語形と意味

「x ガ流レ{ル，タ，テイル，テイタ}」の意味の多変量解析を用いた比較を通じて

李 在鎬[1]　永田 由香[2]　鈴木 幸平[3]　黒田 航[1]　井佐原 均[1]

（ [1] 情報通信研究機構　[2] 京都大学大学院　[3] 神戸大学大学院）

1. はじめに（問題の所在）

従来の語彙意味論では、語（特に動詞）の意味を記述する際、その語の語形毎に異なる意味構造を保持する、という想定はほとんど行ってこなかった。自然言語処理の言語資源 (例えば、日本語語彙大系[1]や IPAL 辞書[7]や格フレーム辞書[2]) においても実情としてはかわらない。

この種の記述の背景には、語の意味に対する静的見方があり、そこでは文脈から独立に語の意味記述が可能であると考えている (この種のアプローチの問題点については黒田・井佐原[5]を参照) 。このようなアプローチにおいては、接辞のような要素は単なる統語派生の副産物とされ、語彙記述の対象とは認識されていない (例えば「日本語基本動詞用法辞典[6]など)。しかし、この種の見方においては (1) のような事例を記述する際、経験的問題を引き起こす。

(1) a. 契約が流れた。
　　b. 契約の流れ。

(1) において興味深いのはいずれの例も同じ「流れる」の派生形ではあるが、(1a) では契約がキャンセルされたことを表すのに対して(1b) では契約の進行を表している。この種の非対称的分布は語基の「流れる」の意味から直接には予測できない。

また、(2) や (3) においても同じ「流れる」の派生形ではあるものの、語形によって異なる容認度を示す。

(2) a. 子供が流れた（cf. 流産の意味）
　　b. ??子供が流れていた。
(3) a. *斜面が北に流れた。
　　b. 斜面が北に流れていた。

以上の事実は、語形をめぐる事実の分布は派生的アプローチでは十分な記述ができないことを示唆する。本稿ではこうした事実を確認し、記述すべく、コーパスベースの調査を行った。ケーススタディとしては、鈴木・李・黒田[8]の「流れる」の事例分析を行った。

調査に際しては「新潮文庫」、「新潮新書」、「読売新聞」の三つコーパスから「流れる」の KWIC データを収集した。そして「タ形、ル形、テイル形、テイタ形」の語形別にデータを整理した。データ解析にあたっては、各々のサンプル文に表れている共起名詞 (e.g.,「川が流れる」の「川」) を 12 変数で特徴づけたあと、主成分分析とクラスタ分析を行った。そして、所属クラスタと実際の語形でクロス集計を取り、どのような一般化が得られるか考察した。

2. 先行研究と問題設定

本稿の問題意識およびデータセットをめぐる直接的な先行研究として金田一[3]や工藤[4]によるテンス・アスペクトモデルが挙げられる。この種のアプローチでは、動詞の意味と (動詞が内在的に有する) 時間的特徴の相関を記述することを目的としている。意味記述においては、構成的アプローチを基本にしており、lemma レベルの意味を保持させた上で、接辞の意味を積み木式的に計算していく。しかし、この種のアプローチの問題点を示唆する例として (4) が挙げられる。

(4) a. 橋が流れている。
　　b. ある組織に金が流れている。
　　c. 川が流れている。

(4) で注目すべきは、いずれの例においても、同じ語形を共有しているが、意味解釈においては異なる。金田一[3]の分析によれば、「流れる」は継続動詞にグループ化され、「テイル」と共起することで、動作が進行中（動作継続）であることを表すとされる。この分析では (4c) は予測可能であるが、(4a) や (4b) については経験的問題を引き起こす。というのも、(4a) などでは継続の読み以外にも、結果として橋が流れてしまっている状態をも表しているからである。詳細は言及しないが、工藤[4]の分析においても同様の問題が指摘できる。

