Vol.2010-NL-199 No.20
2010/11/19

視線情報を含むマルチモーダル協調作業対話コーパスの構築と利用

安原 正晃[†1] 石川 真也[†1]
飯田 龍[†1] 徳永 健伸[†1]

対話における非言語情報の重要性が認識され，韻律情報，ジェスチャー，視線などを含む，様々なマルチモーダル・コーパスが構築されてきた．我々は，2 名が協調して図形パズルを解く課題を与え，その時の発話に同期してパズル・ピースの位置情報，対話参加者の動作，さらに対話参加者の視線情報を記録したマルチモーダル・コーパスを構築した．特に日本語の協調作業対話における参照表現を研究するために対話中に出現するパズル・ピースを指示する参照表現を抽出し，その指示対象と所属性のアノテーションを人手でおこなった．本稿では，コーパスの構築と分析結果について述べる．

Construction and analysis of multimodal collaborative task dialogue corpora including eye-gaze information

YASUHARA MASAAKI,[†1] ISHIKAWA SIN-YA,[†1] IIDA RYU[†1]
and TOKUNAGA TAKENOBU[†1]

Having been recognised the importance of non-linguistic information in dialogue, there have been numerous attempts to construct multimodal corpora including prosody, gesture and eye-gaze information. We are developing a multimodal corpus collected from collaborative task dialogues where two participants solve geometric puzzles. As extra-linguistic information, the position of puzzle pieces, participants' actions and participants' eye gaze were recorded in synchronisation with utterances. For the purpose of studying human reference behaviour, referring expressions referring to puzzle pieces were manually annotated with their referents and attributes. This paper presents the details of the corpus construction and its analysis.

1. はじめに

言語理解における身体性の重要性が強調されて以来，ジェスチャー，韻律情報，視線情報などの非言語情報を利用した対話システムの研究がおこなわれている[16]．そのための基礎データとして様々な非言語情報を含むマルチモーダル・コーパスが作成されてきた[38),41)]．我々がこれまでに構築してきた協調作業対話コーパスも対話参加者の動作や操作対象の位置情報などの非言語情報が発話と同期して記録されている[59)]．本稿では，対話をおこなう課題については従来の設定を踏襲し，従来の非言語情報に加えて新たに視線情報を記録したマルチモーダル・コーパスの構築と利用について述べる．

認知科学の分野では，視線は人間の心的過程を反映する心への窓である[34)] という前提でさまざまな心理実験がおこなわれてきた[49)]．特に最近では視線計測技術の進展のおかげで被験者に過度の負担をかけることなく，容易に視線の計測が可能となったことから，視線情報と種々の認知過程の関係について急速に研究が進んでいる[21)]．たとえば，問題解決過程と視線の動きの関係の分析[20),25),39)] や言語理解や言語生成と視線の関係の分析[19),27),29),40),43),44),51),54),58)] などがおこなわれてきた．言語と視線との関係で言えば，とりわけ参照表現から対象物を同定する時の視線の動き，あるいは逆に指示対象を与えられてそれを指示する参照表現を発話する時の視線の動きなどに高い関心が払われている．たとえば，Meyer らは線画で描かれた 2 つのオブジェクトを被験者に提示し，それを “X and Y” という並列名詞句として発話する際の視線の動きを分析している[44)]．Bock らは被験者に時計を見せて，その時間を発話させ，その時の短針と長針に視線が固定されるタイミングを調査している[7)]．これらは名詞句の発話のプランニング，特に語彙アクセスのタイミングを調査するための実験であるが，Griffin らは犬が人間を追いかけている絵を被験者に見せて，その状況 (イベント) を文として発話させ，その時の視線の動きを分析している[26)]．これらの分析に共通した結論として，人がオブジェクトに言及する時，そのオブジェクトに視線を固定してから約 700 ～ 900m 秒遅れて，参照表現を発話するという結果が得られている．ただし，Kaur らの実験によれば，この「遅れ」は同一人物ではさほどばらつきがないものの，個人差が大きい (150m 秒 ～1,500m 秒) という報告もある．逆に参照表現を聞いて指示対象を同定する課題でも，指示対象に視線を固定するのに約 1,000m 秒を要するという報告がある[1)]．

†1 東京工業大学 大学院情報理工学研究科
Department of Computer Science, Tokyo Institute of Technology

Vol.2010-NL-199 No.20
2010/11/19

このような認知科学の知見を利用して，言語処理の観点からも言語理解や言語生成に視線情報を使う研究が始まっている．たとえば Campana らは，視線固定から発話までの遅れが約 900m 秒であるという Griffin らの実験結果[26] で得られた知見を利用し，対話システムの照応解析に視線情報を利用するインタフェースを提案している[12]．このインタフェースでは，言語情報を使った照応解析に失敗した時にバックアップとして視線を利用する．

対話研究，特に多人数会話 (multi-party conversation) の研究では，話者交替を制御したり[9),17),30),33)]，発話の受信者を同定するため[24),57)] に視線情報を利用しようという研究がおこなわれている．また，CG 合成されたアバターとの対話で視線情報を使う研究も始まっている[45),56)]．

Kelleher らは 3 次元の CG 画像をにおいてシーン中のオブジェクトの顕現性を計算するモデルを使った参照解析手法を提案している[35)–37)]．Kelleher のモデルでは視線情報は直接使用していないが，視線情報を取り込む可能性について言及している．Prasov と Chai は部屋の中を描いた静止画を見ながら対話する状況において，視線情報を利用することによって，参照表現の解析精度が向上する例を報告している[46)]．また，Qu と Chai の一連の研究では，対話における未登録語を理解するのに視線情報が有効であるという結果が報告されている[47),48)]．

認知科学の観点からの研究では，視覚刺激として実験を制御しやすいという理由で静止画を利用することが多いが，言語処理，特に対話研究の観点からは動的な状況における対話の中で視覚情報を扱うことが多く，視覚情報が動的に変化するという特徴がある．このような対話は「状況に依存する対話」(situated dialogue) と呼ばれている．認知科学の分野では動的に変化する状況で視覚情報をどう扱うかの分析は始まったばかりである[21)]．実際，Bard らは，静止画を使った Map Task[2)] における対話者の視線の分析[5)] から，最近では 2 名が協調して block-pattern copying task[28)⋆1] を解くといった動的な視覚情報を扱った領域に研究対象を移している[6)]．従来の認知科学の実験が一被験者に刺激を与え，視線情報を計測するという手法が主だったのに対し，Bard らのように，最近では 2 名の被験者が協調して課題を解く際の視線の相互関係を分析した研究結果も報告され始めている[8),50),52),55)]．

認知科学の研究では実験をおこないデータを収集し，それを分析するという研究手法が取られることが多い．収集したデータは公開されることもあるが，データに他の言語情報などを付与してコーパスとして整理することは少ない．これに対して言語処理の分野では，前述したとおり視線の含む種々の非言語情報を付与したマルチモーダル・コーパスの作成が盛んにおこなわれている[38),41)]．ただし，視線情報を含むマルチモーダル・コーパスは 20 年近く前からその可能性については指摘されていたが[60)]，その構築は緒についたばかりで，その数は多くない[6),10),11),13),15)⋆2]．

以上のような背景をふまえ，我々は，2 名の協調作業対話における対話参加者の発話，動作，視線を同期して記録した日本語のコーパスを構築している．特に対話中の参照表現に注目し，参照表現へのアノテーションをおこなっている．以下，2 節ではコーパス作成のためのデータ収集実験について述べ，3 節では，収集したデータからコーパスを構築する過程を特にアノテーションの詳細を中心に述べる．4 節では，このコーパスを使い，視線情報を利用した参照解析手法の予備実験について述べる．

