

## 多義動詞用例の分類結果に基づいた語義の規定手法

山口昌也

国立国語研究所

masaya@kokken.go.jp

**概要** 多義語の語義の分類基準は、辞書を構築する際に必要不可欠である。多義語の深層構造を考慮しつつ語義分類するには、なんらかの形で人手で分類基準を規定する必要がある、その際に客観的な分類基準を規定することが求められる。本論文では、用例をクラスタリングする際のクラスタ間の類似度を語義分類基準と位置付け、複数の分類者による用例の分類結果から語義分類基準を獲得する手法を提案する。獲得される語義分類基準は、語義を規定する用例クラスタ組を分類者が同一の語義であると認識する確率として表現される。本手法は、あらかじめ特定の分類基準を規定しない状態から分類基準を規定することができるとともに、複数の分類者を想定して、分類基準に客観性を持たせている。本稿では、提案した手法で多義動詞用例のクラスタリングを行い、得られたクラスタを分析、考察を加えた。

**キーワード：** 多義語, 語義, 用例のクラスタリング

## A Method of Discriminating Meanings of a Polysemous Word Based on Results of Discriminating Examples

Masaya YAMAGUCHI

Section for Lexcography, The National Language Research Institute

masaya@kokken.go.jp

**Abstract** The criterion for discriminating meanings of a polysemous word is essential for the construction of a semantic lexicon, and is required to be an objective one. In this paper, we propose a method of acquiring the criterion for discriminating meanings of polysemous word based on results of discriminating examples by more than one subject. The criterion is used for calculating the similarity between example clusters in the example clustering algorithm, and is defined as by the probability where a couple of example clusters would be recognized as the same meaning by the subjects. Our method allows us to acquiring the objective criterion without defining the criterion for the discriminations of subjects in advance. We discriminate examples of Japanese polysemous verbs using the acquired criterion, and analyze each cluster to evaluate the results.

**Keywords:** Polysemous Word, Meaning, Clustering of Examples

## 1 はじめに

多義語の語義分類は、国語辞典や計算機用の意味解析辞書をはじめとして、辞書を記述する際の基本的な分類である。本研究は、意味解析用辞書を構築するための前段階として、用例の語義分類結果に基づいて多義語の語義分類基準を構築することを目的とする。ここでは、語義分類基準構築の対象として、日本語多義動詞を扱う。

多義語の語義分類基準としては、さまざまな方式が提案されている [1]。まず、表層的な情報を語義分類に利用する手法として、表層的な共起情報の類似性に基づいて用例をクラスタリングする手法 ([2], [3], [9])、動詞の表層的な格フレームにより類別する手法 [5] がある。これらは、用例から自動的に獲得できるという利点があるが、表層的な情報からだけでは、深層構造の違いまで考慮した語義の分類は困難である。

一方、深層格構造を考慮した手法としては、表層格フレームと深層格フレームとの関係を分類基準とする手法 [4], [7] がある。文の意味表現として深層格フレームは従来から使われており、語義分類するための基準としても有効である。しかし、適切な深層格や格要素の意味属性を設定することが難しいという問題がある。また、特定の意味表現を想定するのではなく、多言語コーパスという表層的な情報を用いることにより、動詞の表層的な格フレームを獲得する研究 [6] も行われている。この手法は表層上の情報以外に他の言語の知識を使って、語義を分類していることになり、機械翻訳などのタスクを実行するためには有効な方法である。しかし、一般的な語義の分類として適切かどうかはわからない。

以上のことから、深層構造まで考慮した語義分類を行う際には、(a) なんらかの形で人手が必要であるということ、(b) あらかじめ特定の深層構造を規定しておくことは困難であるということ、を考慮すべきであると考えられる。

そこで、本研究では、用例をクラスタリングする際のクラスタ間の「類似度」を語義分類基準と位置付け、人手による用例の語義分類結果から、語義分

類基準を獲得するという方法を用いる。獲得される語義分類基準は、分類者が用例組を同一の語義と認識する確率を基本し、二つの語義が同一の語義であると認識する確率で表現される。語義分類基準獲得のために用例分類を行う際は、個々の分類者独自の基準により分類を行うが、複数の分類者を想定し、個々の分類者の主観を排除する。