これらの事実を踏まえ、鈴木・李・黒田[8]では、語形の相違による意味の多様性の問題は、lemma レベルの抽象的な語彙からの派生では、捉えきれないと考えた。この種の多様性の実態を確認するため、「流れる」と共起するガ格名詞の問題に着目し、二つの実験を行った。第一実験では、56 人の被験者に「○○が流れた」、「○○が流れていた」という形式の文を提示し、○○に入りうる語を制限時間、1 分間で書き出す、という事例産出課題を行い、実際に挙げられた回数とその平均順位を算出した。第二実験では 48 人の被験者に「X が流れた」と「X が流れていた」の「X」の部分に名詞を入れた刺激文カードを 12 枚 (「流れた」と「流れていた」の 2 組) を配布し、用法の類似性に基づいてグループ分けを行うように指示した[1]。結果を行列化し、多次元解析を行った。

実験の結果、両語形に共通する特徴として、液体の移動を表す用法がもっとも顕著であった。その次に顕著だったのは、「流れた」では「出来事のキャンセル (e.g., 計画が流れた)」や「物体の移動」を表す用法、「流れていた」では「連続する抽象物の継続」を表す用法が顕著であることが明らかになった。また、第二実験の結果、両語形では異なるグループ化の傾向が観察された。

鈴木・李・黒田[8]の分析は単一の動詞であっても、語形によって共起する名詞に有意な差があることを示した点では、大きな意味を持つ。しかし、この研究では、統制された状況において我々は一つの動詞であっても語形が違えば、違う (意味の) ものとして認識していることを確認しただけで、その背景にどのような要因が、どれだけ強く関与しているのか、さらにこうした効果がどの程度確実に観察されるのか、という問題については必ずしも明確には示せなかった。この問題を受け、本稿では、コーパス分析に基づく検証を行うことで、鈴木・李・黒田[8]を補強する。

3. データと分析方法

調査に際しては、三つのコーパスを利用した。それぞれのコーパスを茶筌で形態素解析した。解析においては KH Coder を使用した[2]。

コーパス	延べ語数	異なり語数
1. 新潮文庫	4,621,329	61,459
2. 新潮新書	1,847,791	48,913
3. 読売新聞	4,606,346	52,557

表 1. 形態素解析結果

KWIC 検索の結果、1 で 218 例、2 で 36 例、3 で 68 例が収集された。そして文末に生じている「ル、テイル、タ、テイタ」形 (終止形) のみを取り出した結果、表 2 のサンプルが得られた。

コーパス	語 形				合計
	タ	テイタ	テイル	ル	
1. 新潮文庫	50	52	33	19	154
2. 新潮新書	4	2	11	9	26
3. 読売新聞	17	5	18	7	47
合 計	71	59	62	35	227

表 2. 抽出サンプルの集計

[1] 使用した刺激文は下記の通りである。
1. 重苦しい沈黙が[流れた・流れていた]
2. あれから3年の月日が[流れた・流れていた]
3. 冬の冷たい空気が[流れた・流れていた]
4. 雲が[流れた・流れていた]
5. 雲間から月の光が[流れた・流れていた]
6. 頭から血が[流れた・流れていた]
7. ある組織に金が[流れた・流れていた]
8. 家々の前を川が[流れた・流れていた]
9. 足の上を砂が[流れた・流れていた]
10. 旅行の計画が[流れた・流れていた]
11. 迷惑メールが[流れた・流れていた]
12. 蛇口から水が[流れた・流れていた]

[2] KH coder は大阪大学（日本学術振興会）の樋口耕一氏によって作成されたコーパスツールである。詳細は http://khc.sourceforge.net/ を参照されたい。

収集データを 12 変数で特徴づけた。作業は第一著者、第二著者、第三著者が人手で行った。変数として具体物 (e.g. 紙幣、死体)、抽象物(e.g. 歳月、情報)、自然物(e.g. 春風、小川)、人工物(e.g. 銃、票)、事(e.g. クイズ、ムード)、液体(e.g. 汗、血)、固形(e.g. 落下傘、木片)、人間の分泌物(e.g. 涙、汗)、情報媒体(e.g. 映像、歌)、可視物(e.g. 文字、白光)、可触物(e.g. 尿、砂)、流体(e.g. 雲、煙)を用いた。

