自由発話の音声信号から語句候補を切り出し意味付けするアルゴリズムの精度評価

The evaluation of the accuracy of an algorithm which segments candidates for phrases from free speech and acquires their meanings

林口 円[1*]　北川 憲[2]　左 祥[1]　小野 広司[1]　岡 夏樹[1]
HAYASHIGUCHI Madoka[1]　KITAGAWA Ken[2]　ZUO Xiang[1]
ONO Kouji[1]　OKA Natsuki[1]

[1] 京都工芸繊維大学工芸科学研究科
[1] Graduate School of Science and Technology, Kyoto Institute of Technology
[2] 京都工芸繊維大学工芸学部
[2] School of Engineering and Design, Kyoto Institute of Technology

Abstract: We implemented an algorithm which recognizes syllable strings from free speech in a maze navigation task, detects pairs of repeated strings from them as target phrases for meaning acquisition, splits them into groups of similar strings, finds a situation which highly co-occurs with the members of each group, and attaches the situation to the group as its meaning. We report the result of an evaluation experiment on the accuracy of the grouping and the meaning acquisition.

1　はじめに

現在のエージェントの多くは，人間との対話においてあらかじめ入力された語句しか使用することができない．人間とのインタラクションを通じて，未知の語句の意味をエージェントに学習させることができれば，より円滑な対話ができるはずである．未知の語句をエージェントが獲得し，その語句の意味を学習するためには，連続した音声信号から語句候補を切り出して，その語句候補に意味を与える必要がある．

そこで本研究では，人がエージェントを音声で誘導する場面において，その発話の中から類似区間を検出し，それを語句候補として切り出し，切り出された語句候補に対して共起性に基づいて意味を付与するアルゴリズムを実装した．

続いて収集した音声データに対して，本アルゴリズムを適用し，語句候補切り出しの精度と，意味付与の精度を評価した．

*連絡先：京都工芸繊維大学工芸科学研究科情報工学専攻
〒 606-8585 京都府京都市左京区御所海道町
E-mail: m8622027@edu.kit.ac.jp

2　関連研究

2.1　自由発話からの意味学習

Roy[1] は，繰り返される音声と画像を結びつけることで語彙を獲得するモデルを提案したが，その中で，発話の繰り返しの検出は次のようにして行われた．まず入力音声から RNN と HMM によって最尤音素系列を求める．この音素系列の中で，画像入力と共起しているものを残し，それらの中から類似区間を検出した．そして，得られた類似区間に対して相互情報量を用いることで，その類似区間に属する画像と音声の関連性の高さを調べ，音声と画像を対応付けるということを行った．タスクの内容は，母親が自分の赤ちゃんに物を見せて，その物を赤ちゃんに教えるというもので，比較的文法が簡単で，短い発話の音声データが得られ，それを用いて解析を行った．

Yu と Ballard[2] は実験協力者に自分のやっている作業を説明させるというタスクにおいて，腕の位置や，目の動きといった情報から，実験協力者の行動の特定，注目している物体の抽出を行った．入力音声から RNN, TIMIT を用いて音素認識を行い，行動と注目する物体の意味ごとにカテゴリに分け，カテゴリと同時に出現する音素列を 1 グループとした．そして，各音素に特徴量を与え，音素間の Hamming 距離を取り，DP マッチングを用いることで，類似区間の検出を行った．最

後にカテゴリごとにクラスタリングを行い，EM アルゴリズムを用いることで語彙を獲得した．

これらの研究は，知覚や行動を単語と対応付けようとするものであり，他にも多くの研究が行われてきているが，言葉の社会的な働きとしての意味の獲得はあまり扱われてこなかった．

これに対して，鈴木ら [3] は，行動の指示と行動の評価という異なる社会的働きを持つ言葉が混在する中での意味学習を目指した．鈴木らは，迷路を移動するロボットを，自由発話ではないあらかじめ決めておいた何通りかの教示により誘導するタスクを設定し，ある状況でどのような行動が適切であるかの知識をロボットが利用できることを前提として，ロボットによる教示意味の学習モデルを提案した．

我々も，鈴木らと同様に，行動の指示と評価のような異なる社会的働きを持つ言葉が混在する中での意味学習を目指すが，自由発話された音声信号からの語彙獲得を目指す点で異なる．我々の方法では，自由発話を人手で書き起こした語句に対して，鈴木らの研究よりも高い意味付与の精度が得られている [4]．

