないことが分かる。

表3. IPAL語義と文型の対応

分類 動詞	「語義対」文型の関係				保留	合計
	l対l	l対n	m対l	m対n		
あがる	8	5	1	3	1	18
あげる	14	2	1	1	3	21
だす	8	9	5	4	1	27
でる	13	3	10	4	2	32
合計	43	19	17	12	7	98
%	43.9	19.4	17.3	12.2	7.1	100%

(4) パターン対作成上の問題点

用例作成では、特に英訳を考慮しなかったため、下記の例のように、文要素の対応がつきにくいものもあった。

例：彼は樽の栓をあけた。He tapped the barrel.
彼は護衛の任についた。

He was assigned as a bodyguard.

薬の効果が現れる。The medicine takes effect.

また、下記の例のように、一般パターンとしてどんな形にすればよいか判断に迷うものもあった。

例：「包丁で指が切れた。」(主語はIでよいか)

I cut my fingure with a knife.

「私は彼と話が合う。」(「馬が合う」、「話が合う」、「息が合う」などすべて慣用とするか)

He and I have common interest to talk about.

(5) 作成されるパターン対数の見込み

表2では、23動詞に対して、IPAL用例を含む1,528用例から577のパターン対が作成されているから、2.6用例当たり1パターン作成されたことになる。これに対して、表1では、806動詞に対して、IPAL用例を含め15,740の用例が収集されているから、これらの動詞に対して作成されるパターン対の数は、約6,000件(15,740÷2.6)になると予想される。

(6) パターン対作成の生産性

パターン対作成を支援するため、

①対訳用例文から日本語側、英語側のパターン候補自動作成。

②格要素の名詞に連結した意味属性の構造の表示。

③パターン登録後の検証試験の半自動化

などの機能を持つ支援システムを開発したことにより、パターン対作成の生産性は、従来の人手中心の方法に比べて、約6倍(10~20件/日)に向上した。このため、6,000件のパターン対は、2人年程度の工数があれば作成できる見込みである。

6. あとがき

和語動詞の意味的用法については、和英辞書の情報よりも人の知識の方が多彩で柔軟であると考えて、人の知識を内省する方法で動詞の用例を作成し、それを元に日英機械翻訳に必要な結合価パターン対の作成を試みた。その結果、一般パターンについて見れば、和英辞書から得られたパターン対の3~4倍の数のパターン対が得られることが分かった。

また、得られたパターン対を見ると、日英翻訳用のパターン対では、日本語の意味分類(IPAL辞書)に比べて、約2倍の細かさで意味分類することが必要であること、しかも、日本語内に閉じた意味分類を考えるだけでは不十分で、英語との関係で意味分類をすることが重要であること、などが分かった。なお、806の和語系動詞に対しては、約6,000件のパターン対が得られそうである。

謝辞 本検討に対して貴重な議論を頂いた当研究所横尾主任研究員、並びに、用例作成を手伝って頂いた武智しのぶ氏ほか株式会社スバルインターナショナルの方々へ深謝する。

参考文献

- (1) Almuallim, Akiba, Yamazaki, Yokoo and Kaneda: A Tool for the Acquisition of Japanese to English Machin Translation Rules using Inductive Learning Techniques, CALA94, pp.194-201, San Antonio, Texas, March 1994
- (2) 宇津呂、松本、長尾：二言語対訳コーパスからの動詞の格フレームの獲得, 情報処理論文誌(1992)
- (3) 北村、松本：対訳テキストからの翻訳知識の獲得と機械翻訳システムへの応用, 言語処理全国大会, pp. 289-292(1995)
- (4) 池原、宮崎、横尾：日英機械翻訳のための意味解析用の知識とその分解能, 情報処理Vol. 34 No. 8(1993. 8)
- (5) 情報処理振興事業協会技術センター：計算機用日本語基本動詞辞書IPAL(Basic Verbs), 解説編&辞書編(1987)
- (6) 白井、横尾、池原、井上：日英翻訳用構文意味辞書の記述精度の向上と作成支援, 48情報大6Q-9(1994. 3)