社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2004－NL－164（20）
2004／11／5

# 料理教示発話の理解と作業構造の自動抽出

柴田　知秀†　　黒橋　禎夫†

† 東京大学大学院情報理工学系研究科　〒113-8656 東京都文京区本郷 7-3-1
E-mail: †{shibata,kuro}@kc.t.u-tokyo.ac.jp

**あらまし**　実世界情報，映像情報などの高度な利用のためには，その内容の構造的理解が必要であり，そのためには，話し言葉を現場を含めた広い文脈の中で正確に解釈することが重要になる．このような問題意識から，作業教示映像，具体的には料理番組映像を対象として，その発話（クローズドキャプション）の構文・格・省略・談話構造解析を行い，作業構造を自動抽出した．また，談話構造解析を頑健なものとするための知識として，用言の類義・共起関係をコーパスから自動獲得した．

**キーワード**　談話構造，話し言葉，映像情報，文脈，類義表現

# Understanding of Cooking Instruction Utterances and Automatic Extraction of Action Structure

Tomohide SHIBATA† and Sadao KUROHASHI†

† Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
E-mail: †{shibata,kuro}@kc.t.u-tokyo.ac.jp

**Abstract**　In realizing flexible utilization/access of real-word information or video contents, the crucial point is the structural analysis of their contents, which requires the interpretation of utterances based on wider contexts including the scene. This paper describes a method of analyzing cooking instruction utterances (closed caption texts) and extracting action structure. We also acquired synonymous expressions and co-occurrence relations from Web corpus to achieve robust discourse structure analysis.

**Key words**　discourse structure, spoken-language, video contents, context, synonymous expression

## 1. はじめに

計算機パワーの飛躍的進歩によって，実世界情報，あるいはそれを映し取った映像情報が大規模に蓄積・利用されるようになってきた [1]．映像という形態は，料理，園芸，工作，あるいは電化製品・電子機器の利用などを説明する／してもらう場合には非常に適した形態である．たとえば，料理番組を見るのと料理の本を読むのとでは（上級者は別にして一般には）明らかに前者の方がわかりやすい．

映像情報を高度に利用するには，それが何に関する映像であるか，何が映っているか，何が行なわれているかなどの情報を明示的に（記号として）付与する必要がある．これは現在のところほとんど人手で行なわれており，大変コストがかかるものとなっている．将来的に，そのような情報の自動付与を考えた場合，映像中のナレーションや登場人物の発話が大きな手がかりとなる．現在のところそれが決定的に難しいのは音声認識の問題であるが，それが徐々に解決されていくと，次に本格的に問題となるのは話し言葉の取り扱いであろう．すなわち，現在の，書き言葉中心に鍛えられた自然言語処理技術では，言語的文脈や現場の文脈に大きく依存し，また登場人物間でダイナミックにやり取りされる話し言葉を正確に取り扱うことは難しい．

このような問題意識から，映像情報中の話し言葉の解析を行った．具体的には，作業教示映像，特に料理番組映像を対象とし，当面は音声認識の問題をさけ，番組のクローズドキャプションを利用している．まず，作業教示発話を構文・格解析し，話し言葉で頻繁に起こる省略を解析する．さらに，各発話のタイプを解析し，それらの情報を統合して発話全体の談話構造を求め，作業構造を自動抽出した．