3 語句候補の切り出し

我々は以下の手法により，新規獲得語句の候補となる音声中の類似区間の検出を行った (図 1)．

1. 入力音声中，300ms 以上の無音部分で挟まれた部分を 1 発話として，Julius を用いて音節列へと変換する (図 1 の①)．この新しい音節列を入力音節列と呼ぶ．
2. 入力音節列と (過去の音節列＋入力音節列自身) との組合せ全てについて，Confusion Matrix から生成したスコア行列を用いた Smith-Waterman のアルゴリズム [5] によるローカルアラインメントを行い，設定した閾値（本研究では 0.3）を超えた部分 (2 文字以上の音節列) について類似区間と見なし検出結果に追加する (図 1 の②)．
3. 入力音節列を過去の音節列群に追加する (図 1 の③)．
4. 以下同様に 1. から 3. を新しい発話が入力されるごとに繰り返す．
5. 以上で得られた語句候補から類似グループを作成する．

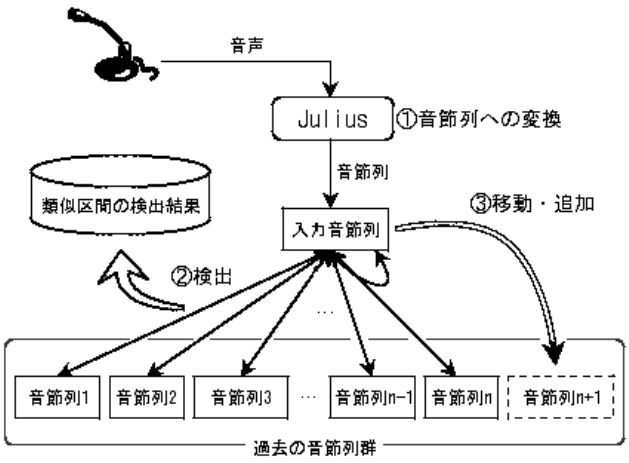

図 1: 類似区間検出手法

3.1 Smith-Waterman のアルゴリズム

入力された音声は Julius によって音節列へと変換され，Smith-Waterman のアルゴリズムと Confusion Matrix から計算したスコア行列を組み合わせることで，近似文字列マッチングを実現し，それによって類似区間の検出を行った．

Smith-Waterman のアルゴリズムは，動的計画法によるローカルアラインメントを行う．具体的には，文字列 $A = \{a_1, a_2, \cdots, a_m\}$ と $B = \{b_1, b_2, \cdots, b_n\}$ に対して，次のようにして類似する区間を求める．

1. $(m+1) \times (n+1)$ 行列 H を作成し，初期値として

$$H_{i0} = H_{0j} = 0 \quad (0 \leq i \leq m, 0 \leq j \leq n) \quad (1)$$

を与える．

2. 続いて，残りのノードのスコアを次の式に従って埋めていく (図 2)．

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + s(a_i, b_j) \\ \max\limits_{k \geq 1} \{H_{i-k,j} - W_k\} \\ \max\limits_{l \geq 1} \{H_{i,j-l} - W_l\} \\ 0 \end{cases} \quad (2)$$

3. 全てのノードについてのスコアの計算完了後，閾値を超えるスコアを持つノードからスコアがゼロになるノードまでを逆向きに辿り，類似区間を求めることができる．

ここで，$s(a_i, b_j)$ は文字 a_i と文字 b_j の類似度をもとにしたスコアである．本手法では，文字同士の類似度は後述の Confusion Matrix から計算する．