2. データ収集

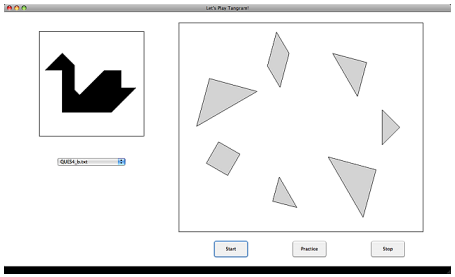
互いに知人関係にある 2 名を一組として大学生・大学院生から 2,000 円の報酬で 58 名の被験者を集め，2009 年 11 月から 12 月にかけてデータ収集実験をおこなった．各組には，与えられた図形パズルを協力して解くように指示した．用意した図形パズルは (1) タングラム，(2) ポリオミノ，(3) ダブル・タングラムの 3 種類である．被験者の発話，動作，視線を同期して正確に記録するために，コンピュータ・ディスプレイ上で簡単なマウス操作によってパズルピースの操作をおこなえる図 1 に示すようなシミュレータを実装した．シミュレータ画面は目標図形の提示領域 (左側) とピースを操作する作業領域 (右側) から構成されている．作業領域中のピースはマウス操作によって「移動」,「回転」,「裏返し」のパズルを解くために必要十分な操作がおこなえる．こららのパズルの目標は，いずれも作業領域中のピースを配置して，目標図形と同じ図形を作ることである．

1 組の被験者には「指示者」と「作業者」の 2 つの異なる役割を与えた．指示者は目標図形を構成するためのピースの配置を考え，各ピースをどのように操作するかを作業者に口頭で伝える．作業者は指示者の指示にしたがってピースをマウスによって操作する．両者は図 2 (a) のように各自のディスプレイの前に並んで座り，シミュレータの作業領域を共有しながら課題を解く．ただし，指示者 (図 2 (a) 左側) には目標図形は与えられるがマウスは与えない．一方，作業者 (図 2 (a) 右側) にはマウスは与えられるが目標図形は与えない．作業

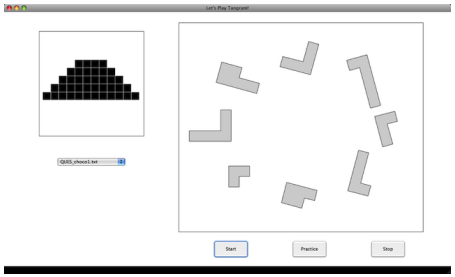
⋆1 モデルとして与えられたとおりにパズル・ピースを配置する課題.

⋆2 最初の AMI コーパス[13)] には視線情報は含まれないが，これを使った Frampton らの研究[24)] の中で視線情報が付与された．また，Byron らのグループの Quake コーパスは人間の視野は記録されているが，これは厳密な意味での視線ではない.

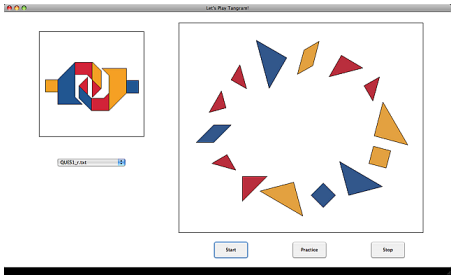
Vol.2010-NL-199 No.20
2010/11/19



(1) タングラム



(2) ポリオミノ



(3) ダブル・タングラム

図 1 パズル・シミュレータ

(a) 対話収録中の様子

(b) 頭部固定の例

図 2 実験環境

者が指示者のディスプレイの目標図形を盗み見ないように両者の間には衝立を設置した．つまり，両者のインタラクションは音声対話とシミュレータの作業領域の映像によっておこなわれる．対話の内容に関しては特に制限を設けなかった．操作者から質問をしたり，配置についての提案をすることも自由である．

この課題は古典的な block-pattern copying task[28] と似ているが，2 名が協調しておこなう点と単に配置をコピーするのではなく，配置を自分で考える問解題決の要素が入っている点が異なる．Foster らは，block-pattern copying task を 2 名で協調して解く課題に拡張して同じようなコーパスを収集しているが，適切な配置を考えるという問題解決の要素は入っていない[22]．また，Foster らの実験では両者が対等な関係で協調作業をおこなう設定になっており，言語表現を使わずに直接対象を各自が操作できる．より多くの参照表現を引き出すために，我々はこのような非対称な役割を与えた．

被験者を 4 つのグループに分け，各グループにタングラム (ヒントあり)，タングラム (ヒントなし)，ダブル・タングラム，ポリオミノのいずれかの課題を与えた．独力で問題解決をする際の振舞いを記録するためにタングラムではヒントなしのグループも設けた．同一グループ内の各組が解く問題はすべて同じで，1 組が解く問題数はタングラムとポリオミノでは 4 問，ダブルタングラムでは 6 問である．問題の半数を解いた時点で被験者は役割を交替させた．

視線位置の計測は Tobii 社の Tobii T60 Eye Tracker を用いた．T60 は，$1,280 \times 1,024$ ピクセルの液晶ディスプレイの下部に赤外線カメラを備えており，精度 0.5° で被験者の視線をディスプレイ上のピクセル座標として 1/60 秒毎に計測することができる．T60 から視線位置のピクセル座標をタイムスタンプと一緒に獲得するために視線計測プログラムを作成した．パズル・シミュレータから 1/65 秒毎に出力されるマウス操作やピース位置などの情報と Tobii から出力される視線位置の情報の時間同期を取るために，シミュレータと視線計測プログラムは同一のコンピュータで動作させ，すべてのタイムスタンプにはこのコンピュータの時間を使った．

被験者にはあらかじめ課題遂行中の視線を計測する旨伝え，課題を開始する前に Tobii 社のキャリブレーション・プログラムを使って 9 点による視線計測のキャリブレーションをおこなった．また，視線計測をするのでできるだけ頭部は動かさないように指示を与えた．シミュレータのマウス操作に慣れさせるために，本課題に入る前に目標図形の正解を与えた上でピースを配置する練習問題を解かせた．Tobii は頭部の移動に対してある程度の耐性はあるが，練習セッションの様子から過度に頭部を動かす被験者については図 2 (b) のように

頭部を軽くゴム紐で固定し，頭部を動かしにくくした．

各問題におけるピースの初期配置は毎回ランダムに配置した．1 問に使える制限時間としてタングラム，ポリオミノでは 15 分，ダブル・タングラムでは 10 分を設定し，これは事前に被験者に伝えた．パズルを解くために指示者が考え込んでしまい，発話がなくなってしまうことを避けるために，タングラム (ヒントあり) とポリオミノでは，途中で正解ピースの位置をヒントとして提示した[⋆1]．ダブル・タングラムの課題では正解の配置が最初からほとんど与えられているのでヒントは与えなかった．課題は目標図形が完成するか制限時間が経過すると終了する．

課題遂行中の 2 人の会話はヘッドセット・マイクを通してチャンネル分離してステレオ録音し，音声と同期してシミュレータから出力されるの各ピースの位置情報，およびすべてのマウス操作を 1/65 秒毎に記録した．視線位置は，Tobii から出力される 1/60 秒毎のディスプレイのピクセル座標を前述の視線計測プログラムによって記録した．

視線計測の技術が進展したとはいえ，視線を正確に安定して計測することは必ずしも容易ではない．Tobii は各計測時間において視線位置の座標と一緒に計測状態も出力することができる．状態は両眼計測成功，片眼のみ計測成功，エラーのいずれかである．収集したデータのうち対話時間の 40%以上で計測状態がエラーであった対話は破棄してコーパスに含めないこととした[⋆2]．各課題ごとの被験者の組数，実験条件，コーパスに収録した対話数をまとめたものを表 1 に示す．

表 **1** コーパスの概要

コーパス ID	パズル	問題数	組数	制限時間 [分]	ヒント	対話数
T2009-11	タングラム (ヒントあり)	4	10	15	5 分おきに 2 回	27
N2009-11	タングラム (ヒントなし)	4	5	15	なし	8
P2009-11	ポリオミノ	4	7	15	3 分おきに 4 回	24
D2009-11	ダブル・タングラム	6	7	10	なし	24