本論文の構成は、次のとおりである。まず、次節において、語義分類基準、および、用例のクラスタリングの方法について述べる。第3節では、複数の分類者による用例の分類結果から語義分類基準を獲得する方法について説明する。第4節では、提案した手法を用いて語義分類基準を獲得する実験を行う。さらに、クラスタリングされた各クラスを分析し、クラスタリングの妥当性を評価する。第5節では、本研究のまとめを述べる。

## 2 語義の分類基準

すでに述べたように、本論文では研究では、用例をクラスタリングする際のクラスタ間の「類似度」を語義分類基準と位置付けている。この節では、まず、用例のクラスタリングの流れを述べる。さらに、本論文で想定する用例の条件、および、語義の同一性判断の基準を述べたあと、語義の分類基準の指標となる同一語義確率を定義する。

### 2.1 語義と用例のクラスタリング

本論文では、語義は用例の集合から規定されるものとする。語義を規定する用例の集合は、用例をクラスタリングすることにより得る。語義分類基準は、クラスタリングにおけるクラスタ間の類似度である。以後、この類似度のことを同一語義確率と呼ぶことにする。

用例をクラスタリングする際には、表層情報からクラスタ間の類似度を計算する方法がよく用いられるが (例えば, [2],[3]), 本論文では、分類者による用例の分類結果から同一語義確率を計算する。

用例のクラスタリングの流れを次に示す。

- (1) クラスタ集合  $C$  を初期化する。クラスタリング対象の用例集合  $E$  の要素数が  $N$  個のとき、 $N$  個のクラスタ  $C_1, \dots, C_N$  を用意し、クラスタ  $C_i$  に用例  $e_i \in E$  を入れる。
- (2)  $(C_i, C_j) = \operatorname{argmax}_{C_i, C_j} (P_c(C_i, C_j))$  を求める。そして、 $C_i$  と  $C_j$  を合併し、新たなクラスタ  $C'$  とする。ここで、 $P_c$  は同一語義確率である。同一語義確率については、この後の節で説明する。
- (3)  $|C| = 1$  になるまで、(2) を繰り返す。

## 2.2 用例の条件、および、語義の同一性判断の規準

同一語義確率を定義する前に、本論文で想定する用例の条件、および、語義の同一性判断の規準を述べる。

まず、用例は対象とする多義語の語義を一意に決定できる範囲の文脈を含むものとする。

次に、語義の同一性を判断する際の基準だが、意味的な要素に限定し、構文的な要素については考えないものとする。例えば、**E1a**, **E1b** は、能動態、受動態の関係にあり、構文的には異なるが、深層構造は同一であり、同じ語義であるとする。また、動詞「消す」の用例 **E2a**, **E2b** も格要素「香辛料」の表層上の格は、ガ格とデ格で異なるが、深層格は同一であると考え、同じ語義であるとする。

**E1a** 先生が太郎を誉める

**E1b** 太郎が先生に誉められる

**E2a** 香辛料が肉の臭みを消す

**E2b** 香辛料で肉の臭みを消す

なお、意味解析用の辞書を考えた場合、構文的な要素を記述しておくことは必要だが、本研究の目的は語義の分類規準を獲得することであるため、上記のような条件にした。ただし、最終的には、クラスタリングの結果から構文的な情報（例えば、格フレーム）を導き出すという立場である。

## 2.3 同一語義確率

同一語義確率は、二つのクラスタ間の類似度であり、クラスタに含まれる用例によって規定される語義が同一のものとして認識される確率として表現される。3節で述べるように、同一語義確率は分類者が用例を分類した結果から求められる。ここでは、同一語義確率の定義だけ述べる。

### 2.3.1 同一語義確率の定義（用例間の同一語義確率）

同一語義確率は、用例クラスタ間に定義されるものであるが、まず、用例間の同一語義確率（要素数が1のクラスタ間）を定義する。用例  $e_1, e_2$  との間の同一語義確率  $P_c(e_1, e_2)$  を次のように定義する。

用例  $e_1, e_2$  を分類者に提示したときに二つが同一の語義として認識される確率

この値は、分類者ごとに個人差が存在することが予想される。実際、国語辞典を見てみると、辞典ごとに語義の認識の仕方が異なる。例えば、動詞「投げる」の語義分類を広辞苑 [10] と国語大辞典 [11] で比較してみると、前者の場合、次の用例を別々の語義として記載しているが、後者では「空中に物をほうり出す」ということで両者を同一の語義としてまとめている。