W_k および W_l は挿入・削除のペナルティ(ギャップペナルティ) であり，次の式で定義されるリニアギャップペナルティを用いる．

$$W_k = dk \quad (d \text{ は } -0.2 \text{ とした}) \tag{3}$$

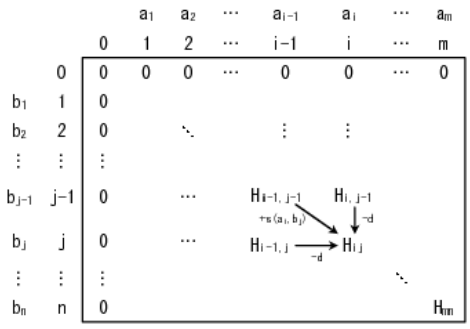

図 2: Smith-Waterman のアルゴリズム

3.2 Confusion Matrix

Confusion Matrix とは，あるクラス ω_i $(1 \leq i \leq n)$ に属するパターンが入力として与えられたとき，これがクラス ω_j $(1 \leq j \leq n)$ であると識別された回数を要素とする $n \times n$ 行列である [6]．我々は，音声波形中に音節 α(に相当する波形) が現れたとき，これが Julius によって音節 β と認識された回数を Confusion Matrix とした (図 3)．そして次のような対数尤度に基づいたスコアを求めた．

ともに長さ n の文字列 $X = \{x_1, x_2, \cdots, x_n\}$，$Y = \{y_1, y_2, \cdots, y_n\}$ があったとする．まず二つの文字列の間に何の関連性もないとした場合 (Random Model)，x, y の対応付けが起こる確率 $P(x, y|R)$ は，文字 a がある位置にランダムに現れる確率を q_a とすると

$$P(X, Y|R) = \prod_{i=1}^{n} q_{x_i} \prod_{j=1}^{n} q_{y_j} \tag{4}$$

となる．一方，二つの文字列の間に何らかの関連性が存在する場合 (Match Model)，文字 a と文字 b がある位置に同時に現れる確率 p_{ab} を用いて，x，y の対応付けが起こる確率 $P(x, y|M)$ を

$$P(X, Y|M) = \prod_{i=1}^{n} p_{x_i y_i} \tag{5}$$

と表すことができる．x，y の類似度 (文字列全体のスコア) はこれら二つの尤度の比

$$\frac{P(X, Y|M)}{P(X, Y|R)} = \prod_{i=1}^{n} \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \tag{6}$$

で定義される．さらに，対数をとることで加法性を持つスコアリング・システムが得られる．

$$S = \sum_{i=1}^{n} s(x_i, y_i) = \sum_{i=1}^{n} \log \left(\frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \right) \tag{7}$$

したがって，ある文字 a と文字 b に対するスコアは

$$s(a, b) = \log \left(\frac{p_{ab}}{q_a q_b} \right) \tag{8}$$

となる．

α \ β	ん	あ	い	う	え	お	か	…
ん	14620	105	381	323	98	114	6	
あ	12	2284	46	22	43	32	112	
い	57	30	10245	64	889	11	6	
う	23	7	47	1246	14	162	0	…
え	12	17	308	20	2919	1	1	
お	45	75	35	268	41	4684	3	
か	6	28	9	5	17	2	6210	
⋮				⋮				⋱

図 3: Confusion Matrix の一部．α が音声波形中の音節，β は変換後の音節

3.3 得られた語句候補から類似グループを作成する

以上より，入力音声データは，Julius を用いて音節列に変換され，語句候補の切り出し (図 4) が行われる．

次に，全ての語句候補を類似した音節列同士でグループに分けていく．これは後述する意味付与アルゴリズムにおいて，類似した語句でできたグループを同じ語句の集まりとして扱い，そのグループ一つずつに意味を与えるために行う．このグループ化は，すでに述べた近似文字列マッチングを利用して行う．

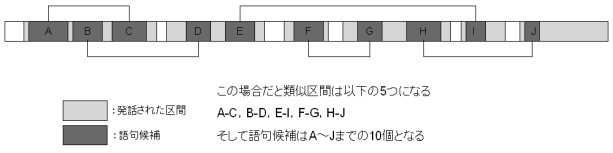

図 4: 類似区間検出と語句候補の切り出しの例

類似グループ作成の手順は以下の通りである (図 5):

