決定木を用いた多義語分析：多義動詞「出る」を例に

李在鎬 ((独)情報通信研究機構) ・ 伊藤健人 (群馬県立女子大学)

1. はじめに (背景と目的)

本稿の目的は、李(他)(2007)の延長で、コーパスデータの定量的・定性的分析モデルに基づき、日本語の多義語を分析することである。具体的には伊藤 (2003) が行った「出る」を含む移動表現を分析する。分析手法としては、複数のコーパスから分析サンプルを収集したあと、決定木 (decision tree) で解析する。さらに、決定木分析の妥当性を検証すべく、判別分析(discriminant analysis)を導入し、結果の信頼性を検討する。調査の結果から、1) 「出る」の語義決定に関わる要素を明示化できたことを報告する。そして、このことの理論的示唆として、2) 多義語の語義単位で優先される制約が異なること、3) 多義現象は構文(的制約)のみでも説明できなければ、語彙(的制約)のみでも説明できないことを示す。

2. 先行研究と問題の所在

2.1. 移動表現をめぐって

認知言語学において移動表現は事態認知の根源的特徴を反映する現象として多くの研究者によって注目されてきた (例えばLangacker (1991) や松本(1997) やTalmy (2003) など) 。とりわけ論争の焦点になっている記述的問題として以下の二点を挙げることができる。

(1) 移動事象がどの要素によって記号化されるのか。

(2) 移動から非移動への意味的シフトはどのようにして起こるのか。

(1)は文の意味として「移動」という事象がどのような文内要素によって表現されるのかという問題である。このことをめぐっては複数の流れが存在する。一つ目として動詞を中心とした一般化があり、上野 (2007) や松本 (1997) などが挙げられる。二つ目として格パターンからなる「構文(Construction)」を中心にした一般化があり、李 (2002) などが挙げられる。三つ目として、充分に確立した見解ではないが、Lee&Isahara (2006) などでは、文の意味決定に名詞の意味クラスが体系的に関わっていることを定量的に示しており、移動表現に関して言えば (文の意味とは独立に定義できる) 名詞の場所性の問題が深く関わっていることを明らかにしている。以下の事例から考えてみよう。

(3) a. 太郎が部屋に入った。
b. 太郎が部屋にいた。

(4) a. 生徒が教室の隅に固まった。
b. 恐怖で生徒が固まった。

(5) a. 船が闇の中に消えた。
b. 市長が凶弾に消えた。

(3)から(5)において注目すべきは、a の事例は b の状態表現に対して何らかの移動を伴う事態を表している点である。しかし、その移動が何によって、記号化されているのかという点をめぐってはいくつか複雑な問題が存在する。というのは、(3)の事例におけるコントラストを考慮した場合、動詞による意味的効果と捉えることができる。一方(4)の事例においては、(4b)が示すように動詞に直接的な移動を仮定することは困難であり、動詞以外の要素によってもたらされていると捉えるべきである。具体的には李(2002)が主張する慣習化された格パターンの意味的機能として移動事象が記号化されていると捉えることができる。しかし、(5)においては、動詞や格パターンのいずれの一般化も十分に捉えられないことを示唆する。というのは、(5a)では、状態動詞の「消える」が使用されているにも関わらず、移動事象を喚起しており、 (4a)同様、格パターンによる意味的機能と考えることができる。しかし、 (5b)においては「X ガ Y ニ消える」パターンにおいて生起しているにも関わらず、移動の解釈ができず、従来のいずれの分析枠組みからも十分な一般化できないことを示す。これは、Lee&Isahara (2006)の主張として、「ニ格」名詞が持つ場所性の問題を考慮しない限り、移動表現全体に対する充分な分析ができないことを示唆する。

以上で示した(1)の問題に加え、(2)の問題として、移動表現の興味深いところとして、(6)の現象が報告されている。

(6) a. 子供たちが城壁の上を走った。
b. 子供たちが非行に走った。
c. 石畳が寺内を縦横に走る。

(6)の事例において、注目すべきは移動動詞の「走る」の具体例であるにも関わらず、移動性・動作性をめぐる解釈は一枚岩ではない。というのは、(6a)のように完全な移動表現の具体例として使用される事例がある一方で、(6b)や(6c)のように移動表現としては捉えられない用法が存在する点である。こうした現象をめぐっては、主体化 (subjectification) の視座から、意味拡張を説明する研究の流れがある。以上の事実は、移動事象を表す言語的要素は動詞や格パターン、さらには名詞の意味クラスの問題など、多岐に渡っていることを示す。