**E2a** 太郎がボールを投げる

**E2b** 次郎が太郎を投げる

このようなことが起こるのは、語義を一般化する程度によると考えられる。上記の例の場合、「投げる」のヲ格の格要素に対して、広辞苑ではボールなどの物体と人間を区別して語義分類をしているが、国語大辞典では「ボール」と「人間」をそれぞれ「物」という解釈で一般化して、語義を一つにしている。

仮に一人の分類者だけを仮定した場合、その分類者固有の同一語義確率が決定してしまうが、本論文では、複数の分類者を仮定することによりそれを防いでいる。

### 2.3.2 同一語義確率の定義（用例クラスタ間の同一語義確率）

用例クラスタ  $C_a, C_b$  間の同一語義確率  $P_c(C_a, C_b)$  は、次のように定義する。本論文では、一方の用例クラスタに含まれる用例と他方のクラスタに含まれる用例との間の同一語義確率の平均値であるとして定義した。

$$P_c(C_a, C_b) = \sum_{e_{ai} \in C_a} \sum_{e_{bj} \in C_b} P_e(e_{ai}, e_{bj}) / |C_a| |C_b|$$

## 3 同一語義確率の獲得

この節では、同一語義確率を分類結果から獲得する手法を示す。まず、分類者が分類を行う際の分類回数を削減する方法を二つ示し、その後、同一語義確率を獲得するまでの流れを示す。

### 3.1 複数用例の同時分類

同一語義確率を計算するには、用例間の同一語義確率  $P_e$  を実験的に収集する必要がある。この際問題となるのが、分類者の手間が大きさである。単純な一対比較を繰り返し行うことを考えた場合、組合せ数は、 ${}_n C_2$ （用例数が  $n$  個の場合）となる。例えば、 $n = 50$  のとき、1225 回もの一対比較が必要になってしまう。

そこで、次の仮定の下に、複数の用例を一度に提示して分類することにする。

**仮定**  $|E| < th$  のとき、 $\forall i, \forall j, e_i, e_j \in E$  に対して、次の式が成り立つものとする。

$$P_c(e_i, e_j | \{e_i, e_j\}) \cong P_c(e_i, e_j | E)$$

なお、 $P_c(e_i, e_j | E)$  は、用例集合  $E$  を提示したときに、 $e_i, e_j$  を同一語義として認識する確率であるとする。

この仮定のもとでは、 $n$  個の用例を提示し、 $m$  個のクラスタに分類されたとき、用例  $e_i, e_j \in C_k$  の条

件が満たされていれば、 $e_j, e_j$  は同一の語義と分類者が判断したとすることができる。

しきい値  $th$  については、本論文では経験的に決定し、議論することはしない。ただし、分類者の語義分類規準が定まっている状態になっていないと、上記の仮定が成り立たなくなる可能性が高いため、実際の分類に際しては、分類者の語義分類規準が定まっている状態になるようにする。

### 3.2 類似用例の同一化

用例の分類をする際に、非常に似ている用例は、一つの用例としてまとめることにより、分類回数を減らすことができる。例えば、用例 **E3a** と **E3b** は非常に類似した用例であり、両者を別々に **E3c** と比較分類するのではなく、どちらか一方を使って、分類すればよい。

**E3a** 太郎は赤いボールを投げる。

**E3b** 次郎は白いボールを投げる。

**E3c** 三郎はその試合を投げた。

ただし、二つの用例が類似しているか否かは、同一語義確率が得られてからでないと判断できない。そこで本論文では、最終的に得られた分類結果に対して、同一語義確率を計算し、その結果に基づいて同一化の処理を行うことにする。この場合、分類結果が得られた段階で同一化を行うため、実験回数の削減にはならないが、同一化したことにより、実質的に分類回数が増えたことと同等なので、同一語義確率の精度の向上を図れると思われる。

同一化の処理は、次の手順で行う。まず、分類度数のことを考慮し、用例  $e_i$  と  $e_j$  を同一化する条件を次のように定め、

$$P_c(e_i, e_j) > t_p, \text{ かつ } f(e_i, e_j) > t_n$$

上記の条件が成り立ったときに、次のような同一化の処理を行う。ここでは、用例  $e_i, e_j$  と  $e_k$  との間の同一語義確率を  $e_i$  の同一語義確率のほうにまとめている。任意の  $k (\neq i, j)$  に対して、同様の処理を