1. すべての語句候補の中から二つを選び，近似文字列マッチングを行い語句候補同士の類似度を出す．
2. 1. をすべての語句候補の組み合わせについて行う．

3. 2. で得られた類似度で，最も高い類似度を持つ二つの語句候補のどちらかをグループの基準にする．

4. 基準となった語句候補と他の語句候補との類似度を見て，設定した閾値（本研究では 0.5）を超えた語句候補をグループとする．

5. 4. で得られたグループに入っている語句候補を次のグループ作成時の対象から除く．

6. グループに入れる語句候補がなくなるまで，3. に戻ってまたグループに分ける作業を繰り返す．

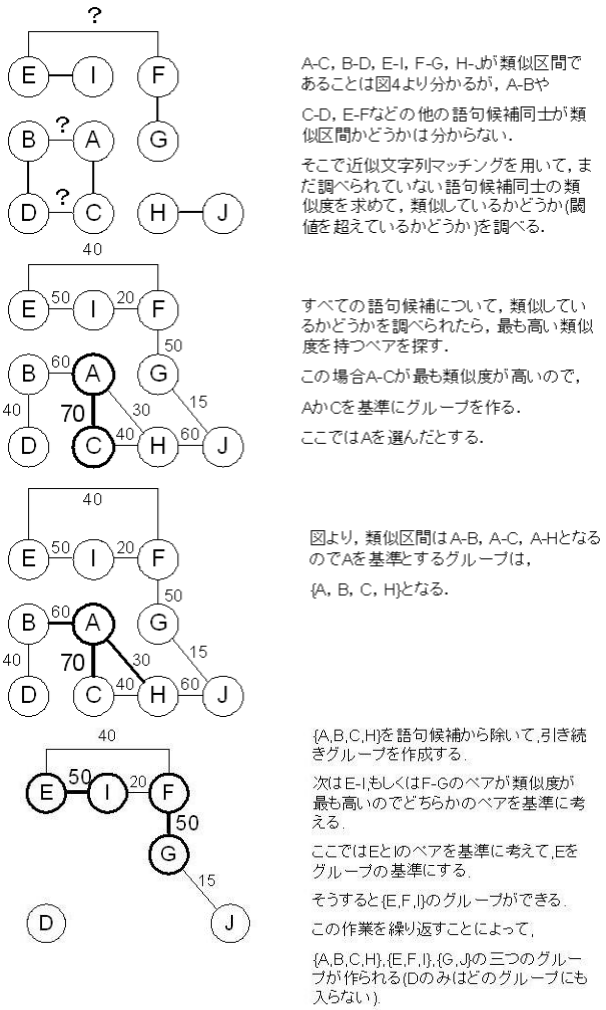

図 5: 類似グループの作成手順

4 語句候補への意味付与

1. グループ毎に状況との対応を表にまとめる．前節で作成された類似グループそれぞれについて，それぞれの語句候補がどのような状況で発話されたかを，全ての語句候補に対して調べ，表にまとめる．

2. 共起頻度に基づき意味を付与する．語句候補と状況の共起頻度に基づき，語句候補と状況を対応付け，対応付けされた状況を語句の意味とする．本研究では，迷路誘導タスクを用いて評価実験を行ったが，そこでの状況として,《次に取るべき行動は上／下／左／右への移動である》《直前の行動は適切（○）／不適切（×）であった》の 6 種類を想定した．対応付けのために，周辺度数（ある特定の状況やある状態の度数）から特定の組度数（ある言葉とある状況が共起する度数）が得られる確率を，Fisher の直接確率計算法を用いて算出する．詳しくは文献 [4] を参照されたい．

5 評価実験

5.1 評価実験に用いたデータ

PC 画面上の迷路において，教示者に，エージェントをゴールまで音声で自由に誘導してもらい，その様子をビデオで記録した．図 6 に迷路誘導ゲームの画面を示す．

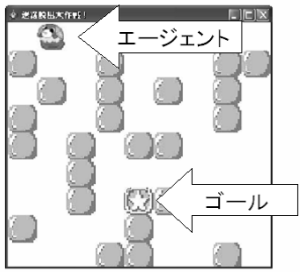

図 6: 迷路誘導ゲームの実行画面

5.2 評価実験に用いたツール

評価を行うにあたって音声データを手作業で書き起こしたものと，Julius で音節列に変換した語句候補との対応を見やすくするためにビデオ分析ツール Anvil[7] を用いた．(図 7)