⋆1 タングラムのヒントは新しいヒントが出た時点で前のヒントは消えるが，ポリオミノのヒントは一度出たヒントは次のヒントが出ても消えない．

⋆2 Bard ら[5] はさらに厳しい 30%以上の基準を課しているが，本稿では収録対話数をできるだけ多くするために 40%としている．また，タングラム (ヒントなし) では，問題解決の分析もおこなうことを目的としているので，2 組分の対話をすべて収録することを優先したため，エラー率が 40%を越えるものも 1 つ (N01) 含まれている．

3. アノテーション

3.1 アノテーションの概要

実験によって収集したデータは，ビデオや音声などを時間同期してアノテーションできるアノテーション・ツール ELAN[4)⋆3] を用いてアノテーションをおこなった．まず，各対話のシミュレータの出力と視線計測プログラムの出力から，パズル・ピースに ID を付与した上ですべての操作を再現し，かつ 2 名の視線を重畳したビデオを作成した．このとき視線については平滑化するために 6Hz のローパスフィルタをかけた．音声とこのビデオを ELAN に読み込み，ビデオと音声を参考にして対話を書き起した．発話単位は ELAN に標準で用意されているセグメンテーション機能を用い，最小無音区間 400m 秒，最小発話区間 300m 秒で自動分割した．自動分割がうまく機能しない箇所は人手で修正した．また，ひとつの参照表現が複数の発話に分断されてしまった場合は発話をひとつにまとめ，参照表現が複数の発話にまたがらないように修正した．

表 **2** ELAN で管理するアノテーション層

層の名前	意味
OP-UT	操作者の発話
SV-UT	指示者の発話
OP-REX	操作者の使った参照表現
OP-Ref	その指示対象
OP-Attr	その属性
SV-REX	指示者の使った参照表現
SV-Ref	その指示対象
SV-Attr	その属性
Action	ピースに対する操作
Target	その操作の対象ピース
Mouse	マウスカーソルの位置
OP-GZE-P	操作者の視線停留点の中心座標
OP-GZE-N	操作者の視線停留点に一番近いピース
SV-GZE-P	指示者の視線停留点の中心座標
SV-GZE-N	指示者の視線停留点に一番近いピース

∗ 先頭の字下げはアノテーション層の親子関係を表わす．

次にシミュレータから出力されるピースの操作とマウス位置，視線計測プログラムから出

⋆3 http://www.lat-mpi.eu/tools/elan

力される視線位置の情報を ELAN のアノテーション層に加え，これらの情報を参考にしながら人手で発話中の参照表現のアノテーションをおこなった．ELAN で管理するアノテーションの一覧を表 2 に示す．この他にシミュレータの出力から得られる情報として各パズル・ピースの位置情報がある．ピースの位置は変化する都度，タイムスタンプと共に記録する．サンプリング・レートはマウス・カーソルの位置と同じ 1/65 秒である．ただし，パズル・ピースの位置は ELAN には取り込まず，別ファイルとして管理している．

3.2 ピース操作

ピースの操作は，操作中の時区間を Action 層に，その操作の対象となるピース ID を Target 層に記録する．Target 層は Action 層の子供として定義されている．操作の種類は，移動 (Move)，裏返し (Flip)，回転 (Rotate) のいずれかであり，このラベルを Action 層の時区間に付与する．ただし，裏返し操作に関しては，マウスをダブルクリックした瞬間に実行されるので，時区間を 1m 秒の固定幅としている．

3.3 マウス位置

シミュレータから出力されるマウス・カーソルの位置は，1/65 秒のサンプリング・レートですべて記録されているが，ELAN 上でアノテーションする際は，マウス・カーソルがいずれかのパズル・ピース上にある時区間のみを記録した．ここで，マウス・カーソルの中心から 8 ピクセル以内の正方形領域がピースの領域と重なる場合にマウスがピースの上にあると定義した．Mouse 層にはマウスがピースの上に乗っていた時区間がピースの ID と共に記録される．

3.4 視 線 位 置

視線位置は，Tobii T60 により 60Hz のサンプリング・レートで計測しており，各時刻におけるディスプレイ上の視線のピクセル座標とその有効性が利用できる．視線の動きには滑らかに移動する Smooth pursuit と瞬間的に移動する Saccade があり，Saccade 中は人間の認知活動が抑制されるという報告がある[42]．また，視線の停留は，視線の先への注意を示唆すると考えられていることから[51]，Tobii の出力データから停留区間を抽出することとした．ここでは，停留を，許容誤差の範囲に収まる連続した点の集合と定義する．許容誤差 D は，Tobii のカタログ仕様の測定誤差 (0.5°)，画面から被験者までの平均的な距離 (50cm)，および，画面解像度 $(1,280 \times 1024)$ から計算し，16 ピクセルとした．また，Richardson らの実験[51] にならい，視線がこの許容誤差内の領域に 100m 秒以上留まる時に停留するとみなした．

まず，停留となりうる区間を抽出する．停留開始時刻 (初期値：タスク開始時刻) から，停留を構成する点集合に連続する時刻の点を 1 つずつ加えていき，次の条件を満たしている限り停留が継続しているとする．

- 停留を構成するすべての点が，停留の重心 (構成するすべての点の平均) から距離 D 以内に存在する
- 計測エラー・レコードを含まない

これらの条件が満たされなかった場合，それまで条件を満足していた区間を停留とし，条件が初めて満足されなくなった時刻を新たな停留の開始時刻として，次の停留を探索する．これをタスク終了時刻までおこなう．

次に計測エラー・レコード区間を補完する．停留構成に必要な時間 (100m 秒) 未満の計測エラー・レコード区間は，その前後の停留区間の重心間の距離が距離 D 以内であれば，エラー・レコード区間とその前後の停留区間をつなげて 1 つの停留区間とする．これは，停留構成に必要な時間 (100m 秒) 以下のエラー・レコード区間では，前後の停留区間と異なる位置で停留が形成されるとは時間的に考えられないためである．また，これは，散発する短いエラーによって停留が細切れになってしまうことを防ぐ効果もある．

最後に停留構成に必要な時間 (100m 秒) 以上の停留のみを抽出し，区間の開始・終了時刻と構成点集合の重心のディスプレイ上のピクセル座標を用い停留を記録する (ELAN の *-GZE-P 層)．また，各停留区間において，どのピースに注目しているかを判断するため，停留の重心から各ピースまでの距離を計算し，一番近いピースの ID を*-GZE-N 層に記録する．ここで，ピースまでの距離は，停留の重心からピース外周までの最短のユークリッド距離とした．

3.5 参 照 表 現

参照表現のアノテーションは指示者 (SV)，操作者 (OP) ごとに参照表現区間 (SV-REX, OP-REX)，その指示対象 (SV-Ref，OP-Ref)，その参照表現の属性 (SV-Attr, OP-Attr) を各アノテーション層に付与した．*-Ref 層と*-Attr はいずれも*-REX 層の子供として定義されている．参照表現の認定基準は以下のとおりである．

- 原則として作業領域中のパズル・ピースを指示する名詞句を対象とする．したがって，目標図形を説明する発話中に出現する表現は対象としない．
- 対象ピースの同定に必要な言い誤りの訂正や情報の追加を含む場合は，訂正・追加部分も表現に含める．
- 同じピースを指す表現を連続して使っている場合は個別にアノテーションする．

- 参照表現が入れ子になっている場合は一番外側の名詞句のみを対象とする．[⋆1]
- 相手に対する発話ではなく，ひとりごと中の表現は対象としない．
- 発話途中で言い留まるなど，不完全な表現は対象としない．
- ピースの一部を指す表現は対象としない．

指示対象は 1 文字の英数字で表現されるピースの固有 ID の並びによって表わす．上述したように，ピースの ID はアノテーション用に作成したビデオ中では各ピース上に表示されている．複数のピースを指す表現ではピースの ID を並べるが，指示対象が不定の場合は可能性のあるピース ID を並べ，先頭に 0 を追加することによって表現する．たとえば，まったく同じピースが 2 つ (ID=1, 2) ある場合，ピース 1 とピース 2 の 2 つのピースを指示する表現には “12” を付与するが，このいずれを指すか不定の場合は “012” を付与する．