また，談話構造解析を行う上で問題となる表現のズレを吸収するために，「火を弱める」と「弱火にする」といった用言の類義表現をコーパスから自動獲得した．

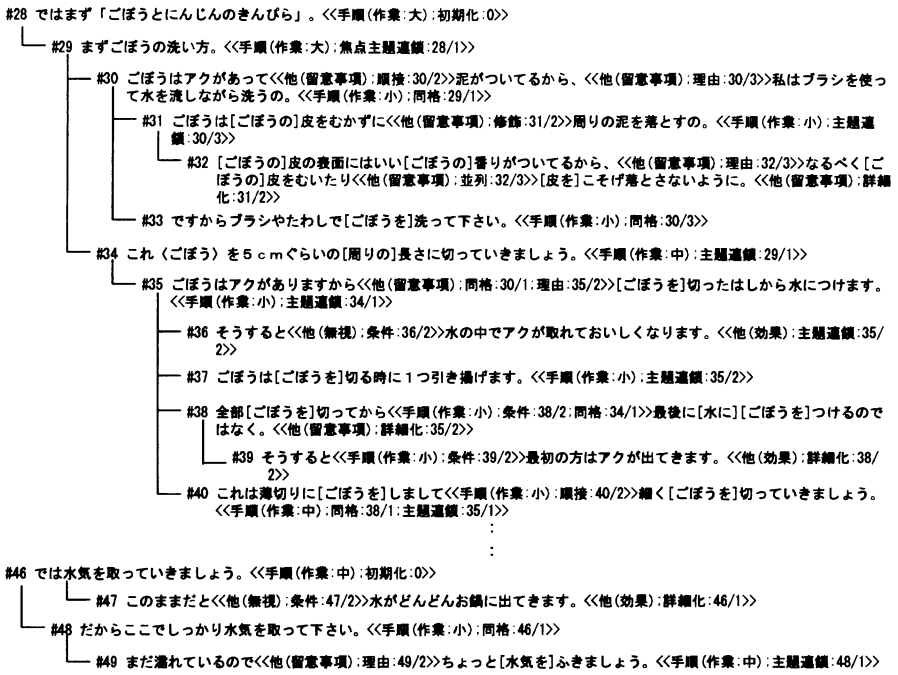


図１　料理番組のクローズドキャプションの解析例

## 2. 作業教示発話の解析

本研究で対象としている料理番組のクローズドキャプションの解析例を図１に示す．図において，文中の括弧 ([]) で示されたものは省略要素が補われたものであり，節末の括弧 (<<>>) は発話のタイプ，結束関係，親の節の文番号／節番号を示すものである．この例に示すように，話し言葉では主語，目的語などが頻繁に省略され，その一方で，説明が何度か繰り返されるという冗長性もある．また，作業の説明だけでなく，コツや注意点などの説明も含まれている．このような話し言葉の情報に対して，単純に単語マッチングのような検索を行っても，意図する映像を正確に取り出すことは難しい．

そこで，料理分野の「常識」に相当する知識を自動構築し，それを用いて発話中の明示されない関係の検出を行なう．さらに，各発話のタイプを解析し，それらの情報を統合して発話全体の談話構造を求める．これは，作業教示発話においては，作業の構造を反映したものとなっており，映像の要約や，検索システムのための構造化されたインデックスとして利用可能なものとなる．

### 2.1 格フレームの自動構築と関係解析

言語表現中には省略されている（明示されていない）関係が多数存在する．図１の例では「皮」は「ごぼうの皮」であり「切る」は「ごぼうを切る」である．このような省略は，書き言葉にもあるが，話し言葉ではより頻繁におこる．

このような関係を計算機によって検出するためには，「皮」というものは「動物や植物の皮」であり，「切る」ときには「何か切る」ということ，すなわち，各名詞，動詞に対して必須的に関係するものとしてどのようなものがあるかという常識的知識が必要となる．このような知識は格フレームと呼ばれている．

表１　動詞格フレームの例

| 用言 | 格 | 用例 |
|---|---|---|
| 切る (1) | ガ | 【主体】 |
| | ヲ | 豚肉，大根，こんにゃく，・・・ |
| | ニ | 正方形，櫛形，三角形，・・・ |
| 切る (2) | ガ | 【主体】 |
| | ヲ | 水気，水分，汁気，・・・ |
| | ノ | なす，豆腐，肉，・・・ |
| 入れる (1) | ガ | 【主体】 |
| | ヲ | 塩，油，野菜，・・・ |
| | ニ | 鍋，ボール，容器，フライパン，・・・ |
| 入れる (2) | ガ | 【主体】 |
| | ヲ | 包丁・・・ |
| | ニ | 魚，腹，付け根，・・・ |