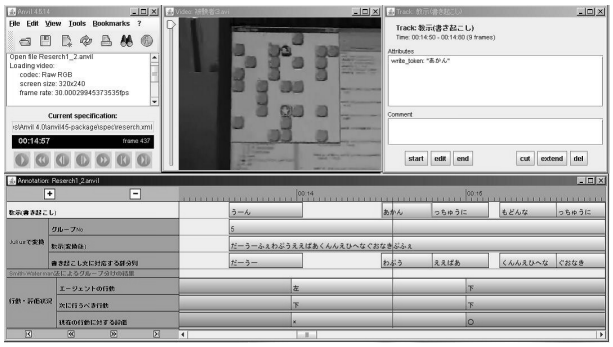

図 7: ビデオ分析ツール Anvil

5.3 結果

本研究では，語句候補を類似グループに分け，そのグループ毎に意味付与を試みた．そこで，グループ分けがどれだけうまくできたかという評価と，グループそれぞれへの意味付けがどれだけ正しくできたか，という 2 種類の評価を行った．

グループ分けについては以下の二つの点から評価した．

1. グループ中の頻度が最大の語句候補が，グループの要素の何%を占めるか (precision rate とする)
2. ある語句の最大何%が一つのグループに入ったか (recall rate とする)

グループそれぞれの語句候補の構成を調べた結果，平均で precision rate は 47.9 %（語句候補の数で重み付けしたグループ毎の precision rate の分布は図 8），recall rate は 36.1 %となった．

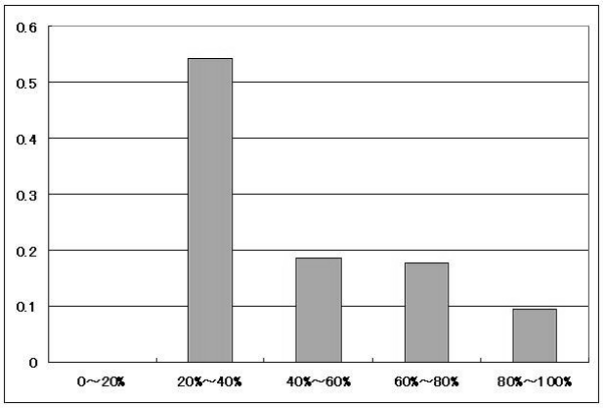

図 8: precision rate の分布

次に，意味付与の正しさについては，グループ分けされた語句候補の，各状況との共起度合いの分布と，意味付与した結果の分布とのカルバック・ライブラー距離を算出し，評価の指標とした．カルバック・ライブラー距離 d の度数分布（語句候補の数で重み付けしたグループの分布）は図 9 のようになった．

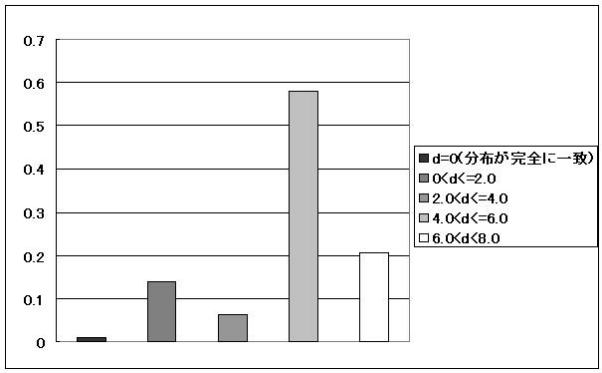

図 9: カルバック・ライブラー距離の度数分布

6 考察

6.1 実験結果からの考察

6.1.1 グループ分けの精度

図 8 の precision rate について見ると，20〜40 %にもっとも多く分布しており，80 %以上のものは全体の約 9 %である．平均的には一つのグループ内の約半数が書き起こし文の同じ文字列に対応していると言える．recall rate について見ると，一種類の書き起こし文は平均的には 3 つ以上のグループに分かれている．

このグループ分けの結果を意味学習に用いるため，理想的には，一つのグループは書き起こし文の一種類の文字列に対応する必要がある．そのためには，息や笑い声といった意味を与える際にノイズとなるような，余計な語句候補をグループに入れないようにするということと，同じグループ内にはなるべく一つの書き起こし文に対応する語句候補のみを集める必要がある．これらを実現する方法として，ノイズの入りにくい実験環境の整備や，グルーピングアルゴリズムの見直しが必要であると我々は考える．

6.1.2 意味付けの精度

意味付与の正しさについて見ると，カルバック・ライブラー距離 d が 2.0 以下であれば，図 10 のように分布の度数が最大のもの（この場合だと《直前の行動は不適切であった（×）》）が同じである可能性が高く，分布の度数が 0 のものは 0 に近かった．よってある程度正しく意味付けできていると考えられる．これは全