参照表現に付与する属性の一覧を表 3 に示す．これらの属性は，同様の課題を使って我々が過去に構築したコーパス[59] の属性を修正したものである．

表 3　参照表現の属性

記号	意味	例
dpr	指示代名詞	「これ」,「それ」,「あれ」
dad	指示形容詞	「その三角形」
pnn	形式名詞	「大きいの」
siz	大きさ	「大きい三角形」
col	色	「赤い三角形」
typ	タイプ	「四角形」
dir	方向	「左を向いてる三角形」
prj	投射型空間関係	「右の三角形」
tpl	位相型空間関係	「四角形の近くの三角形」
ovl	重なり	「大きいのの下の三角形」
act	動作	「さっき回転した三角形」
cmp	補集合	「もうひとつの三角形」
sim	類似性	「同じやつ」
num	数	「2 つの三角形」
rpr	修正	「右の大きい，いや小さい三角形」
err	誤り	「四角形」(三角形を指示して)
nest	入れ子	「四角形の右にある三角形」
meta	比喩	「足の部分」
nul	属性なし	(上記の属性が付与されない表現に付与するダミー)

⋆1 ELAN は同一層で入れ子のアノテーションを許していない．

3.6 アノテーションの評価

アノテーションの信頼度を評価するために，アノテーションした参照表現の区間と，それに付与した指示対象と属性のタグについて，一致度を計算した．T2009_11 (ヒントありタングラム) の 27 対話中 9 対話について，著者の 1 人 (A_1) がアノテーションした結果と，独立した 2 人のアノテータ A_2，A_3 がアノテーションした結果を比較した．重複しておこなったアノテーションは，(A_1, A_2) 組で 4 対話，(A_1, A_3) 組で 5 対話である．

複数のアノテーションの一致度には κ 係数が用いられることが多いが[14]，参照表現区間の一致度を計算するためには部分一致も考慮する必要があるため，本来範疇素性の一致度を計算するための κ 係数はなじまない．このため，本稿では β 係数[3] を用いる．β 係数は，アノテーション結果の部分一致を考慮し，独立した確率分布に従う複数のアノテータ間の一致度を示す係数であり，κ 係数と同じように，観測された一致率 A_{obs} と偶然による一致率 A_{exp} を用いて，式 (1) で定義される．

$$\beta = \frac{A_{\mathrm{obs}} - A_{\mathrm{exp}}}{1 - A_{\mathrm{exp}}} \tag{1}$$

観測された一致率 A_{obs} の計算のために，参照表現区間の一致は書き起しされた文字単位で考え，一致の度合いによって異なる重みを与える．個々のアノテーション箇所において，2 人の区間が完全一致する場合 (match) 重み 1，一方の区間が他方を包含する場合（subsume）重み 2/3，2 人のアノテーション区間が一部のみ重なっている場合（overlap）重み 1/3 を与える．それ以外（片方のアノテーションのみで，他方に対応するアノテーションがない）の場合（mismatch）は重み 0 とする．この重みは Foster ら[23] と同じ値を用いた[⋆2]．

偶然による一致率 A_{exp} の計算は，以下の式 (2) で計算する．

$$\begin{aligned} A_{\mathrm{exp}} = {} & (1 - AP_a)(1 - AP_b) \\ & + AP_a AP_b \sum_{l_s=1}^{L_{\max}} P(l_s) \sum_{l_a=1}^{l_s} \sum_{l_b=1}^{l_s} \sum_{i_a=0}^{l_s - l_a} \sum_{i_b=0}^{l_s - l_b} w(l_a, l_b, i_a, i_b) P_a(l_s, l_a, i_a) P_a(l_s, l_b, i_b) \end{aligned} \tag{2}$$

⋆2 実際には Foster らは不一致に対する重みを 0, 1/3, 2/3, 1 の 4 段階で与えている．

$$w(l_a, l_b, i_a, i_b) = \begin{cases} 1 & l_a = l_b \wedge i_a = i_b \text{ (match)} \\ 2/3 & (l_a \neq l_b \wedge i_a \leq i_b \wedge i_b + l_b \leq i_a + l_a) \\ & \vee\ (l_a \neq l_b \wedge i_b \leq i_a \wedge i_a + l_a \leq i_b + l_b) \text{ (subsume)} \\ 1/3 & (i_a < i_b < i_a + l_a < i_b + l_b) \vee (i_b < i_a < i_b + l_b < i_a + l_a) \\ & \text{(overlap)} \\ 0 & \text{otherwise (mismatch)} \end{cases}$$

AP_a，AP_b はアノテータ a，b がある発話に対してアノテーションする確率で，片方でもアノテーションがなされた発話の数に対するそれぞれのアノテータが実際にアノテーションした発話数の割合で推定する．l_s はアノテーションされた発話の長さ，$L_{\max}$ はアノテーションされた発話の最大長，$P(l_s)$ は長さ l_s のアノテーションされた発話の割合である．長さ l_s の発話に対して，それぞれ長さ l_a，l_b のアノテーションを位置 i_a，i_b から開始するものとし，その確率はそれぞれ，$P_a(l_s, l_a, i_a)$，$P_a(l_s, l_b, i_b)$ で表わす．$P_a(l_s, l_a, i_a)$ と $P_a(l_s, l_b, i_b)$ は，実際のアノテーションにおいて，長さ l_s の発話に長さ l_a，l_b のアノテーションをする確率と開始位置の分布 (一様分布を仮定する) から計算する．また，各場合における一致度は，A_{obs} の計算と同じく，一致の度合い（match，subsume，overlap，mismatch）によって重み $w(l_a, l_b, i_a, i_b)$ を与える．各アノテータの組 (A_1, A_2)，(A_1, A_3) について β 係数を計算した結果を表 4 に示す．β 係数の値は，いずれも 0.7 前後の値となっており，参照表現の区間のアノテーションは安定していると考えられる．

表 4　参照表現区間の一致度 (β 係数)

アノテータの組	A_{obs}	A_{exp}	β
(A_1, A_2)	0.752	0.250	0.669
(A_1, A_3)	0.824	0.298	0.749

次に指示対象と属性のアノテーションの一致度を式 (3) で定義される κ 係数で評価する．指示対象や属性も部分一致の可能性があるが，これらの情報は完全に一致していることが重要だと考えられるので κ 係数を用いた．

$$\kappa = \frac{A_{\text{obs}} - A_{\text{exp}}}{1 - A_{\text{exp}}} \tag{3}$$

ここでは，参照表現区間が mismatch とならなかったアノテーションすべてを対象とする．アノテータは，指示対象については，指示対象として 7 つの各ピース含むか否かに加え，定／不定フラグの計 8 種類の組合せ $2^8 = 256$ 通りからひとつのタグを，属性については，18 種類の属性 (dpr, dad, siz, col, typ, dir, prj, tpl, ovl, act, cmp, sim, num, rpr, err, nest, state[⋆1], meta) の組合せ $2^{18} = 262,144$ 通りからひとつタグを選択することになる．

指示対象と属性のアノテーションについては観測された一致率 A_{obs} は，アノテータ間で付与したタグが完全に一致した割合を用いる．偶然による一致率 A_{exp} は，式 (4) により計算する．

$$A_{\text{exp}} = \sum_{X \in \text{Candidate}} P_a(X) P_b(X) \tag{4}$$

ここで，Candidate はタグ候補の集合を表し，$P_a(X)$，$P_b(X)$ はアノテータごとにタグ X を付与する確率を表す．$P_a(X)$，$P_b(X)$ は，実際に各アノテータが付与したタグの頻度から最尤推定する．指示対象，属性の一致度を表 5，6 に示す．いずれも κ 係数は，0.81 ～ 1.00 の高い値であり，指示対象と属性のアノテーションは安定しているといえる．なお，属性についてはアノテーションの検証段階で，実際に現れなかった特徴や不一致の多い特徴を見直し，コーパスでは最終的に表 3 の 19 種類の属性を用いている．

表 5　指示対象の一致度 (κ)