名詞，動詞は少なくとも数万，専門分野での特有の用法などを考えるとそれ以上になり，格フレームを人手で与えることはとてもできない．これに対して，我々は，国語辞典やコーパス（大規模なテキスト集合）から格フレームを自動構築する手法を提案している [3], [4]．例えば，「皮」の国語辞典の定義文は「動物や植物などの外がわをつつんでいるもの」であり，また具体的に「レモンの皮」「みかんの皮」などがコーパスで頻繁に出現することから，「皮」という名詞は「動物か植物」の皮である

| 手順 | その他 |
|---|---|
| [作業:大]<br>・さ，では，ステーキの材料にかかります．<br>[作業:中]<br>・強火で油を温めましょう．<br>・じゃあ炒めていきましょう．<br>[作業:小]<br>・お鍋にお水を入れます． | [留意事項]<br>・最初に肉をパラパラに炒める事がポイントです．<br>[代替可]<br>・もし半個ぐらいでしたら，手で搾って頂いても結構です．<br>[食品・道具提示]<br>・材料は，牛ひき肉，百五十グラムです．<br>[雑談]<br>・暑くなってきましたね．<br>[効果]<br>・そうすると最初の方はアクが出てきます． |
| [料理状態]<br>・ニンジンの水分がなくなりました． | |

図 2　発話タイプの分類

ことが自動的にわかる．同様にして，コーパスの「切る」の用例を収集し，うまくクラスタリングすることによって「切る」には「豚肉，大根」などを切ることと「水気，水分」などを切ることがあるとわかる．表 1 に自動構築した動詞格フレームの例を示す．ここでは，コーパスとして Web から「料理」などのキーワードによって収集した 280 万文を用いた．

このような格フレームを用いて，まず，必須的な要素が欠けていることを認識し，次に，そこに何が補われるべきであるかを文脈中から探し出す．これは，いま注目しているものに構造的に近いものから順にチェックし（単純に単語数などではなく，構文解析の結果の木構造中の距離において近いものから），格フレームの用例と十分に類似しているものであれば補われるべきものと解釈する．語と語の類似性については，大規模なシソーラスが存在し，その中での近さから計算することができる [5]．

このような関係の解析の精度は 65%程度である．

### 2.2　発話タイプの解析

作業教示発話では，基本的には作業が順をおって説明されるが，中にはコツ，注意点や，雑談のような発話もある．これらのタイプを正確に認識しておくことは談話構造の解析のためにも，また検索のためのインデキシングとしても重要である．

発話のタイプは節ごとに考える(注1)．ただし，「〜しながら」「〜しつつ」などといった副詞的で，他の節に包含されるような節は分割しない．

[6] を参考にし，発話タイプを料理の「手順」に関係するかどうかという観点で分類し，また手順に関係するものはさらに作業に関係するかものか料理の状態を説明するものかに分類した．図 2 に発話のタイプと例を示す．

作業:大，作業:中，食品・道具提示，代替可，留意事項，効果，雑談については節末の表現のパターンを記述することで認識することができる．例えば，代替可については「〜しても結構です／構いません／よい」など．留意事項については「〜できます」「〜しやすいです」「〜を目安にしてください」などのパターンである．パターンマッチングのシステムは，形態素を単位として記述し，各形態素について語，品詞，活用形，シソーラスの意味素性をチェックする能力をもつものである．

作業:小，料理状態については，料理ドメインに対してすべての述語を列挙するという方法も考えられるが，他のドメインへの移植性を考え，自動詞，形容詞＋「なる」などを料理状態，それ以外を作業:小とする一般的な規則を用いている．このような規則を約 500 個記述したところ，発話タイプの分類精度は約 90%であった．

### 2.3　談話構造解析

談話構造のモデルとして，2.2 節で述べた節を一つのノードとし，関係するノードがリンクされたグラフ構造を考える．談話構造の解析の様子を図 3 に示す．

節間の関係として以下に示すものを考える．

（1）　一文内における節の係り受け関係

（2）　任意の 2 つの節における同格の関係 (同一文内に限らない)

（3）　主節 (「〜と思う」などといった節を除いた，一文で最後の節) 間の関係

以下に談話構造解析の手順を示す．

まず，一文内で係り受け関係にある節間の結束関係を決定する．付与する関係は以下のものであり，括弧内に示すような表層パターンにより決定される．

- 順接 (〜て，(連用形))
- 並列 (KNP で並列構造とされたもの)
- 理由 (〜から，〜ので)
- 条件 (〜と，〜たら)
- 修飾 (〜ずに)

次に，任意の 2 つの節において，用言の原形とその格要素が一致するかどうかをチェックし，同格の関係にあるものを検出する．チェックする格要素は，他動詞の場合はヲ格を，自動詞の場合はガ格とする．