体の約 14 %だった．d が 4.0 より大きい場合は，図 11 のように分布の度数が最大のものが異なっていることが多く，意味付けを誤ったと言ってよいと考えられる．これは全体の約 79 %であった．

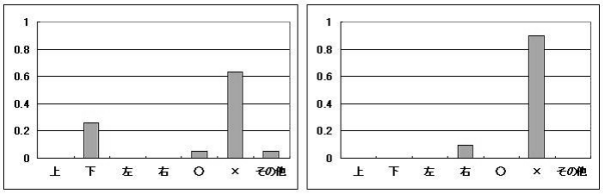

図 10: グループ内の語句候補と状況との共起割合の分布（左図）と意味付与した結果（右図）の一例 (d=1.65)

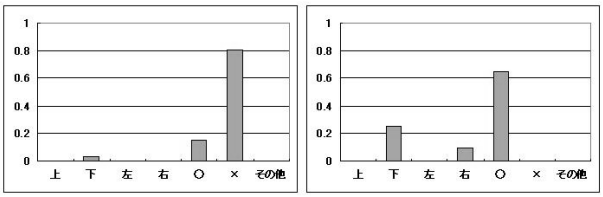

図 11: グループ内の語句候補と状況との共起割合の分布（左図）と意味付与した結果（右図）の一例 (d=5.11)

6.2 関連研究との比較

意味付けの精度に関しては，本研究と同じ実験データを用いて，人手で発話された内容を書き起こしその語句に意味を与えた場合，約 65 %の意味付けの精度が得られている [4]．これは本研究の意味付けの精度よりも高い．

本研究で意味付けの精度が悪かった原因として，まず一つ挙げられるのは，Roy[1] は実験時にヘッドホンと防音室を用いることで，ノイズを極力おさえていたが，我々は研究室のデスクトップ型 PC の前で，ビデオカメラによる録音を行っていたためにノイズが入りやすかったという音質の差がある．そして，Roy[1]，Yu と Ballard[2] は語句候補を検出する際に，画像，行動と同時に発話され，かつ同じ画像に対する語句候補毎でカテゴリに分けることで，余計な語句候補を省くということを行い，さらにカテゴリごとにクラスタリングを行っているが，我々はグループ分けを行う際のクラスタリングしか行っておらず，また，グループの大きさを決めるパラメータの大きさが固定であった等，クラスタリングの精度も悪かったということが原因であると考えられる．

7 おわりに

考察より，今後は，まず，音質の良いデータを取るために実験環境を整える，そして，語句候補を検出する際に，エージェントの行動状況，評価状況と発話をセットにしたものを，それぞれカテゴリに分け，さらに，グルーピングアルゴリズムの見直し，パラメータの調整を行いクラスタリングの精度を高めることにより，意味付けの精度を向上させることを目指す．

謝辞

本研究の一部は，科学研究費補助金基盤研究 (C) 17500093 の支援を受けた．

参考文献

[1] Roy, D.: “Learning from Sights and Sounds: A Computational Model”, Ph. D. Thesis, MIT Media Laboratory (1999)

[2] Yu, C. and Ballard, D. H.: “Learning to recognize human action sequences”, In IEEE Proceedings of the 2nd International Conference on Development and Learning, pp. 28-34, Boston, U.S. (2002)

[3] 鈴木 健太郎, 植田 一博, 開 一夫: “自律的な行動学習を利用した評価教示の計算論的意味学習モデル”, 認知科学 9(2), pp. 200-212 (2002)

[4] 岡 夏樹, 増子 雄哉, 林口 円, 伊丹 英樹, 川上 茂雄: “Fisher の直接法を用いたインタラクションデータからの意味学習”, 知能と情報, Vol. 20, No. 4, pp. 461-472 (2008)

[5] Smith, T.F. and Waterman, M. S.: “Identification of Common Molecular Subsequences”, J. Mol. Biol. Vol. 147, pp. 195-197 (1981)

[6] Parker, J. R.: “Rank and Response Combination From Confusion Matrix Data”, Journal of Information Fusion, Vol. 2, pp. 113-120 (2001)

[7] The video annotation research tool Anvil: http://www.anvil-software.de/