ユビキタスセンサによるインタラクション・コーパスの収集とインタラクション分析への応用

坊農真弓[†‡] 鈴木紀子[‡] 片桐恭弘[‡*]

神戸大学大学院総合人間科学研究科[†] ATR メディア情報科学研究所[‡] 神戸大学*

1. はじめに

これまで，音声言語の解析や音声認識の技術向上を目的に，音声発話や音声対話を集めたコーパス作成が試みられてきた[1,2 等]．これらの音声コーパスの収集環境は実験室などの閉鎖的なものが大半であった．本発表では，角ら[3]によって収集されたインタラクション・コーパスの分析を行う．このコーパスはポスタープレゼンテーション会場という参加者が流動的・開放的な環境で，視野画像や音声情報，身体配置を示すセンサ情報など，多岐に渡った情報の収集を試みている．ここでは，ポスター会場で来訪者がどのようにコミュニケーションの場を作成し，その中でどのような振る舞いをするかに焦点を当て，社会構造の動的な変遷を観察する．このような社会学的理論を背景にした観察は，今後重要視されてくるであろう，ユビキタス・ウェアラブル環境で収集したデータの分析や利用に有益な示唆を与えるものである．

2. 社会学的アプローチ

2.1. 会話場 (conversational space)

我々が対面して会話を行う際，対面しあっている者同士の間には物理的な空間が生まれる．本発表ではそれを「会話場 (conversational space)」と呼ぶ．会話に対し社会学的なアプローチを試みる Kendon[4]は，会話参与者(participant)の活動は参与者それぞれの「操作領域(transactional segment)」で行われるものとしている．この領域は参与者の前方の拡がる空間，つまり視野領域である．Kendon は人々が協同で何かを達成する際，各参与者の操作領域(視野領域)が重なり合って共同の空間が生成されると述べている．このような会話場の生成には各参与者の視野情報が重要な要因であると考えられている[5].

2.2. 参与構造(participation structure)

相互行為の局面を分析対象にしてきた社会学者は参与者らの視野領域の重なりによって形作られる会話場の内部構造を「参与構造(participation structure)」(又は，構造まで至らず存在自体を「参与枠組」「参与フレーム」(participation framework)など)と呼ぶ[5-9]．複数の参与者によって協同に構築されていく会話では，発話の権利を譲渡しあうことで「話し手」「聞き手」「それ以外」という会話上の役割が瞬時に交換される．これらの構成素（参与者）によって構成される会話の構造（参与構造）は決して静的なものではなく流動的である．Clark[8]はあるひとつの発話が発せられる際のその会話場付近にいる人々の役割を以下のようなモデルを挙げ，説明している．

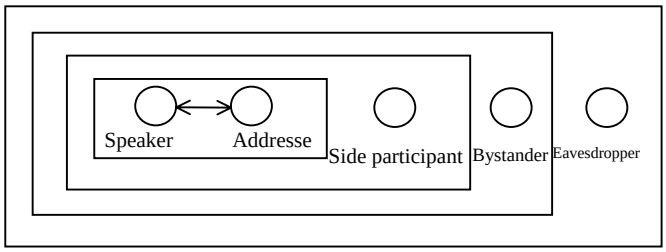


図 1. Clark[8]で提案されるモデル

このモデルでは Speaker(話し手)を発話を産出する者，Addressee(聞き手)を話し手が産出した発話が割り当てられた者，Side participant(傍-参与者)を会話には参加しているが発話が割り当てられていない者とする．また Clark は Bystander と Eavesdropper を合わせて Overhearer と呼び，会話場にはいるものの会話に参加する権利や責任を負わない者とする．Bystander は目に見える形で会話場にいるが会話の一部ではなく,Eavesdropper は話し手に気づかれることなく話を聞いている者として定義される．

2.3. 2 段階の参与

本発表では，2.2 節に記した参与構造は 2 段階の参与によって形成されるものと考える．「会話場への参与」「会話への参与」である．今回使用するコーパスのポスタープレゼンテーショ

Interaction analysis of poster presentations based on interaction corpus of ubiquitous sensor data

†Graduate School of Kobe University, Cultural Studies and Human Science.