多変量解析として、三つの手法を使った。主成分分析とクラスタ分析と判別分析である。主成分分析は、データの分岐に関わる主要な特徴を抽出するため利用した。クラスタ分析は、サンプルのグループ化のために用い、判別分析はクラスタ分析の精度を評価するために用いた。

4. 結果と考察

主成分分析の結果 3 つの主成分が抽出された (第 2 因子までの累積寄与率はおよそ 69.2%)。

	成分		
	1	2	3
具体物	.859	-.243	.275
抽象物	-.889	.163	.156
自然物	.694	.551	.304
人工物	-.076	-.668	.118
事	-.610	.475	.466
液体	.839	.282	-.320
固形	.094	-.848	.240
人間の分泌物	.630	.375	-.328
情報媒体	-.587	.075	-.303
可視物	.768	-.168	.453
可触物	.851	-.355	-.130
流体	.673	.500	.028

表 3.　　主成分分析の結果

寄与率の大きく、明示的な解釈が可能な第一・第二主成分の得点を計算し、プロットしてみた。図1の結果が得られた。

図 1 によって、左上に人工物で形状を有する具体物の対象が、右上に自然物で流動性をもった対象が、左よりの下に情報性を持った抽象物の対象が、右よりの下に自然物で固形ではない対象が分布している。

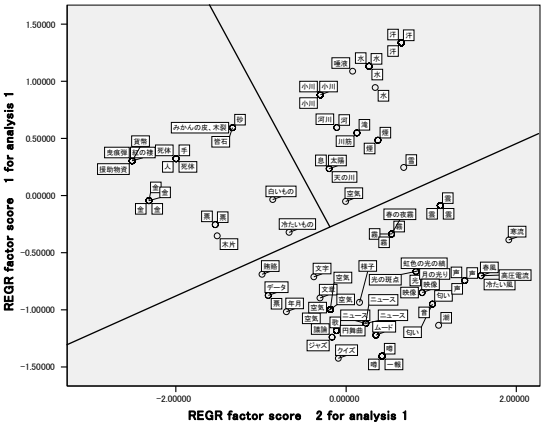

図 1　　主成分得点による分布

次に階層法によるクラスタ分析を行った (Ward 法、平方ユークリッド距離)。最適なクラスタ数の決定においては所属クラスタを従属変数に、元の変数群を独立変数にして、判別分析を行った。その結果、4 つのクラスタで最適な分離が得られた (交差確認済みのグループ化されたケースのうちに 98.7%が正しく分類され、もっとも高い値を示した) 。

クラスタ	度数 (%)	具体例
1	73(32.2)	噂、観測、議論、曲
2	30(13.2)	金、砂、品物、落下傘
3	37(16.3)	雲、霧、光、匂い、春風
4	87(38.3)	涙、川、血、水、煙
合計	227(100)	

表 4.　　クラスタの度数分布

次に各々のクラスタと語形をクロス集計した。その結果、図 2 の分布が観察された。

図 2 の分布においてまず注目すべきは、クラスタ 3 (自然物の流体) の分布である。クラスタ 3 は、テイル形とテイタ形のいずれの語形においてももっとも生産的なクラスタであることが分かる。一方、タ形やル形においては生産性が低い。またクラスタ 1 (抽象物) に関しては、タ形とテイル形においてもっと生産的なのに対して「テイタ形」や「ル形」に関しては生産的でなく、鈴木・李・黒田[8]と同様の結果が得られた。そして、クラスタ 2 (具体物) においては語形によ