アノテータの組	A_{obs}	A_{exp}	κ
(A_1, A_2)	0.834	0.126	0.810
(A_1, A_3)	0.942	0.151	0.932

表 6　属性の一致度 (κ)

アノテータの組	A_{obs}	A_{exp}	κ
(A_1, A_2)	0.892	0.257	0.855
(A_1, A_3)	0.890	0.183	0.866

3.7　コーパスの概要

アノテーションが完了しているタングラム課題の対話から構築した 2 つのコーパス (T2009-11 と N2009-11) の概要を述べる．図 3 は ELAN でアノテーションしたコーパスのスクリー

⋆1 state 属性はアノテーション終了後の検証過程において削除することしたので，最終的なコーパスには含まれないが，最初のアノーテションの時点では使用した．

Vol.2010-NL-199 No.20
2010/11/19

ンショットである．左上のビデオ画面中の赤丸は指示者の，青丸は作業者の視線位置を表わしている．表 7 は各対話ごとの総対話時間，発話数，参照表現数，表 8 は参照表現の属性の分布をまとめたものである．

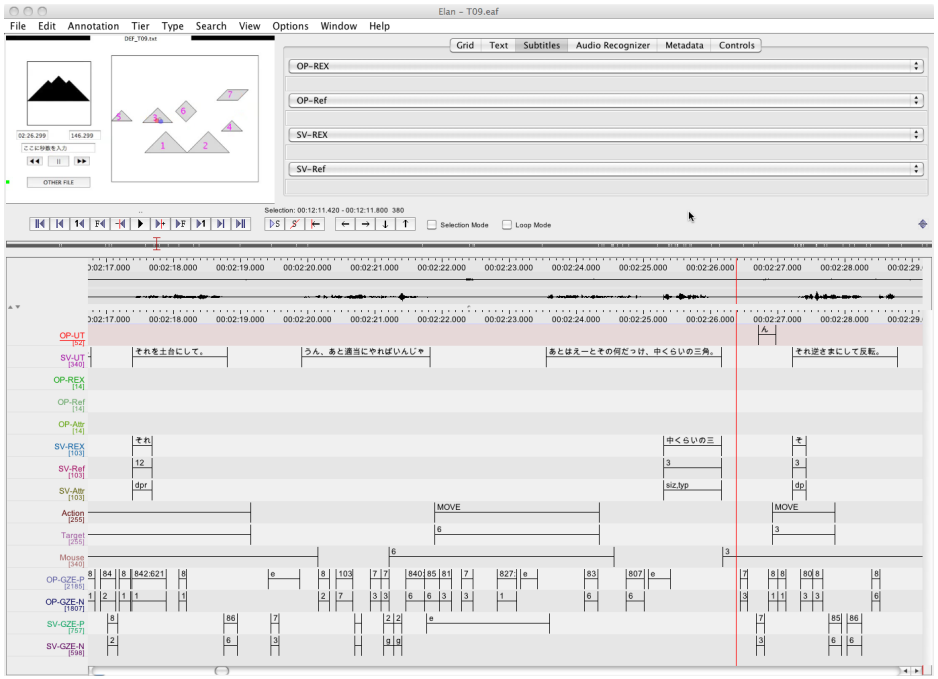


図 3　ELAN によるコーパスのアノテーション

4. コーパスの利用

作成したコーパスを使って，視線情報を利用した参照解析手法の予備的実験をおこなったので，その結果について述べる．

4.1 参照解析手法

1 つのピースを指示対象とする参照表現を対象に，参照解析をおこなう．参照表現の同定にはランキング・モデルに基づいた手法を用いる．ランキング・モデルとは，指示対象候補

表 7　T2009-11，N2009-11 の概要

対話 ID	対話時間	OP-UT	SV-REX	OP-REX	SV-REX
T2009-11					
T06	9:30	78	160	0	32
T07	10:52	53	201	6	61
T08	2:52	9	64	0	17
T09	15:00	52	340	14	103
T10	10:45	46	238	11	92
T11	4:55	91	94	18	25
T12	11:05	175	159	29	41
T13	15:00	252	328	41	94
T14	13:38	212	287	18	68
T15	9:00	116	202	20	57
T16	12:56	167	320	28	81
T17	8:14	49	141	2	29
T18	5:29	11	88	0	24
T19	10:06	51	144	4	40
T20	9:53	51	158	2	48
T21	6:41	34	66	2	17
T22	6:11	9	57	1	19
T23	10:25	16	126	0	25
T24	7:10	16	67	0	17
T29	13:02	149	198	8	40
T30	10:57	117	172	11	34
T31	15:00	197	313	20	54
T32	8:29	96	154	0	33
T33	15:00	0	190	28	40
T34	3:02	28	50	0	19
T35	9:29	78	161	6	42
T36	7:39	52	135	1	40
合計	4:22:20	2,205	4,613	270	1,192
平均	9:43	81.7	170.9	10.0	44.1
SD	3:32	69.4	86.8	11.5	24.8
N2009-11					
N01	14:49	201	233	24	85
N02	15:00	195	286	24	93
N03	14:06	132	276	29	77
N04	11:40	116	194	26	46
N17	15:00	146	206	20	59
N18	15:00	141	227	22	65
N19	15:00	139	212	16	48
N20	7:10	49	82	7	24
合計	1:47:45	1,119	1,716	168	497
平均	13:28	140	215	21	62
SD	2:48	47.4	62.6	6.9	22.8

Vol.2010-NL-199 No.20
2010/11/19

表 8　属性の分布

属性	T2009-11	N2009-11
dpr	673	345
dad	152	57
pnn	60	31
siz	358	170
typ	690	244
dir	2	1
prj	92	19
tpl	4	2
ovl	0	0
act	48	25
cmp	25	12
sim	2	0
num	16	15
rpr	4	0
err	2	1
nest	7	1
meta	12	7
nul	0	1

すべての中でどれがもっとも指示対象らしいかを，ランカーを用いてランク付けし，その 1 位を指示対象と判定するという，多値分類をおこなうモデルである．具体的には，以下のように参照解析をおこなう．

(1)　各参照表現について，その表現の発話開始時までの各ピースの状況を，次項で説明する素性を用いて表現し，ピースごとに特徴ベクトルを作成する．

(2)　Ranking SVM[32)⋆1] を用いて，訓練データからランカーを作成する．2 値分類を行う通常の SVM に対し，Ranking SVM は指定したグループ内でのランク付けを行うことができる．ランカーの学習では，人手でタグ付けをした指示対象を 1 位，その他のピースを 2 位として，学習をおこなう．

(3)　(2) 作成したランカーを利用し，テストデータの特徴ベクトルをランク付けする．

(4)　(3) で 1 位となった特徴ベクトルに対応するピースを，その表現の指示対象とする．

以上の手順で各参照表現につき 1 つのピースを指示対象として同定する．

4.2　特徴ベクトルに用いる素性

Iida らの手法[31)] で用いられた素性 (談話履歴情報，オンマウス情報，操作履歴情報) に加え，視線情報の素性を用いる．談話履歴情報 (D) は，従来の照応解析手法と同様に，談話の先行文脈から得られる情報を素性として用いる．オンマウス情報 (M) は，ELAN の Mouse 層の情報を使い，作業者の操作するマウスがピース上に乗っている状態をオンマウス状態として，これを素性として用いる．オンマウス状態は，人間同士の実環境における直示 (指差しなど) の情報に近いものと考えることができる．

操作履歴情報 (A) は，ELAN の Action 層の情報を参照し，作業者が何らかのピースを操作しているという情報を利用する．本環境では，作業者はピースの移動，ピースの回転，ピースの反転の 3 つの操作を選択しておこなうことができるが，ここではこの 3 つの操作を区別せず，操作を行っているか否かという粒度でこの情報を扱う．