例えば，30 文目の第 3 節の「ごぼうを洗う」と 33 文目の第 1 節の「[ごぼうを] 洗う」や，30 文目の第 1 節の「ごぼうはアクがあって」と 35 文目の第 1 節の「ごぼうはアクがありますから」が同格の関係にあることが検出される．

最後に，主節間の関係を求める．談話構造の初期状態として初期節点を考え，初期節点に接続することは，その発話から新しい話題が始まることを意味し，この時の関係を “初期化” とする．初期化以外の主節間の結束関係としては，並列，対比，理由，条件，主題連鎖，焦点主題連鎖，詳細化，理由，原因結果，例示，質問応答などの関係を考える．主節間の関係は，種々の

(注1)：「(〜と) 思う」などの節は無視した

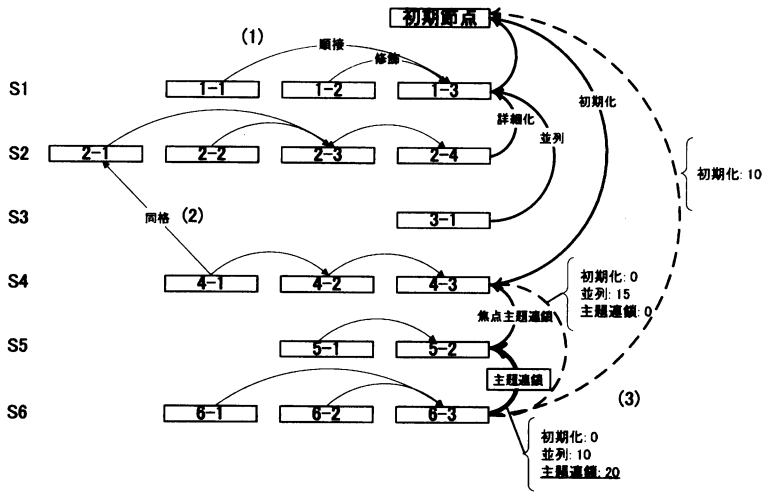


図 3　談話構造のモデル

表 2　談話構造解析のルール

| 結束関係 | スコア | 適用範囲 | 接続可能文パターン | 入力文パターン |
|---|---|---|---|---|
| 並列 | 5 | 1 | ～ | そして～ |
| 並列 | 40 | * | [並列] | ～さらに～ |
| 対比 | 30 | 1 | ～ | むしろ～ |
| 詳細化 | 30 | 1 | ～ | すなわち～ |
| 詳細化 | 15 | 1 | ～ | <留意事項> |
| 焦点主題連鎖 | 25 | 1 | <食品提示> | <個別作業> |
| 理由 | 30 | 1 | ～ | ～からだ |

表層的手がかりをもとに，各入力文に対して，関係をもつ以前の発話 (接続文) とその間の結束関係を逐次的に求める [7]．新しい話題が導入された後に古くなった話題に接続することはないという仮定をおき，入力文は談話構造の一番最後の子供の発話にのみ接続可能と考える．そして，さまざまな接続可能文との間のさまざまな結束関係を考慮し，最終的に最も高い確信度を得た関係を採用する．

確信度は表 2 に示すようなルールにより決定される．表 2 において，適用範囲とはどれだけ離れた発話との関係まで考えるか，接続可能文パターン，入力文パターンは，それぞれに対する表層表現，節間の結束関係 ([ ] で括られたもの)，発話タイプ (<>で括られたもの) などのパターンである．ルールが一致した場合には，結束関係欄の関係に対して，スコア欄の点数が与えられる．

談話構造解析の結果の具体例は図 1 に示したものである．

次節では，談話構造解析をより頑健なものとするための知識として，用言の類義・共起関係を自動獲得することについて述べる．

## 3.　用言の類義・共起関係の自動獲得

前述したように，話し言葉では何度か同じ内容の発話が繰り返されるという冗長性がある．2.3 節では，2 用言において，用言の原形とその格要素が一致するかどうかをチェックすることにより，繰り返しの発話を検出したが，そのようなものだけでなく，「水につける」と「水にさらす」，「弱火にする」と「火を弱める」などといった表現の類義関係を認識する必要がある．