る明示的な差は観察されなかった。

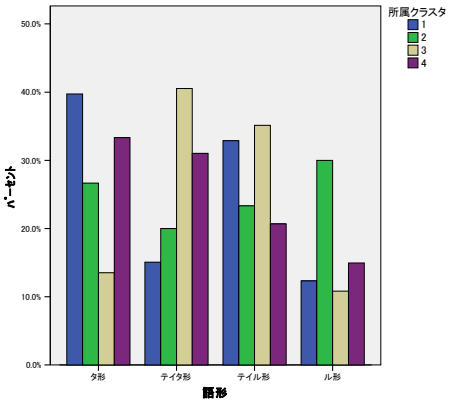

図 2　　語形と所属クラスタのクロス集計結果

図 2 の分布は間接的ではあるが、次の事実を示唆する。クラスタの生産性と語形の対応を考えた場合、クラスタ 1 に関しては「タ」形と「テイル」形が同じ名詞を多く共有しているのに対して、クラスタ 3 においては、「テイタ」形と「テイル」形が同じ名詞を共有している。クラスタ 4 に関しては「タ」形と「テイタ」形が同じ名詞を共有しており、語形による意味的類似・非類似の複雑な関係が示唆される。これは、従来のテンスアスペクトモデルが言うような「過去」対「非過去」あるいは「完了」対「進行」のような単純化したモデルでは説明できない。少なくとも従来のモデルが予測する分布は、図 2 では確認できない。これには、全体の事態に対する捉え方の問題が多分に関与していると思われるが、紙幅の都合上、詳細は別の機会に議論したい[3]。

5. まとめと課題

本論では、コーパスベースに語形と共起名詞の意味クラスタの対応を調査することで、各々の語形による非対称的意味の分布構造を明らかにした。理論的示唆として従来のテンスアスペクトモデルの問題点を示し、複雑な対応関係が存在することを示した。最後に二点の課題を述べる。

1. コーパス間の比較を行う。
2. 他の動詞に対しても同様の調査を行う。

1 に関連して、部分的ではあるが、読売新聞コーパスでは、クラスタ 1 の生産性が際立って高く、クラスタ 4 が低いことを確認している。コーパス間の比較に耐えうるデータ量が集まれば、コーパス間の比較を行うことで、本稿の観察がコーパスのバイアスによるものなのか、それともコーパスのバイアスを超え、普遍的に見られるものなのか、考察したい。2 に関連しては、本稿のケーススタディが流れるのみであることは、一般化としてたぶんに問題がある。この点を補うために、他の動詞に対しても同様の調査分析を行い、結果を検証していく必要がある。

〈参考文献〉

[1] 池原 悟, 宮崎 正弘, 白井 諭, 横尾 昭男, 中岩 浩巳, 小倉 健太郎, 大山 芳史, 林 良彦 (1997)『日本語語彙大系』岩波書店.

[2] 河原大輔, 黒橋禎夫 (2001)「用言と直前の格要素の組を単位とする格フレームの自動構築」『自然言語処理』Vol.9, No.1, pp.3-19.

[3] 金田一春彦(編) (1976)『日本語動詞のアスペクト』むぎ書房.

[4] 工藤真由美 (1995)『アスペクト・テンス体系とテクスト』ひつじ書房.

[5] 黒田 航・井佐原 均 (2006)「複層意味フレーム分析 MSFA を用いた文脈に置かれた語の意味の多次元的表現法」『認知言語学論文集』Vol.6, pp.171-181

[6] 小泉 保, 船城道雄, 本田晶治, 塚本秀樹 (1989)『日本語基本動詞用法辞典』大修館書店.

[7] 情報処理振興事業協会技術センター (1987)「計算機用日本語基本動詞辞書 IPAL 解説編」

[8] 鈴木幸平, 李在鎬, 黒田航 (2006)「実験に基づく「流れる」の語形の意味グループ」日本語用論学会 第 9 回大会 発表資料, pp.26

[3] この点に関連し、加藤鉱三氏 (信州大学) から次の指摘があった。描写モードの相違として「タ」形は事態の包括描写(事態を一括してひとかたまりに描写)であり、「テイル」形は事態の状況描写(事態を「現在ただ今」のものとして描写)である。