2.2. 問題の所在

さて、本研究が取り上げる「出る」の具体例に関しては問題になるのは、以下の多様性が存在する点である(伊藤 2003)。

(7) a. 地震に驚いた人々が部屋から廊下に出た。
b. 太郎が会議に出た。
c. 我が校から優勝者が出た。
d. 太郎は住み慣れた街を出た。

(7)の事例において注目すべきこととして、移動動詞「出る」の具体例であるにも関わらず、前節の(6)同様に、移動性・動作性に関して大きな差が見られる点である。その証拠として、(7)を(8)のように言い換えた場合、その容認度に差が見られる点が挙げられる(#は「出た」の容認文ではあるが、

言い換え表現としての同一性が保持されないことを表す)。

(8) a. 人々が部屋から廊下に{移動する,?現れる,移る,*誕生する,?去る}
b. 太郎が会議に{?移動する, 現れる,?移る,*誕生する,??去る}
c. 我が校から優勝者が{#移動する,#?現れる, #移る,誕生する,#去る}
d. 太郎は住み慣れた街を{?移動する,*現れる,?移る,*誕生する,去る}

(7)が示す文意の問題を(8)に示した容認度の差から捉えた場合、(7a)は「位置変化」を表す用法であると考えることができる。そして、(7b)は移動の結果としての「出現」を表す用法、(7c)は状態変化としての「発生」を表す用法、(7d)は移動とその結果の融合として「離脱」を表す用法として捉えることができる。本研究では、伊藤(2003)の分析を踏まえ、(7)に見られる「出る」の具体例とその用法のクラスを表 1 のように定義する。

表 1　「出る」の用法のクラス

用法のクラス	定義
位置変化	主体が起点領域から経路を経て着点領域に移動すること
出現	既にある主体が着点領域に移ること
発生	もともとなかった主体が着点領域に生まれること
離脱	主体が起点領域と着点領域の境界線を越えること

表 1 を(7)に照らし合わせ、考えてみた場合、「位置変化」は「ガ格」でマークされるトラジェクター「人々」が「カラ格」でマークされる起点「部屋」から「ニ格」でマークされる着点「廊下」に物理的な移動する事態に対応する。一方、「出現」は、事態レベルの骨格は「位置変化」と同様の構造を持つが、着点となる「会議(場)」にプロファイルが当たる点で、「位置変化」とは異なる[1]。「発生」は、起点と着点を結ぶパスが移動ではなく、変化であり、着点にプロファイルがあたる事例である。最後に「離脱」は、「ガ格」でマークされるトラジェクター「太郎」が「ヲ格」でマークされる起点「住み慣れた街」から移動することを表しており、着点については言語化されないという特徴を有する。

本研究では、以上の問題を踏まえ、前節で示した記述的問題(1) を中心に考察する。具体的には表 1 に示した用法のクラスがどのような要素によって、エンコードされるのかという問題をコーパスデータに対する実験的分析手法で考察する。このことから記述的問題 (2) に対する示唆として、中心から周辺への意味拡張という単純化したモデルではなく、個々の用法の個別性を重視した分析が必要なことを示す。そして、方法論レベルで用法基盤モデル(Usage-Based Model; Langacker 1999) がどのようにして実践されるべきかという問題に対する具体的な方法論を提案する。

3. 分析方法

本研究では、「出る」の文レベルの意味の問題と、それを動機づける文内要素の貢献度の問題をコーパスベースに考察する。しかし、この考察においては二つの方法論的難題が存在する。

[1] プロファイルの相違を証拠づけるものとして、「位置変化」と「出現」では「ニ格」の必須度が異なる点が挙げられる。というのは、「出現」用法は「ニ格」を省略した場合、「位置変化」用法と違って、元の文意が保持されず、解釈が困難なものになる。