本稿で加える視線情報 (G) は，視線の方向が人の注目情報を表す[34)] という仮定に基づいている．物体を指示する場合には，その指示物体へ注目していると考えることができるため，視線がそのピースへ向いていると考える．具体的には，前節で述べた処理方法により抽出した停留の情報を用いて，ピースへの注目を定義する．各停留では，停留の重心の最も近傍にあるピースまたは目標図形を見ていると考える．ELAN の*-GZE-N 層を参照し，発話開始の一定時間前から発話開始までに発生した停留を見た場合に，合計時間が最も長いピースであるか (G1)，停留の回数が最も多いピースであるか (G2)，一度でも停留が発生したピースであるか (G3) の 3 つの素性を考える．なお，発話開始の一定時間前から発話開始までの間，視線計測が不安定で停留がまったく抽出されなかった場合は，視線情報を用いることができないため，‘unknown’ としている．これらの素性を抽出する時に考慮する発話までの区間については，発話開始 1,500m 秒前から発話開始時までとした．これは，静止画において視線を利用した参照解析をおこなった Prasov らの手法[46)] で用いられた値と同じである．

4.3　評価実験：実験設定

前節で説明した参照解析手法を用い，コーパス中に出現する参照表現の指示対象をどの程度自動的に同定可能か評価実験をおこなった．実験には，今回構築したタングラムを課題とした 2 つのコーパス (T2009-11 と N2009-11) に出現する参照表現中の指示対象がピース 1 つのもの 1,847 表現を使用した．ただし，代名詞を含む表現 911 表現とそれ以外 936 表現を分け，個別に評価実験をおこなった．これは，代名詞が他の参照表現と比較して直示や先行詞との時間的な近さの影響を受けやすい点を考慮したためである[53)]．特徴の異なる代名詞とそれ以外を区別し，それぞれの特徴に合ったランカーを作成することで，精度の向上を図った．Denis らも，参照表現を特徴ごとに区別した specialized model を利用することで，精度が向上したことを報告している[18)]．

以上の理由から，代名詞を含む表現と代名詞以外のランカーを個別に学習するが，個別に

⋆1 http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

素性名	値	説明
D1	yes, no	最後にされたピースか
D2	yes, no	最後に言及されてからの経過時間が 10 秒未満のピースか
D3	yes, no	最後に言及されてからの経過時間が 10 秒以上 20 秒未満のピースか
D4	yes, no	最後に言及されてからの経過時間が，20 秒以上のピースか
D5	yes, no	以前に一度も言及されていないピースか
D6	yes, no, unknown	参照表現の持つ属性 (形・大きさ) が，ピースの属性と矛盾しないか
D7	yes, no	最後にそのピースを指す表現が，ヲ格として用いられているか
D8	yes, no	最後にそのピースを指す表現が，二格として用いられているか
D9	yes, no	参照表現が代名詞の時，その直前にピースが代名詞以外の表現で参照されているか
D10	yes, no	参照表現が代名詞以外の表現の時，その直前にピースが代名詞で参照されているか

※ D6 の yes は属性が矛盾していない場合，no は矛盾している場合，unknown は表現に形と大きさの属性が共になく (例 「それ」) 判断できない場合をそれぞれを表す．

表 9 談話履歴情報の素性

素性名	値	説明
M1	yes, no	発話開始時刻にオンマウスされているピースか
M2	yes, no	発話開始時刻にオンマウスしていないとき，直前にオンマウスしていたピースか
M3	yes, no	最後にオンマウスされてからの経過時間が 10 秒未満のピースか
M4	yes, no	最後にオンマウスされてからの経過時間が 10 秒以上 20 秒未満のピースか
M5	yes, no	最後にオンマウスされてからの経過時間が 20 秒以上のピースか
M6	yes, no	以前に一度もオンマウスされていないピースか

表 10 オンマウス情報の素性

素性名	値	説明
A1	yes, no	発話開始時刻に操作されているピースか
A2	yes, no	発話開始時刻に操作されていない場合，直前で操作されているピースか
A3	yes, no	最後に操作されてからの経過時間が 10 秒未満のピースか
A4	yes, no	最後に操作されてからの経過時間が 10 秒以上 20 秒未満のピースか
A5	yes, no	最後に操作されてからの経過時間が 20 秒以上のピースか
A6	yes, no	以前に一度も操作されていないピースか

表 11 操作履歴情報の素性

素性名	値	説明
G1	yes, no, unknown	1500m 秒前〜発話開始時の間で，最も停留頻度が高いピースか
G2	yes, no, unknown	1500m 秒前〜発話開始時の間で，最も停留時間が長いピースか
G3	yes, no, unknown	1500m 秒前〜発話開始時の間に発生した停留の最近傍にあったピースか

※ unknown は 1500m 秒前〜発話開始時の間に，まったく停留が無かった場合

表 12 視線情報の素性

解析した場合の有効性を調べるために，代名詞を含む表現とそれ以外を区別せずに学習したランカーの結果も示す．ここで，代名詞を含む表現のみを個別に解析するモデルを代名詞モデル，代名詞以外を個別に解析するモデルを非代名詞モデルと呼ぶ．また，代名詞と代名詞以外を個別に解析し，その結果を足し合わせたモデルを個別モデル，代名詞と代名詞以外を分けずに解析したモデルを統合モデルと呼ぶ．

各参照表現の同定で，どの素性カテゴリが有効か比較するために，談話履歴情報のみをベースラインとして，談話履歴情報 (D) にオンマウス情報 (M)，操作情報 (A)，視線情報 (G) を組み合わせた 8 種類のモデルについて実験をおこなった．なお，評価は 5 分割交差検定でおこなった．今回の実験では RankingSVM は線形カーネル，パラメタ c は 1.0 に設定して評価をおこなった．

4.4 評価実験：実験結果

実験結果を表 13 に示す．列は解析に使用した素性の組合せ，行は各モデルに対応する．また，各セルの上段は正解数，下段は精度である．表 13 より，談話履歴情報のみを用いる場合 (D) に比べ，M, A, G の非言語情報を加えることにより，いずれの場合も精度が向上することがわかる．代名詞モデルにおいては，談話履歴情報のみの場合 (D) に，オンマウス情報を加える (D+M) ことで，29.8 ポイントの精度の上昇がみられる．次いで，すべての素性を用いた場合 (D+M+A+G) に，27.3 ポイントの精度の上昇がみられた．代名詞モデルにおいては，特にオンマウス情報 (M) が有効に働くことがわかる．一方．非代名詞モデルにおいては，談話履歴情報のみの場合 (D) に，視線情報を加える (D+G) ことで，10.3 ポイント精度が向上した．すべての素性を用いた場合 (D+M+A+G) も同じく，10.3 ポイント精度が向上した．非代名詞モデルでは，視線情報 (G) が有効に働くことがわかった．

個別モデルと統合モデルでは，非言語情報を加えたすべてのモデルで，個別モデルが統合モデルを上回った．すなわち，参照表現を特徴で分け，非言語情報の使い方を分けることに

より，精度を向上させることができた．

	表現数	D	D+M	D+A	D+M+A	D+G	D+M+G	D+A+G	D+M+A+G
(a) 代名詞モデル	911	436	708	558	698	563	685	627	700
		0.479	0.777	0.613	0.766	0.618	0.752	0.688	0.768
(b) 非代名詞モデル	936	612	644	621	637	709	702	702	709
		0.654	0.688	0.663	0.681	0.757	0.750	0.750	0.757
個別モデル	1,847	1,048	1,352	1,179	1,335	1,272	1,387	1,329	1,409
(a)+(b)		0.567	0.732	0.638	0.723	0.689	0.751	0.720	0.763
(c) 統合モデル	1,847	1,057	1,299	1,132	1,297	1,215	1,302	1,260	1,308
		0.572	0.703	0.613	0.702	0.658	0.705	0.682	0.708

表 **13**　各モデルの正解数と精度

すべての情報を用いたモデル (D+M+A+G) で作成したランカーの各素性の重み (5 分割の平均) を表 14 に示す．代名詞モデルで談話履歴情報 (D)，オンマウス情報 (M)，操作情報 (A)，視線情報 (G) のすべての素性カテゴリから 1 つ以上の素性にトップ 5 の重みが付けられたが，非代名詞モデルでは，トップ 5 の重みが付いた素性は，談話履歴情報 (D) と視線情報 (G) の素性のみであった．