‡ ATR Media Information Science Laboratory.

* Kobe University.

ンの例を挙げれば，ポスター近辺，ポスター説明員の音声が聞こえる(聴覚的)範囲，ポスターの活字が読める(視覚的)範囲に接近するという物理的な距離の移動が，会話場への参与(non-participant から Bystander への昇格)である．その会話場の中で発話権を取る(Speaker への昇格)，話し手に発話を割り当てられる(Addressee への昇格)，説明員と向き合う(Side participant への昇格)などの行動が会話への参与である．

よって，本発表では Bystander，Eavesdropper は会話場にのみ参与した者であり，Speaker, Addressee, Side participant は会話場と会話に参与している者と考える．

2.4. 聞き手デザイン(audience design)

2.2 節で触れた話し手が発話を割り当てるという行為は Clark & Carlson[9]によって「聞き手デザイン(audience design)」と呼ばれる．これは Addressee になりえる人物が複数人存在し，その中の特定の 1 名に向かって話し手が発話を産出する場合，その中の一人しか知りえない話題を発話の中に盛り込むなどといった操作である．このように，その会話場にいる人々のどの人に向かってその発話を産出したのかを示すことは，現時点での話し手による自主的な次話者特定に繋がる．これまで，この聞き手デザインの分析は発話内容分析を通じ，相互行為者間の知識・背景・社会的アイデンティティなどという要素から観察が進められた[7,10].

以下では，参与者の視野や身体配置を示す画像データの分析を通じ，この聞き手デザインという操作が，相手の方へ体を向ける，相手を自分の視野に入れるなどの非言語的手段を用いて実現される可能性を分析する．

3. データ分析

3.1. なぜ参与できなかったのか

次の図 2(事例 1)および図 3(事例 2)の左に示す静止画は，来訪者がポスタープレゼンテーション会場のあるポスター前に立ったにも関わらず(会話場への参与：Non-participant から Bystander への昇格)，会話への参与 (Bystander から Side participant への昇格)を行わなかった事例である．