(9)　文の意味をどのように決定すれば良いのか

(10)　貢献度の有無をどのように判断すれば良いのか

(9)の問題として、表 1 の 4 つの用法クラスは、相互に排他的でない、すなわち連続的関係にあり、生のコーパスのデータの場合 (分析者が意図を持って作成したデータに比べ) 、個々の文の意味を一貫した基準で決定することは容易ではない。次に、(10)の問題として、(Langacker(1987)の Symbolic View の考え方に基づくのであれば) そもそも文の意味解釈に関与しない要素が文に表れることは理論的に許されないことであり、要素の貢献度を正当に評価することは実際問題として簡単ではない。こうした問題をクリアするため、(9)に対しては、複数の母語話者において、もっとも安定した用法のクラスをその文の意味であると仮定し、分析を行う。(10)に対しては、統計的手法を用いて有意差の有無を見る。具体的には決定木や判別分析を用いて、分類精度でもって貢献度を判断する。実際の解析では、(9)に対しては(11)、(10)に対しては(12)を行った。

(11)　アンケート調査: 5 名の母語話者に意味評定を依頼し、サンプル文に対してもっとも一致する用法クラスを調査した。

(12)　コーディングと統計解析: 30 の変数で分析サンプルの形式や意味をコーディングし、異なる条件づけで決定木分析した。

以上の方法で、母語話者のカテゴリー化をもっともうまく説明できる最適な条件づけを探った。最終的な狙いとしては、次の二点を目指す。1) 認知言語学が提案する分析モデル(e.g., Goldberg 1995; 構文モデル, Langacker 1987, 1999;トラジェクター・ランドマークモデルなど)の記述力に対する定量的・定性的評価を行う。2) 解析結果から得られる示唆は「出る」のみの問題ではなく、移動表現全体における多様性を反映していることを示す。

3.1. コーパスデータについて

調査に際しては二つのコーパスからサンプルを収集した。新潮新書 100 冊分 (規模: 1,847,791 語) と読売新聞 11 年間分 (規模: 4,606,346 語) である。それぞれのコーパスを「Mecab」で形態素解析したあと、文末形としての「出る」の用例のみを「chaki(茶器)」で抽出した[2]。その結果、KWIC フォーマットの 577 個の用例が収集された。次に(11)の意味評定を行った。評定においては前節の(7)の四つの用例との類似度に基づいてグループ分けするように指示した。調査では5名中 3 名以上が同じクラスを指定した場合、それをその文の意味であると認定した。

表 2　評定結果

		文意				合計
		位置変化	出現	発生	離脱	
コーパス	新潮新書	25	30	17	16	88
	読売新聞	14	79	391	5	489
合計		39	109	408	21	577

[2] Mecab および Chaki は自然言語処理の分野で開発されたコーパスツールであり、詳細は次を参照されたい。
Mecab: http://mecab.sourceforge.net/, Chaki: http://chasen.naist.jp/hiki/ChaKi/

表 2 の具体例を示す。

(13) a. 警察官に付き添われて章江は表に出た。
b. 問題日本野鳥の会から英語版の野外識別図鑑が出た。
c. 前期比四円五六銭の円高になったため差益が出た。
d. 「親子の縁を切る」と言い残し、自分の写真などをすべて持って家を出た。

(13a)は「位置変化」、(13b)は「出現」、(13c)は「発生」、(13d)は「離脱」の具体例である。次に、(11)の意味評定作業と平行して、(12)の作業として二種類の変数セットを使用し、サンプル文をコーディングした。

表 3　変数一覧

区分	対象	変数の具体例
(A) 格パターン	A-1. 動詞の直前格	ガ、ニ、カラ、ヲ、ヘ、マデ
	A-2. 文の先頭格	
(B) 共起名詞の意味クラス	B-1. トラジェクター	主体、具体物、場所、抽象物、事、抽象的関係
	B-2. 起点	
	B-3. 着点	

(A) は構文レベルの制約を変数化したもので、実際の文に表れた形式に基づいて判断した。具体的には、A-1 は、「出る」の直前に表れている格要素、A-2 は文頭に表れている格要素を対象に行った。次に (B) は語彙レベルの制約を変数化したもので、池原 (編) (1999) の「名詞の意味属性分類」に従って判断した。なお、トラジェクターに関しては (細かい部分での問題はあるが) Langacker (1991)の主張に従って、文における主語、すなわち「ガ格」名詞をトラジェクターとした。なお、起点(Sauce)と着点(Goal)は Langacker(1987,1991)のランドマーク(Landmark)を細分化したものである。トラジェクターのコーディングに関する具体例を示す。