行う。なお、4節の実験では、 $t_p > 0.85$ ,  $t_n > 7$ とした。

$$P_e(e_i, e_k) = (f_s(e_i, e_k) + f_s(e_j, e_k)) / (f(e_i, e_k) + f(e_j, e_k))$$

ただし、 $f(e_i, e_j)$ ,  $f_s(e_i, e_j)$  は、それぞれ用例  $e_i$ ,  $e_j$  の分類を行った回数、その分類の結果同一語義であると判断された回数である。また、 $P_e(e_i, e_k)$ ,  $P_e(e_j, e_k)$  は、次のとおりである。

$$P_e(e_i, e_k) = f_s(e_i, e_k) / f(e_i, e_k)$$

$$P_e(e_j, e_k) = f_s(e_j, e_k) / f(e_j, e_k)$$

### 3.3 同一語義確率獲得の流れ

同一語義確率の獲得の流れを次に示す。ここでは、分類する用例集合を  $E$  とし、 $|E| = N$  とする。

- (1)  $E$  からランダムに  $M$  個の用例を取り出し、分類者に分類してもらう。
- (2) (1) を  $L$  回繰り返す。この後示す実験では、 $|E| = 50$ ,  $M = 10$ ,  $L = 50$  として実験を行った。なお、 $|E| = 50$ ,  $M = 10$  のとき、任意の用例組が取り出される確率は、 $9/245$  である。今回は、各分類者が任意の用例を 2 回程度分類されるような  $L$  の値を設定した。
- (3) すべての分類者の分類結果から同一語義確率を計算する。
- (4) 3.2 節で示した類似用例の同一化の処理を行う。

## 4 実験

この節では、3節で示した手法に基づいて、用例の分類実験を行い、用例間の同一語義確率を獲得する。さらに、2節で示した方法により用例のクラスタリングを行い、その結果を示す。

### 4.1 実験条件

同一語義確率獲得のための用例分類は、次の条件で行った。

- 分類対象の動詞は、IPAL 動詞版 [4] から語義数が 5 以上の動詞をランダムに選んだ。今回は、三つの多義動詞「ぬける」(8), 「よぶ」(5), 「わかる」(5) で実験を行った (カッコの中は IPAL に記載されている語義数)。
- 用例は、上記の三つの動詞それぞれに対して、岩波新書、および、岩波少年文庫 150 冊から無作為に 50 個取り出した。分類者に提示する際は、対象とする動詞の前後 100 文字の文脈をつけた。
- 分類者は、国立国語研究所国語辞典編集室に勤務している非常勤研究員、および、アルバイト 6 名である。いずれも国語学、もしくは、言語学の教育を受けたか、関連する職業に従事した経験がある。
- 分類者による用例の分類は、50 個の用例の中から無作為に取り出した用例 10 個を分類者に提示し、分類してもらった。各分類者の分類回数は、「ぬける」、「よぶ」、「わかる」それぞれ平均 48 回である。なお、分類者の語義分類を確定させるため、実際の分類を行う前に、5 回の訓練用の分類を行っている。

### 4.2 実験結果

「ぬける」、「よぶ」、「わかる」の用例のクラスタリング結果をそれぞれ図 1, 2, 3 に示す。これらの図では、葉ノードには語義説明用に代表的な用例の一部、もしくは、語義説明を示してある。また、葉ノードに付与されている数値は、葉ノードに含まれる用例の数である (ただし、代表的な用例、および、語義説明はあくまで説明用である。後述するように「わかる」は、語義説明・代表的な用例を示すのが難しかったため、示していない。また、同一クラスタ

に含まれるのが妥当とは思えない用例が含まれている場合もある)。さらに、IPALの語義分類と比較するために、右側にIPALの語義見出しとの対応関係を示した。

なお、スペースの関係上、同一語義確率が0.75以上の中間ノードは、展開せず、葉ノードとして表示してある。1.0に達していない葉ノードがこれにあたる。

## 5 考察

ここでは、クラスタリングの妥当性を検証するために、クラスタリング結果中の個々のクラスタに含まれる用例を分析して、そのクラスタに対する解釈を行う。また、IPALとの語義分類の比較を行う。