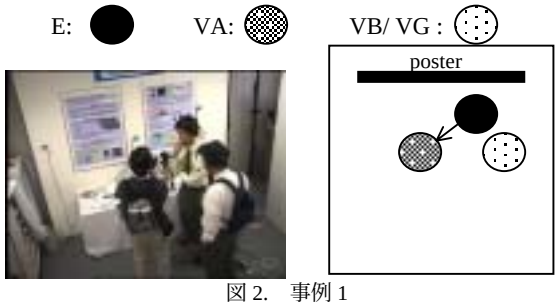


図 2. 事例 1

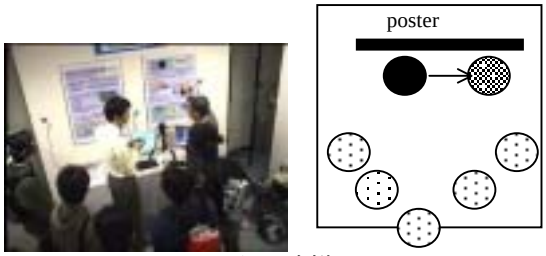


図 3. 事例 2

事例 1 は，3 名の人物がポスターの前に立っている光景である．1 名はポスターの説明員(Exhibitor:E)であり，残る 2 名は説明を聞きに来ている来訪者である．この来訪者 2 名のうち 1 名(VisitorA:VA)は先に会話場へ参与し説明員から説明を受けている．残る 1 名の来訪者(VisitorB:VB)は，E と VA の間にある程度会話が交わされた後，会話場へ接近している．事例 1 から分かるように VB は会話を行っている E と VA のかなり近くまで接近している．しかし，VB は E，VA と会話を交わさず，およそ 27 秒後[1]に会話場から去る．事例 2 は事例 1 の場合と比較的類似しており，もとから E と VA(事例 1 の時の VA とは異なる人物)が会話場におり，その会話場に 5 名で構成される友達同士のグループ(Visitor group: VG)が会話場に参与してくる例である(VG の会話場への参与距離は事例 1 の事例ほど近くない)．VG もまたこの会話場に 126 秒程度[2]参与したが，E と VA によって構成される会話への参与は行わない．第三者的に会話場に参与してきた事例 1 の VB と事例 2 の VG は図 1 の Clark の役割に当てはめると Bystander の役割を担っている．ここではなぜ彼らは Side participant に昇格せず(できず)，会話に参与しなかった(できなかった)のであろうか．

3.2. 2 つの会話モード

3.1 節で観察した 2 つの事例を説明するためには，会話における 2 つのモードの存在を証明することが有益である．その 2 つのモードを本発表ではレクチャーモード(lecture mode: L-mode)とインタラクションモード(interaction mode: I-mode)と呼ぶ．

会話には，ポスタープレゼンテーションのような説明員の役割を担った参与者に発話量負荷がかかるものや，自然なおしゃべりのような 2 名以上の参与者が全く等しい発話権利を持ったものなどいくつかの種類が挙げられる．この L-mode と I-mode はポスタープレゼンテーションでの会話にのみ当てはまるモードの違いではなく，

[1]天井に設置した環境カメラに VB の姿が確認された時間長.
[2] VG メンバー5 名のうち一人目が天井に設置した環境カメラで確認されてから 5 名全員が同環境カメラからフレームアウトするまでの時間長.

すべての会話にこの 2 つのモードは含まれている．ポスタープレゼンテーションでの会話を例に挙げると，ポスターの前に説明員が立ち来訪者が集まってきた場合，(ⅰ) とにかくまず説明員が来訪者に一方的にポスターの内容を説明する．その後，(ⅱ) 来訪者の質疑と説明員の応答という相互行為が開始される．(ⅰ)が L-mode で(ⅱ)が I-mode である．以下の基準に従い，インタラクション・コーパス内の会話を L-mode と I-mode に分類した．

【L-mode】
- 開始点：説明員が，それまで会話を交わしていない来訪者にポスター内容を説明し始める箇所．その箇所の第一発話の冒頭．(説明員の誘いかけか，来訪者の依頼により開始)
- 終了点：説明を受けた来訪者が質問し始める箇所．来訪者の質問の前に発せられた説明員の発話の末尾．
- 話者交替無し

【I-mode】
- 開始点：一定時間続く L-mode の後，来訪者が質問し始める箇所．来訪者の質問発話の冒頭．
- 終了点：会話場にいる参与者によって，会話が明示的に閉じられる箇所．明示的な会話終了発話の末尾．
- 話者交替有り

3.3. 発話持続長と視野情報の分析

L-mode と I-mode，それぞれでの説明員の発話持続長と視野情報はどのような振る舞いを見せるのであろうか．本発表では仮説として以下の表 1 を挙げる．

表 1. 説明員の 発話持続長と視野情報

	発話持続長	視野情報
L-mode	長	ポスター
I-mode	短	ポスター以外

これは，説明員は L-mode の際，発話持続長が長く，ポスターを見ながら説明する傾向にあり，I-mode の際，発話持続長が短く，ポスター以外(来訪者など)を見ながら来訪者の質問に答えるのではないかという予想である．このように L-mode と I-mode には発話持続長，視野範囲という 2 点に関して物理的な差異が見られるかどうか，以下で分析する．

3.3.1. 発話持続長

本研究で用いたインタラクション・コーパスの中から，3.2 節で示した L-mode と I-mode の傾向が見られる箇所[3]を選び出し，L-mode 箇所 (データ 1)30 秒程度，I-mode 箇所(データ 2)80 秒程度の説明員の振る舞いを観察する[4]．これらのデータは説明員の頭部に付けられた視野画像と音声のデータであり，WaveSurfer1.4.6 という画像音声分析ツールを用い，分析した．