また，2.3 節での談話構造解析では，関係をボトムアップに求めるものであったが，例えば，「洗う」「５ｃｍぐらいの長さに切る」「水につける」「細く切る」といった用言の共起関係をトップダウンに利用することで，談話構造解析をより頑健なものとすることが考えられる．

このような用言の類義・共起関係をコーパスから自動獲得することを考える．類義関係については国語辞典の定義文といったリソースを使うことも考えられるが，ここでは 2 用言における前後の用言の分布の類似度に着目した．

### 3.1　用言共起ペアの収集

まず，以下の手順で連用修飾関係にある 2 用言 (用言共起ペアと呼ぶ) を収集する．

（1）　Web から自動収集した料理テキスト 280 万文を構文・格解析する．

（2）　連用修飾関係にある 2 用言のペアを格フレーム単位で抽出する．2. 章で述べた解析を行い，用言のタイプ，用言間の関係のタグをふる．

例えば，「塩少々を加え，弱火にして下さい」という文を構文・格解析しタグを付与することにより，「加える:動 1(作業:小) - する:動 9(作業:小) < 順接>」という用言共起ペアを得る(注2)．

### 3.2　用言の共起関係の自動獲得

得られた用言共起ペアをまとめることにより，用言の共起関係を獲得した．用言の共起度はダイス係数を用いて計算した．得られた用言の共起関係を表 3 に示す．

(注2)：用言の後の "動 1" などといった表記は各用言の格フレームの ID を，括弧内は発話タイプを，山形括弧内は節間の関係を示す．

表 3　用言の共起関係

| 格フレーム 1 | 例 | 格フレーム 2 | 例 | ダイス係数 |
|---|---|---|---|---|
| 冷やす:動 1 | 冷蔵庫で冷やす | 固める:動 0 | 黄身を冷蔵庫で固める | 0.00693 |
| シェイク:動 1 | 材料をシェイクする | 注ぐ:動 1 | グラスに注ぐ | 0.00582 |
| 水洗い:動 0 | 豆を水洗いする | 切り落とす:動 1 | 根元を切り落とす | 0.00402 |
| つける:動 4 | 水につける | もどす:動 0 | 肉をもどす | 0.00315 |
| 溶く:動 2 | 卵を溶く | ほぐす:動 0 | 身をほぐす | 0.00298 |
| 干す:動 0 | 干す | 戻す:動 1 | 水で戻す | 0.00278 |

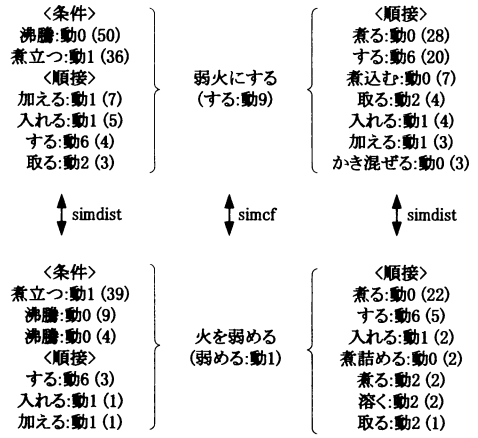


図 4　用言の前後の分布

### 3.3　用言の類義関係の自動獲得

さらに用言の類義関係を自動獲得した．用言共起ペアを各用言についてまとめると，図 4 のような分布が得られる．図中の () は用言の頻度を表している．2 用言の類似度を，用言の前後の分布の類似度と格フレームの類似度で決定する．

2 用言 ($p_1$,$p_2$) の前後の分布の類似度はベクトル空間モデルで定義した．例えば，図 4 において，「弱火にする (する:動 9)」と「火を弱める (弱める:動 1)」の前の分布の類似度は以下のように計算される．

$$simdist(p_1(\text{前}), p_2(\text{前})) = \frac{\sum_i a_{i,p_1} a_{i,p_2}}{\sqrt{\sum_i a_{i,p_1}^2 \sum_i a_{i,p_2}^2}} = \frac{50*9+36*39+\cdots}{\sqrt{(50^2+36^2+\cdots)(39^2+9^2+\cdots)}}$$

また，格フレームの類似度 $simcf$ は [4] で定義されているものを用い，用例パターン間の類似度で計算する．

コーパスから得られた類似度の高い用言を表 4 に示す．なお，2 用言の類似度を $simdist(\text{前}) + simdist(\text{後}) + simcf/2$ で定義した．