非代名詞の参照表現は，代名詞の参照表現に比べ，それまで操作していたピースから離れて，新規に他のピースを指示対象とする場合に用いられることが多いため，それまでの操作の影響が含まれるオンマウス情報や操作情報ではなく，影響の少ない視線情報が有効に働いたのではないかと考えられる．

5. おわりに

本稿では，2 名の被験者に図形パズルを協調して解かせる課題を通じて収集した対話とその対話中の被験者のパズル・ピースの操作および視線を時間同期して記録したマルチモーダル・コーパスの構築について紹介した．このコーパスでは，人間の参照表現の生成・理解に関する研究をおこなうために，パズル・ピースを参照している表現に人手でアノテーションをおこなった．タングラム・パズルの課題についてはコーパスのアノテーションが完了したので，視線情報を使った参照表現解析の予備的な実験をおこない，視線情報が参照表現の解析に有効利用できる可能性を確認した．今後は，他の図形パズルのコーパスについてもアノテーションを完成させ，視線情報の利用に関するさらに詳細な分析，課題による参照行動の違いの分析などをおこなう予定である．

	代名詞モデル		非代名詞モデル	
ランク	素性	重み	素性	重み
1.	M1	0.487	D6	0.638
2.	G3	0.265	G3	0.339
3.	M3	0.258	D2	0.101
4.	A1	0.186	G2	0.078
5.	D1	0.179	D1	0.075
6.	A3	0.135	G1	0.059
7.	D6	0.134	M1	0.058
8.	A2	0.130	D7	0.053
9.	D2	0.125	M3	0.051
10.	G2	0.090	A4	0.035
11.	G1	0.089	A3	0.026
12.	D9	0.086	M4	0.021
13.	M2	0.051	D8	0.004
14.	D7	0.026	D9	0*
15.	D10	0*	A1	−0.002
16.	M6	−0.028	D3	−0.006
17.	D8	−0.029	D10	−0.016
18.	A6	−0.035	A5	−0.024
19.	D3	−0.038	M2	−0.026
20.	A4	−0.038	A2	−0.029
21.	D5	−0.039	D5	−0.030
22.	D4	−0.048	M6	−0.035
23.	A5	−0.062	M5	−0.036
24.	M4	−0.083	A6	−0.038
25.	M5	−0.148	D4	−0.064

* D10 は条件に「参照解析している表現が代名詞以外の表現の時」と条件を付けているので，代名詞以外の表現の時には素性の値に yes が 1 つも付かないから当然重みもない．D9 に関しても同様である．

表 **14**　(D+M+A+G) における各素性の重み (平均)

参 考 文 献

1) Allopenna, P.D., Magnuson, J.S. and Tanenhaus, M.K.: Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models, *Journal of Memory and Language*, Vol.38, pp.419–439 (1998).
2) Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H.S. and Weiniert, R.: The HCRC Map Task Corpus, *Language and Speech*, Vol.34, No.4, pp.351–366 (1991).
3) Artstein, R. and Poesio, M.: Kappa^3 = Alpha (or Beta), Technical Report CSM-437, University of Essex (2005).
4) Auer, E., Russel, A., Sloetjes, H., Wittenburg, P., Schreer, O., Masnieri, S., Schneider, D. and Tschöpel, S.: ELAN as Flexible Annotation Framework for Sound and Image Processing DetectorsELAN as Flexible Annotation Framework for Sound and Image Processing Detectors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, pp.890–893 (2010).
5) Bard, E. G., Anderson, A. H., Chen, Y., Nicholson, H. B. M., Havard, C. and Dalzel-Job, S.: Let's you do that: Sharing the cognitive burdens of dialogue, *Journal of Memory and Language*, Vol.57, No.4, pp.616–641 (2007).
6) Bard, E. G. B., Hill, R. and Arai, M.: Referring and gaze alignment: Accessibility is alive and well in situated dialogueccessibility is alive and well in situated dialogue, *Proceedings of the Annual Meeting of the Cognitive Science Society*, pp.1246–1251 (2009).
7) Bock, K., Irwin, D. and Davidson, D.J.: Putting first things first, *The interface of language, vision, and action: Eye movements and the visual world* (Henderson, J. H. and Ferreira, F., eds.), Psychology Press, chapter8, pp.249–278 (2004).
8) Brennan, S. E., Chen, X., Dickinson, C. A., Neider, M. B. and Zelinsky, G. J.: Coordinating cognition: The costs and benefits of shared gaze during collaborative search, *Cognition*, Vol.106, pp.1465–1477 (2008).
9) Brône, G., Oben, B. and Feyaerts, K.: InSight Interaction – A multimodal and multifocal dialogue corpus, *The Proceedings of the Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, pp.157–159 (2010).
10) Byron, D.K.: The OSU Quake 2004 corpus of two-party situated problem-solving dialogs, Technical report, Department of Computer Science and Enginerring, The Ohio State University (2005).
11) Byron, D. K. and Fosler-Lussier, E.: The OSU Quake 2004 corpus of two-party situated problem-solving dialogs, *Proceedings of the 15th Language Resources and Evaluation Conference (LREC 2006)* (2006).
12) Campana, E., Baldridge, J., Dowding, J., Hockey, B.A., Remington, R.W. and Stone, L.S.: Using eye movements to determine referents in a spoken dialogue system, *Proceedings of the 2001 workshop on Perceptive user Interfaces*, pp.1–5 (2001).
13) Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D. and Wellner, P.: The AMI Meeting Corpus: A Pre-announcement, *Machine Learning for Multimodal Interaction*, LNCS 3869, Springer-Verlag, pp.28–39 (2006).
14) Carletta, J., Isard, A., Isardt, S., Kowtkot, J.C., Doherty-Sneddon, G. and Anderson, A.H.: The Reliability of a Dialogue Structure Coding Scheme, *Computational Linguistics*, Vol.23, No.1, pp.13–31 (1997).
15) Carletta, J., l. Hill, R., Nicol, C., Taylor, T., de Ruiter, J.P. and Bard, E.G.: Eyetracking for two-person tasks with manipulation of a virtual world, *Behavior Research Methods*, Vol.42, No.1, pp.254–265 (2010).
16) Cassell, J., Sullivan, J., Prevost, S. and Churchill, E.(eds.): *Embodied Conversational Agents*, The MIT Press (2000).
17) Chen, L., Rose, R. T., Qiao, Y., Kimbara, I., Parrill, F., Welji, H., Han, T. X., Tu, J., Huang, Z., Harper, M., Quek, F., Xiong, Y., McNeill, D., Tuttle, R. and Huang, T.: VACE Multimodal Meeting Corpus, *Machine Learning for Multimodal Interaction*, LNCS 3869, Springer-Verlag, pp.40–51 (2006).
18) Denis, P. and Baldridge, J.: Specialized models and ranking for coreference resolution, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 660–669 (2008).
19) Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C. and Tanenhaus, M. K.: Eye Movements as a Window into Real-Time Spoken Language Comprehension in Natural Contexts, *Journal of Psycholinguistic Research*, Vol.24, No.6, pp.409–436 (1995).
20) Epelboim, J. and Suppes, P.: A model of eye movements and visual working memory during problem solving in geometry, *Vision Research*, Vol.41, pp.1561–1574 (2001).
21) Ferreira, F. and Tanenhaus, M.K.: Introduction to the special issue on language–vision interactions, *Journal of Memory and Language*, Vol.57, pp.455–459 (2007).
22) Foster, M. E., Bard, E. G., Guhe, M., Hill, R. L., Oberlander, J. and Knoll, A.: The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue, *Proceedings of 3rd Human-Robot Interaction*, pp.295–302 (2008).
23) Foster, M.E. and Oberlander, J.: Corpus-based generation of head and eyebrow motion for an embodied conversational agent, *Language Resources and Evaluation*, Vol.41, No.3–4, pp. 305–323 (2007).
24) Frampton, M., Fernández, R., Ehlen, P., , Christoudias, M., Darrell, T. and Peters, S.: Who is "You"? Combining Linguistic and Gaze Features to Resolve Second-Person References in Dialogue, *Proceedings of the 12th Conference of the European Chapter of the ACL*, pp. 273–281 (2009).