### 5.1 「ぬける」に対する分析

図1に示したように、葉ノードのクラスタに対しては、一定の解釈をつけることができた。また、同一語義確率を低いほうから見てみても、まず、「何かが取り除かれる」と解釈できるクラスタ(図1上側)と「何かが通り抜ける」と解釈できるクラスタ(図1下側)といったように解釈することができ、その下のクラスタも何らかの形でクラスタが表す概念を解釈することができる。付与した解釈の正当性を議論することは難しいが、一定の解釈に基づいたクラスタの解釈ができるという意味で妥当なクラスタ構造を獲得できていると考えられる。

なお、図1には示されていないが、専門的内容を扱った用例に対して、明らかに誤って分類している場合があった(2例)。このようなことを防ぐために、分類者が分類を行う際に、同一の語義であるか否かだけでなく、理解できないという場合も想定すべきである。

次に、IPALとの語義分類を比較してみると、「{とげ, 歯, 毛}がぬける」の語義分類の点で差が現れている。これは、IPALでは、ヲ格の格要素がガ格の格要素の体の一部の場合、独立した述語素を規定

しており、「ぬける1」、「ぬける2」というように二つの語義に分類されるための生じる。意味解釈でこのような分類は有用である。しかし、図1では、同一語義確率が0.8であり、分類者が語義の違いとしては認識していない。

また、IPALでは語義として区別されていない慣用表現(例:「ぬけるような空」、「間が抜ける」)は、分類者は分類しているということがわかる。

### 5.2 「よぶ」に対する分析

「よぶ」も「ぬける」と同様に各クラスタに一定の解釈をつけることができた。しかし、「~とよばれる」(「名づける」の意)の部分では、能動態・受動態といった構文的な要素が語義の分類規準に混入する現象が見られた。用例の集合から格フレームを獲得する際には、深層格以外での規準が混じることになり、対策が必要である。

### 5.3 「わかる」に対する分析

「わかる」は、「ぬける」、「よぶ」と異なり、各クラスタに一定の解釈を下すことが難しい、という結果になった。これは、「わかる」自体の語義の差が少ないため、同一の語義か判断する際のばらつきが大きいためだと思われる。このことは、図2に示したとおり、葉ノードの数が23個と「ぬける」、「よぶ」と比較して多く、同一語義確率が高い段階でクラスタが合併することが少ないことからわかる。このような多義語に対しては、語義用例間の同一語義確率を計算した段階で、その分布を検証するなどの対処することを検討中である。

## 6 おわりに

本論文では、多義動詞の語義分類規準を獲得する手法として、複数の分類者の用例分類結果から、語義分類規準を獲得する手法を提案した。提案した手法を評価するために、日本語多義動詞の用例クラス

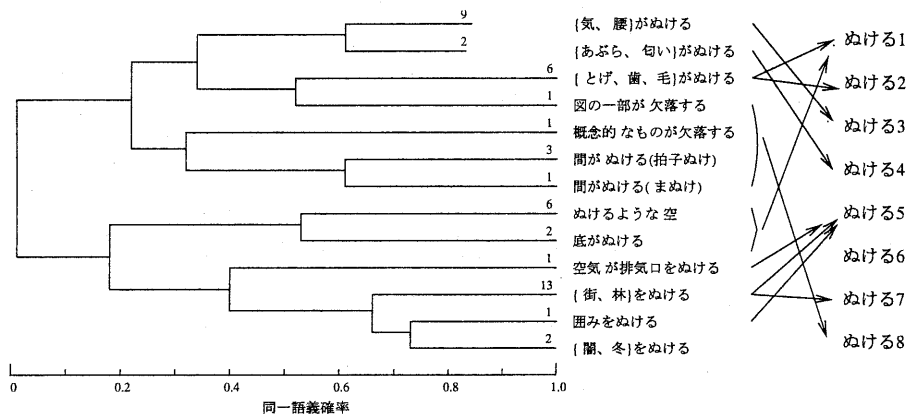


図 1: 「ぬける」のクラスタリング結果

タリングを行った。その結果、各クラスタに対する解釈をつけられるという点から、おおむね妥当な結果を得ることができたが、「わかる」のように語義の差が少ない、動詞に対しては何らかの対策が必要なのことがわかった。