(14) a. 松尾は映画館を出た。 -> 主体
b. 毎年十月に新豆が出る。 -> 具体物
c. 地方圏でも値下がりする地域が出ている。 -> 場所
d. 「政党の生命は政策」といった言葉が何回も出た。 -> 抽象物
e. 苦しい状況になった時に人間の弱さが出る。 -> 抽象的関係
f. 双子の男児を出産させていたことが明るみに出た。 -> 事

3.2. 統計解析

前節で示したデータに対して、統計的手法を用いて解析を行った。解析においては被験者側が行った評定結果を従属変数、分析者側がコーディングした結果を独立変数として投入し、独立変数から従属変数を予測し、分類するタスクを行った。また従属変数を固定し、 (A) , (B) の変数セットを合計 6 通りの条件で投入し、どのような条件づけにおいて、もっとも正しい分類結果が得ら

れるか比較検討した[3]。

- **解析 1: 格パターンのみから、「出る」の用法クラスを分類**
- **解析 2: トラジェクターのみから、「出る」の用法クラスを分類**
- **解析 3: 起点と着点から、「出る」の用法クラスを分類**
- **解析 4: トラジェクターと起点と着点から、「出る」の用法クラスを分類**
- **解析 5: 格パターンと起点と着点から、「出る」の用法クラスを分類**
- **解析 6: 格パターンとトラジェクターから、「出る」の用法クラスを分類**

データ解析に用いたのは、決定木 (decision tree) 分析である。決定木は、人工知能やデータマイニングなどの分野で使われる標準的な予測モデルの一つとされているが、主として多変量データをいくつかのカテゴリーに分類するタスクや分類に影響を及ぼす変量の値により、次々と分類を進めていく手法で、非線形の分類法を与える(cf. 岩崎(編) 2004:81)[4]。分析ツールはSPSS 14を使用しており、木の成長方法はCHAID (Chi-squared automatic interaction detection)を使用した[5]。重み付け(影響度変数)としては評価の一致率を使用した。また、従属変数は、評定結果を、独立変数はコーディングの結果を用いて行った。また、結果の信頼性を検証すべく、判別分析を行った。判別分析は、決定木と同様、従属変数は、評定結果を、独立変数はコーディングの結果を用いて行った。

4. 結果と考察

6 つの変数セットで解析を行った。その結果、観測値(母語話者の評定結果)に対する予測の的中率(正答率)が異なることが明らかになった[6]。6 つの実験による全体の正答率をプロットしてみた。

図 1 では、「出る」の用法クラスを 6 つの条件で分類させた際の正答率の推移が示されている(詳細な値は文末資料参照)。結論的には、解析 6(格パターンとトラジェクターの意味クラスを使用して解析した時)において、もっとも高い正答率、すなわち母語話者のカテゴリー化をもっとも効率よく説明できることが明らかになった。一方、解析 3(着点と起点の意味クラスで解析)は、解析 1 や解析 2 に比べても著しく低い精度を示すことから、相対的に説明力が弱いことが明らかになった。

---

[3] 分析では、7 つ目の条件として表 3 のすべての変数を投入して分析したが、結果的には、解析 6 と同じ値になっているので、ここでは省略する。

[4] 詳細なアルゴリズムは紙幅の都合上、省略するが、コーパス分析における利用法については、玉岡(2006)を参照されたい。

[5] CHAID は樹木モデルにおいてもっとも広く用いられているアルゴリズムの一つであり、変数の分岐基準としてカイ 2 乗統計量や F 検定統計量など統計学で多く用いられている統計量が用いられている。各ステップにおいて、CHAID は、従属変数と最も強い交互作用を持つ独立(予測)変数を選択する。もし有意差が無ければ各予測変数はマージされる形で木を成長させる。

[6] 的中率および正答率とは、母語話者が行った「出る」の用法をめぐるカテゴリー化を格パターンおよび名詞の意味クラスといった文内要素がどの程度正しく予測するかを表す指標であると位置づけられている。