用言単位ではなく，格フレーム単位で結果を収集することにより，用言の多義性が解消されており，良好な結果が得られているといえる．誤って抽出された原因の一つに，各用言に対して格フレームが多い傾向にあることがあげられる(注3)．現在のシステムでは，格フレームをクラスタリングする際の類似度の閾値が 0.9 としており，ここで用いた前後の分布の類似度を利用してさらに格フレームのクラスタリングをすることによりこの問題に対処することを考えている．

このようにして自動獲得した類義表現を利用して談話構造解析を行なった．類義関係と判断する条件を，前，後のいずれかの分布の類似度が 0.3 以上，かつ，格フレームの類似度が 0.7 以上とした．例えば，図 1 の 48 文目の「水気を取る」と 49 文目の「水気をふく」において，前の分布の類似度が 0.395，後の分布の類似度が 0.292，格フレームの類似度が 0.982 となっており，類義表現と認識することができた．

作業の構造を理解するためには，用言の基本的な共起・類義関係に関する知識が必要であり，そのような知識を利用して作業構造の理解を高度にすることができる．そして，その結果からさらに用言の共起・類義関係を獲得するといったサイクルで共起関係の抽出，作業構造の理解をともに高度にできる．本研究では，その第一歩として，基本的な共起・類義関係を獲得し，それを利用して談話構造解析を行ったが，さらに，その結果から作業列を取り出し，それらを汎化，クラスタリングすることにより，さらに共起関係を抽出する予定である．

また，得られた類義表現は，以下で述べる料理映像検索システムにおいて同義表現辞書に登録し，ユーザからのクエリとデータベースとして用いたクローズドキャプション間の表現のズレの解消に役立てている．

## 4.　料理映像の要約と料理映像検索システム

図 1 に示した談話構造解析の結果 (28 文目から 40 文目まで) から，各段の最後の発話で，作業に関するもの (作業:大, 中, 小) を取り出し，そこから主節の用言とその格要素，修飾部を抽出すると，次のように主要な作業からなる要約を得ることができる．

（1）　ごぼうを洗う．
（2）　ごぼうを５ｃｍぐらいの長さに切る．
（3）　ごぼうを切ったはしから水につける．
（4）　ごぼうを細く切る．

このような談話構造解析結果や主要作業の表示機能を備えた，料理映像の自然言語検索システムを作成した（図 5）．

(注3)：「炒める」は 8 個,「切る」は 14 個ある．

表 4　2 用言の類似度

| 格フレーム 1 | 例 | 格フレーム 2 | 例 | 前の分布の類似度 | 後の分布の類似度 | 格フレームの類似度 | 2 用言の類似度 |
|---|---|---|---|---|---|---|---|
| 炒める:動 1 | 肉を油で炒める | 炒める:動 2 | ご飯をフライパンで炒める | 0.940 | 0.825 | 0.777 | 2.154 |
| 混ぜる:動 0 | ご飯に粉を混ぜる | 混ぜ合わせる:動 0 | 材料を混ぜ合わせる | 0.984 | 0.613 | 0.911 | 2.053 |
| 加える:動 1 | 塩を加える | 入れる:動 1 | 鍋に油を入れる | 0.720 | 0.696 | 0.969 | 1.901 |
| 整える:動 1 | 味を整える | 調える:動 1 | 味を調える | 0.814 | 0.588 | 0.985 | 1.895 |
| 沸かす:動 1 | 湯を沸かす | 沸騰:動:C1 | 湯を沸騰させる | 0.621 | 0.773 | 0.980 | 1.884 |
| 熱する:動 1 | 油を中火で熱する | 溶かす:動 1 | フライパンにバターを弱火で溶かす | 0.498 | 0.906 | 0.913 | 1.861 |
| 熱する:動 0 | フライパンに油を熱する | 沸騰:動:C0 | 湯を強火で沸騰させる | 0.948 | 0.472 | 0.865 | 1.853 |
| 混ぜる:動 1 | 塩を混ぜる | 混ぜ合わせる:動 1 | 調味料を混ぜ合わせる | 0.824 | 0.560 | 0.931 | 1.850 |
| する:動 9 | 弱火にする | 弱める:動 1 | 火を弱める | 0.742 | 0.833 | 0.528 | 1.839 |
| 温める:動 0 | 電子レンジ温める | 熱する:動 0 | フライパンに油を熱する | 0.762 | 0.581 | 0.925 | 1.806 |
| つける:動 6 | 両面に小麦粉をつける | つける:動 9 | 衣をつける | 0.638 | 0.803 | 0.712 | 1.797 |
| 散らす:動 1 | 上にパセリを散らす | 飾る:動 1 | 上にパセリを飾る | 0.878 | 0.440 | 0.954 | 1.795 |
| まぜる:動 0 | 全体をまぜる | 熱する:動 0 | フライパンに油を熱する | 0.907 | 0.465 | 0.820 | 1.782 |
| 加える:動 2 | 水を加える | 入れる:動 1 | 鍋に油を入れる | 0.732 | 0.728 | 0.564 | 1.742 |