25) Grant, E.R. and Spivey, M.J.: Eye movements and problem solving, *Psychological Science*, Vol.14, No.5, pp.462–466 (2003).
26) Griffin, Z. M. and Bock, K.: What the eyes say about speaking, *Psychological Science*, Vol.11, No.4, pp.274–279 (2000).
27) Hanna, J. E. and Brennan, S. E.: Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation, *Journal of Memory and Language*, Vol.57, pp.596–615 (2007).
28) Hayhoe, M.M., Bensinger, D.G. and Ballard, D.H.: Task constraints in visual working memory, *Vision Research*, Vol.38, No.1, pp.125–137 (1998).
29) Hegarty, M. and Just, M.A.: Constructing Mental Models of Machines from Text and Diagrams, *Journal of Memory and Language*, Vol.32, No.6, pp.717–742 (1993).
30) Herrera, D., Novick, D., Jan, D. and Traum, D.: The UTEP-ICT Cross-Cultural Multiparty Multimodal Dialog Corpus, *The Proceedings of the Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, pp.49–54 (2010).
31) Iida, R., Kobayashi, S. and Tokunaga, T.: Incorporating Extra-linguistic Information into Reference Resolution in Collaborative Task Dialogue, *Proceedings of 48th Annual Meeting of the Association for Computational Linguistics*, pp.1259–1267 (2010).
32) Joachims, T.: Optimizing serch engines using clickthrough data, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.133–142 (2002).
33) Jokinen, K., Yamamoto, S. and Nishida, M.: Collecting and Annotating Conversational Eye-Gaze Data, *The Proceedings of the Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, pp.125–130 (2010).
34) Just, M.A. and Carpenter, P.A.: Eye fixations and cognitive processes, *Cognitive Psychology*, Vol.8, pp.441–480 (1976).
35) Kelleher, J., Costello, F. and van Genabith, J.: Dynamically Structuring Updating and Interrelating Representations of Visual and Linguistic Discourse, *Artificial Intelligence*, Vol.167, pp.62–102 (2005).
36) Kelleher, J. and van Genabith, J.: Visual salience and reference resolution in simulated 3-d environments, *Artificial Intelligence Review*, Vol.21, No.3, pp.253–267 (2004).
37) Kelleher, J.D.: Attention driven reference resolution in multimodal contexts, *Artificial Intelligence Review*, Vol.25, pp.21–35 (2006).
38) Kipp, M., Martin, J.-C., Paggio, P. and Heylen, D.(eds.): *Multimodal Corpora*, LNAI 5509, Springer-Verlag (2009).
39) Knoblich, G., Ohlsson, S. and Raney, G.E.: An eye movement study of insight problem solving, *Memory & Cognition*, Vol.29, No.7, pp.1000–1009 (2001).
40) Knoeferle, P., Crocker, M.W., Scheepers, C. and Pickering, M.J.: The influence of the immediate visual context on incremental thematic role-assignment: evidence from eye-movements in depicted events, *Cognition*, Vol.95, No.1, pp.95–127 (2005).
41) Martin, J.-C., Paggio, P., Kuehnlein, P., Stiefelhagen, R. and Pianesi, F.: Spetial Issue on: Multimodal Corpora for Modeling Human Multimodal Behavior, *Language Resources and Evaluation*, Vol.41, No.3-4 (2007).
42) Matin, E.: Saccadic suppression: a review and an analysis, *Psychological Bulletin*, Vol.81, No.12, pp.899–917 (1974).
43) Metzing, C. and Brennan, S.E.: When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions, *Journal of Memory and Language*, Vol.49, pp.201–213 (2003).
44) Meyer, A.S., Sleiderink, A.M. and Levelt, W. J.M.: Viewing and naming objects: Eye movements during noun phrase production, *Cognition*, Vol.66, No.2, pp.B25–B33 (1998).
45) Murray, N. and Roberts, D.: Comparison of head gaze and head and eye gaze within an immersive environment, *Proceedings of the 10th IEEE international symposium on Distributed Simulation and Real-Time Applications*, pp.70–76 (2006).
46) Prasov, Z. and Chai, J.Y.: What's in a gaze?: The role of eye-gaze in reference resolution in multimodal conversational interfaces, *Proceedings of the 13th international conference on Intelligent user interfaces*, pp.20–29 (2008).
47) Qu, S. and Chai, J.: Incorporating Temporal and Semantic Information with Eye Gaze for Automatic Word Acquisition in Multimodal Conversational Systems, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, pp.244–253 (2008).
48) Qu, S. and Chai, J.Y.: The role of interactivity in human-machine conversation for automatic word acquisition, *Proceedings of the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2009)*, pp.188–195 (2009).
49) Rayner, K.: Eye movements in reading and information processing: 20 years of research, *Psychological Bulletin*, Vol.124, No.3, pp.372–422 (1998).
50) Richardson, D.C., Dale, R. and Kirkham, N.Z.: The art of conversation is coordination – Common ground and the coupling of eye movements during dialogue, *Psychological Science*, Vol.18, No.5, pp.407–413 (2007).
51) Richardson, D.C., Dale, R. and Spivey, M.J.: Eye movements in language and cognition: A brief introduction, *Methods in Cognitive Linguistics* (Gonzalez-Marquez, M., Mittelberg, I., Coulson, S. and Spivey, M.J., eds.), John Benjamins., pp.323–344 (2007).
52) Richardson, D.C., Dale, R. and Tomlinson, J.M.: Conversation, Gaze Coordination, and Beliefs About Visual Context, *Cognitive Science*, Vol.33, No.8, pp.1468–1482 (2009).
53) Spanger, P., Yasuhara, M., Iida, R. and Tokunaga, T.: Using extra linguistic information for generating demonstrative pronouns in a situated collaboration task, *Proceedings of PreCogSci 2009: Production of Referring Expressions: Bridging the gap between computational and empirical approaches to reference* (2009).

54) Spivey, M.J., Tanenhausb, M.K., Eberhardc, K.M. and Sedivy, J.C.: Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution, *Cognitive Psychology*, Vol.45, No.4, pp.447–481 (2002).
55) Stein, R. and Brennan, S.E.: Another person's eye gaze as a cue in solving programming problems, *Proceedings of the 6th international conference on Multimodal interfaces (ICMI 2004)*, pp.9–15 (2004).
56) Steptoe, W., Wolff, R., Murgia, A., Guimaraes, E., Rae, J., Sharkey, P., Roberts, D. and Steed, A.: Eye-tracking for avatar eye-gaze and interactional analysis in immersive collaborative virtual environments, *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pp.197–200 (2008).
57) Takemae, Y., Otsuka, K. and Mukawa, N.: An analysis of speakers' gaze behaviour for automatic addressee identification in multiparty conversation and its application to video editing, *Proceedings of the 2004 IEEE International Workshop on Robot and Human Interactive Communication*, pp.581–586 (2004).
58) Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M. and Sedivy, J.C.: Integration of visual and linguistic information in spoken language comprehension, *Science*, Vol.268, No.5217, pp.1632–1634 (1995).
59) Tokunaga, T., Iida, R., Yasuhara, M., Terai, A., Morris, D. and Belz, A.: Construction of bilingual multimodal corpora of referring expressions in collaborative problem solving, *Proceedings of 8th Workshop on Asian Language Resources*, pp.38–46 (2010).
60) Voss, C.R., Gurney, J. and Walrath, J.: Exploration in a Large Corpus: Research on the Integrationof Eye Gaze and Speech with Visual Information in a Virtual Reality System, *Intelligent Integration and Use of Text, Image, Video, and Audio Corpora* (Hauptmann, A. and Witbrock, M., eds.), AAAI Technical Report SS-97-03 (1997).