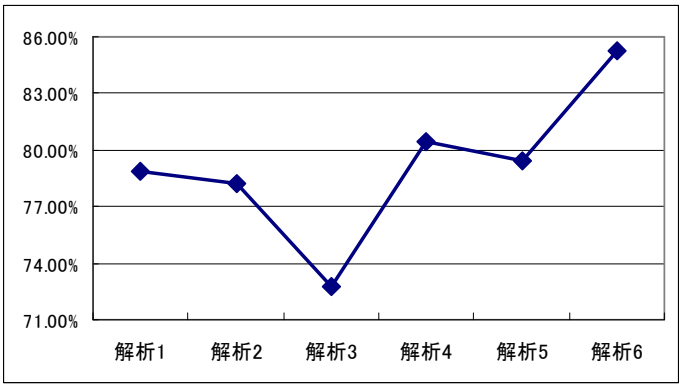
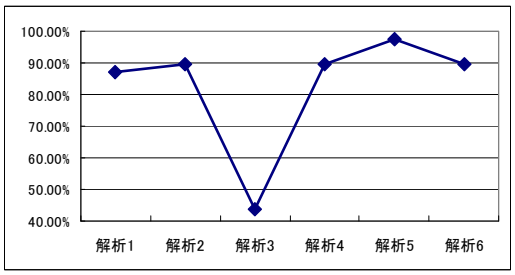
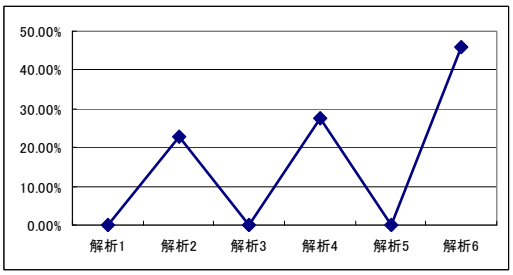


図 1　　決定木による全体の正答率の推移

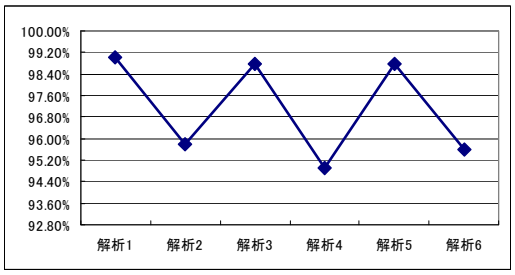
次に、個々の用法クラス別の正答率を見た。詳細を以下に示す。



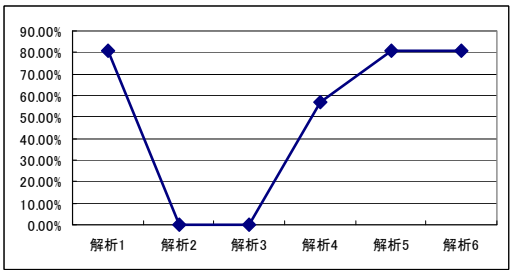
a. 位置変化



b. 出現



c. 発生



d. 離脱

図 2　　クラス別正答率の推移

図 2において注目すべきは、すべての用法クラスが解析 6でもっとも良い精度を示すとは限らない点である。というのも、「位置変化」や「発生」に関しては解析 6 でむしろ精度が下がっている。結論的には、四点の興味深い観察結果が得られた。

(15)　a.　「位置変化」は解析 5（格パターンと起点と着点の意味情報）でもっとも良い精度を出しており、構文情報とランドマークの語彙情報を共に使用したカテゴリー化の傾向が見られる。

b.　「出現」は解析 6（格パターンとトラジェクターの意味情報）で良い精度を出しており、構文情報とトラジェクターの語彙情報を共に使用したカテゴリー化の傾向が見られる。

c.　「発生」はどのような条件においても高い学習精度が得られているが、中でも解析 1（格パターン）が示すように、構文情報によるカテゴリー化の傾向が見られる。

d. 「離脱」は格パターンを入れることで飛躍的に精度が向上しており、構文情報によるカテゴリー化の傾向が顕著に見られる

以上の結果が示唆する一般化の一つとして、用法単位で優先される制約が異なっている可能性が浮上してくる。これらの結果を検討すべく、同様の方法で、判別分析を行った。紙幅の都合上、図 1 に対応するものとして、全体の交差確認済みの正答率の変異のみを示す。