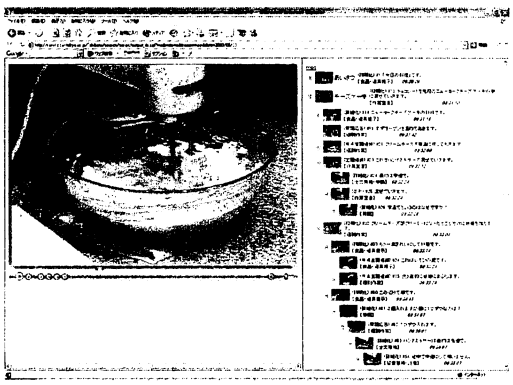

図 5　料理映像検索システム

## 5. おわりに

これまでの自然言語処理研究は，書き言葉の取り扱いが先行する形で進められ，機械翻訳，検索エンジンなど一定の成果がえられた．また，その中で常識的な知識の自動獲得などが少しずつ可能となってきた．

今後は，映像処理技術や音声認識技術の高度化に伴い，人間にとってより身近である話し言葉の解析が重要になっていくものと考えられる．本稿では，作業教示映像を対象として，その発話理解を行なった．

今後，言語解析を高度にするとともに，言語解析と映像解析の統合の研究を進める予定である．われわれはすでにショット・シーンといった映像の構造を利用することにより，談話構造解析の精度を向上させる研究を行っている [2]．映像情報をさらに利用するには映像中の物体認識が不可欠であるが，現在の画像処理技術では強い作り込みを行わない限り相当に難しい．そこで，発話タイプや談話構造といった言語解析を手がかりとして，映像中の物体認識に取り組む予定である．

## 文　献


[1] Rosenfeld Azriel, David Doermann, and Daniel DeMenthon: *Video Mining*, Kluwer Academic Publishers, 2003.

[2] Tomohide Shibata, Masato Tachiki, Daisuke Kawahara, Masashi Okamoto, Sadao Kurohashi, and Toyoaki Nishida: Structural Analysis of Instruction Utterances using Linguistic and Visual Information, In Proceedings of Eighth International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES2004), pp.393-400, Wellington, New Zealand (2004.9).

[3] Ryohei Sasano, Daisuke Kawahara and Sadao Kurohashi: Automatic Construction of Nominal Case Frames and its Application to Indirect Anaphora Resolution, In Proceedings of the 20th International Conference on Computational Linguistics (2004.8).

[4] Daisuke Kawahara and Sadao Kurohashi: Japanese Case Frame Construction by Coupling the Verb and its Closest Case Component, In Proceedings of First International Conference on Human Language Technology Research (HLT 2001), pp.204-210, San Diego, California, (2001.3.18-21).

[5] NTT コミュニケーション科学研究所: 日本語語彙大系, 岩波書店, 1997.

[6] Hidekatsu IZUNO, Yuichi NAKAMURA, and Yuichi OHTA. Quevico: A framework for video-based interactive media. In *Working Notes WS-5 International Workshop on Intelligent Media Technology for Communicative Reality, PRICAI-02 (Seventh Pacific Rim International Conference on Artificial Intelligence)*, pp. 6–11, August 2002.

[7] Sadao Kurohashi and Makoto Nagao: Automatic Detection of Discourse Structure by Checking Surface Information in Sentences, In Proceedings of 15th COLING, Vol.2, pp.1123-1127 (1994.8).