## 参考文献

- [1] Nancy Ide, Jean Veronis: Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art, Computational Linguistics Vol.24 No.1, pp.1-40(1998)
- [2] Hinrich Schutze: Automatic Word Sense Discrimination, Computational Linguistics Vol.24 No.1, pp.97-123(1998)
- [3] Fumiyo FUKUMOTO, Jun'ichi TSUJII: Automatic Recognition of Verbal Polysemy, Proc. of COLING-94, pp.762-768(1994)
- [4] IPA Technical center, IPAL(Basic Verbs) dictionary volume, IPA,1987
- [5] 小泉 保ら: 日本語基本動詞用法辞典, 大修館書店 (1989)
- [6] 宇津呂武仁, 松本裕治, 長尾眞: 二言語対訳コーパスからの動詞の格フレーム獲得, 情報処理学会論文誌, Vol. 34, No. 5, pp.913-924 (1993)
- [7] Takiguchi, N., Xie, J. and Kotani, Y.: Acquisition of Semantic Feature on Case Frame Structure, Proceedings of NLPRS'91, pp. 337-344 (1991).
- [8] 鈴木孝夫: 日本語と外国語, 岩波新書 (1990)
- [9] 河原大輔, 黒橋禎夫: 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獲得, 情報処理学会研究報告 NL-140-18, pp.127-134(2000)
- [10] 新村出編: 広辞苑 (第五版), 岩波書店 (1998)
- [11] 小学館編: 国語大辞典 (新装版), 小学館 (1988)

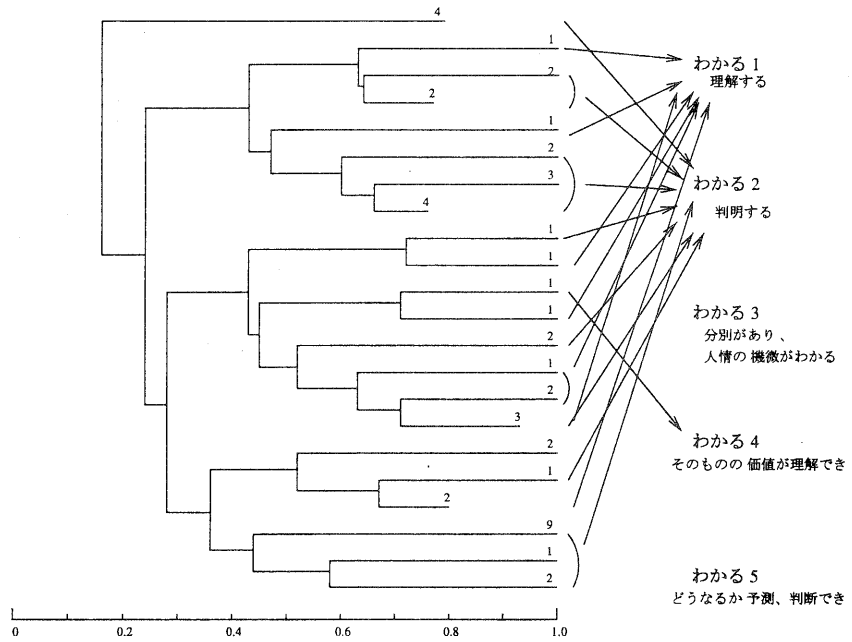


図 2: 「わかる」のクラスタリング結果

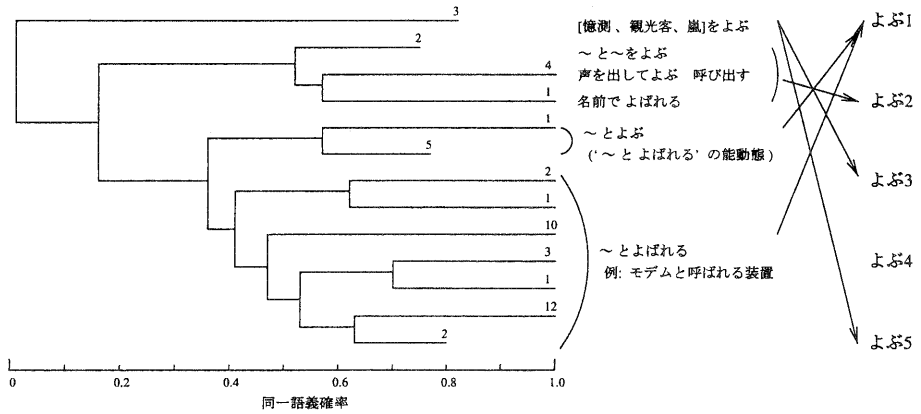


図 3: 「よぶ」のクラスタリング結果