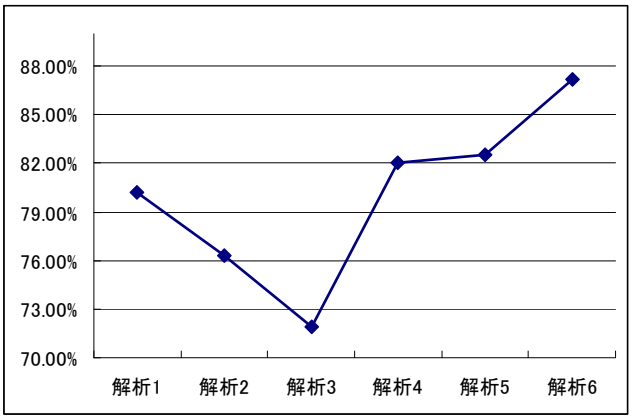


図 3　判別分析による全体の正答率の推移

図 3 に関して注目すべきは、図 1 の結果とほぼ同じ傾向を示している点である。このことが偶然でない限り、図 1 さらには、図 2 で得られた一般化は、決定木という手法に依存することなく、妥当な一般化であることが示唆される。

次に、解析 6 の結果を元に、用法クラスを階層化した。図 4 では、独立変数の値から、合計 8 つのノードに各々の用法がグループ化されていく様子が図示されている。まず、直前格として「ヲ格」かどうかという特徴によって、「離脱」を表す用法(ノード 2)とそれ以外の用法(ノード 1)に分岐する。次に、ノード 1 は直前格として「ガ格」かどうかという特徴によって(直前格として「ガ格」を持つ)「出現」を部分的に含む「発生」の用法(ノード 4)と(直前格として「ガ格」を持たない)「出現」を含む「位置変化」の用法(ノード 3)に分岐する。次にノード 3 はトラジェクターとして「主体」かどうかという特徴によって(トラジェクターが主体である)「位置変化」の用法(ノード 5)とそれ以外の用法(ノード 6)に分岐する。最後に、ノード 4 はトラジェクターとして「具体物」かどうかという特徴によって(トラジェクターが具体物である)「出現」の用法(ノード 8)と抽象物が主である用法(ノード 7)に分岐する。この分岐パターンにおいて興味深いことは、上位階層においては、「出る」の前に生じている格が「ヲ格」なのか「ガ格」なのかといった構文情報によって、カテゴリー化がなされているのに対して、下位階層ではトラジェクターの語彙情報として、「主体」や「具体物」といった特徴によってカテゴリー化されている点である。このことが示唆する理論的問題の一つとして、構文文法の分析モデルは上位階層の分岐を、認知文法の分析モデルは下位階層の分岐を捉える上で有効であると結論づけられる。

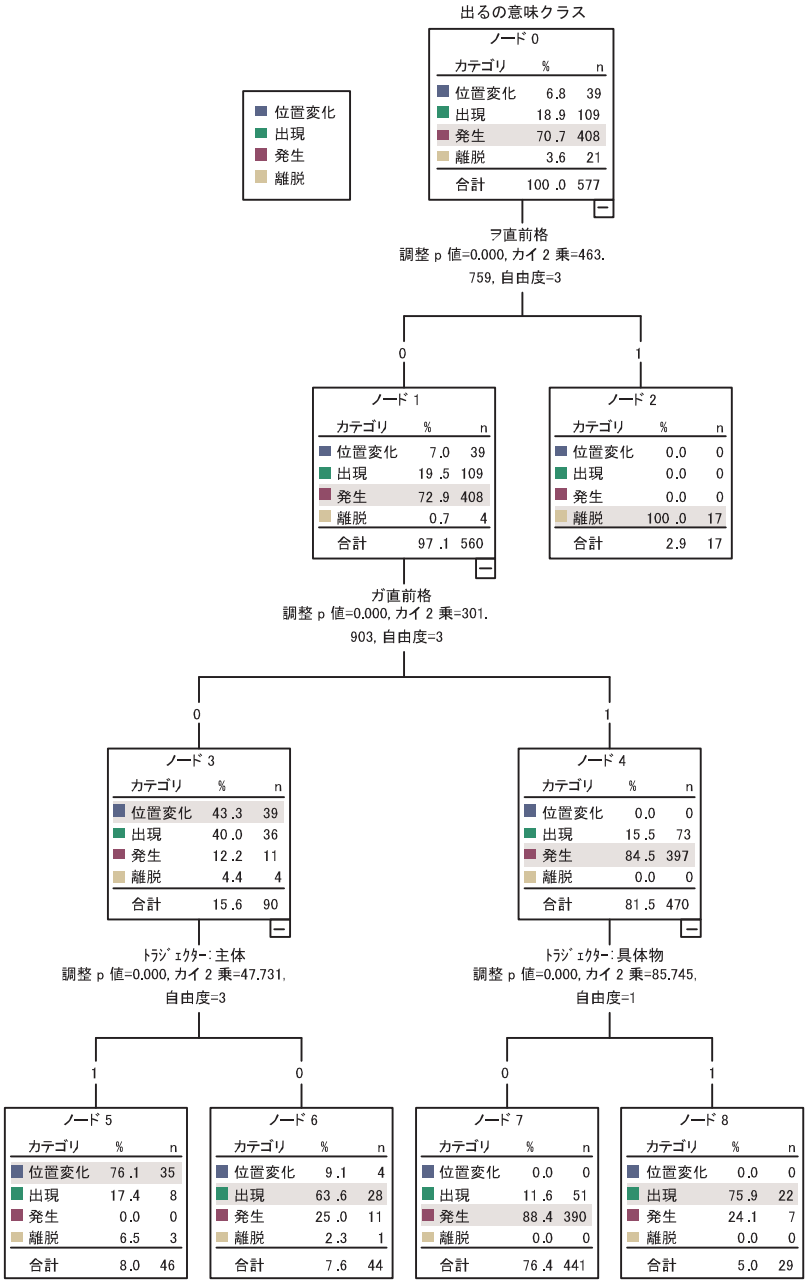


図 4　解析 6 による決定木

5. 最後に

本研究では移動動詞「出る」を例に、母語話者によるカテゴリー化の実態を統計的手法で考察した。この考察によって、明らかになったことの一つとして、動詞の用法単位で優先される制約が異なる点である。この観察が妥当であるなら、従来の認知言語学におけるプロトタイプカテゴリー論による多義語分析は、個々の用法の質的違いを捨象した一般化である可能性が浮上する。そもそも用法によって、受ける制約が異なり、異なる意味の構造を有する以上、それを中心と拡張という単純化した図式で、相互を結びつけることが果たして妥当なことだろうか。

*謝辞: 本研究は博報堂「ことばと文化・教育」研究助成(代表:李在鎬, 06-B-0039)および科学研究費補助金(若手(B), 課題番号: 19720111)の援助を受け行った。感謝申し上げる。

＜参考文献＞

池原 悟 (編) (1999)『日本語語彙大系 CD-ROM 版』,東京:岩波書店.

伊藤健人 (2003)「動詞の意味と構文の意味 ー「出る」の多義性に関する構文文法的アプローチー」『明海日本語』(8) , pp. 39-52.

岩崎 学(他)(編) (2004)『実用 統計用語事典』, 東京:オーム社.

Goldberg, Adele, E. (1995) *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: The University of Chicago Press.

松本曜 (1997)「空間移動の言語表現とその拡張」,『空間と移動の表現』,東京: 研究社出版.

Langacker, Ronald W. (1987) *Foundations of Cognitive Grammar, Vol I, Theoretical Prerequisites.* Stanford, California: Stanford University Press.

Langacker, Ronald W. (1991) *Foundations of Cognitive Grammar, Vol II, Descriptive Application.* Stanford, California: Stanford University Press.

Langacker, Ronald W. (1999) “A Dynamic Usage Based Model”. in Barlow, Michael and Suzanne Kemmer (eds.) *Usage Based Model of Language*. Stanford, California: CSLI Publications, pp.1-64.

李在鎬 (2002)「構文の意味的動機付けに関する一考察:「X が Y に V する」を例に」,『日本言語学会 124 回研究大会予稿集』,pp.226-231.

LEE, Jae-Ho & ISAHARA Hitoshi (2006) "A Cognitive Approach to Japanese Constructional Phenomena: Evidence from Motion Construction and Resultative Construction" International Conference on Japanese Language Education 2006, pp.25-26.

李在鎬・鈴木 幸平・永田由香・黒田航・井佐原均(2007)「動詞「流れる」の語形と意味の問題をめぐって」『計量国語学』(26 巻 2 号), pp.64-74.

Talmy, Leonard (2003) *Toward a Cognitive Semantics: Typology and Process in Concept Structuring, Vol. 2*, London: The MIT Press.

玉岡賀津雄 (2006)「決定木」分析によるコーパス研究の可能性: 副詞と共起する接続助詞「から」「ので」「のに」の文中・文末表現を例に」『自然言語処理』13 (2), pp.169-179.

上野誠司 (2007)『日本語の空間表現と移動表現の概念意味論的研究』、ひつじ書房.

山梨正明 (2000)『認知言語学原理』、くろしお出版.