まず，それぞれのデータの無音区間を探し出し，データ内の会話構成を分析した．データ内の発話持続長を分析するため，「間休止単位(inter-pausal unit: IPU)」の考え方に従い提案される以下の基準を採用した．これは，1. 発話内の 100msec 以下の無音区間をポーズと認めず後続単位に含め，2. 100msec 以上 400msec 以内の無音区間をポーズとし，3. 400msec 以上の無音区間を IPU の切れ目とする考え方である．発話内の文法的要素は考慮しない．

＜間休止単位(IPU)の特定(無音区間)＞[5]

1. 100msec 以内 ：ポーズとして認めない
2. 100msec～400msec：ポーズ
3. 400msec 以上 ：IPU 境界

本発表で関わるのは 3 である．400msec 以上の無音区間で挟まれた音声を 1 単位として考える[2]．データ 1 と 2 を上記の基準に従い，IPU に切り分け，それぞれの IPU 持続長(つまり発話持続長)の平均値・標準偏差を導いた．

表 2. L-mode と I-mode の IPU 平均持続長

	平均	標準偏差
L-mode (6 IPU)	3346.5msec	1161.6msec
I-mode (15 IPU)	2260.9msec	2318.4msec

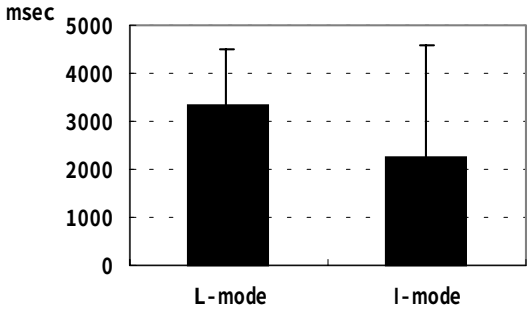


図 4. L-mode と I-mode の IPU の平均発話持続長

以上の結果は表 1 で挙げた仮説を裏付けるものとして解釈できる． L-mode での IPU は持続長が I-mode に比べ長く，各 IPU の持続長が一定している．しかし，I-mode での IPU は持続長が L-mode に比べ短く，各 IPU の持続長は一定していない傾向にある．この原因として，L-mode の場合は話者交替が起こらず，説明員が発話権を維持しているのに対し，I-mode は来訪者の質疑応答の会話であり，発話権が来訪者に渡ることが挙げられる．話者交替により I-mode では極端に短い IPU が確認された．話者交替との関連を今後検討する必要がある．もう一つ大きな理由として，説明員の視野情報との関連が挙げられるのではないか．表 1 の仮説で挙げたように，説明

[3] 説明員 1 名に対し，来訪者 6 名のポスタープレゼンテーション．
[4] L-mode では話者交替がなく，I-mode では話者交替があるため，説明員の発話量をなるべくそろえるために分析のデータ量に差を与えた．
[5] 無音区間によるターンを構成する単位は人工知能学会言語・音声理解と対話処理研究会(SIG-SLUD)における談話・対話研究におけるコーパス利用研究グループの作成した基準に従った．http://www.slp.cs.ritsumei.ac.jp/dtag/

員は L-mode の時はポスター，I-mode の時はポスター以外(来訪者)を注視している可能性が考えられる．I-mode の際，来訪者に目を向け，話者交替が可能であることを来訪者に知らせる場合には，説明員のターンを構成する発話と発話の間にポーズが入りやすく，IPU の再構成が起こりやすくなる可能性がある．つまり，I-mode の会話自体が来訪者の存在を配慮しているために説明員の発話持続長が短くなるという仮説が立てられる．その仮説の適切性を見るために 3.3.2 節では視野情報とモードの関連性を観察する．

3.3.2. 視野情報

前節で使用した L-mode (データ 1)と I-mode (データ 2)箇所の説明員の視野画像を分析する．

データ 1，データ 2 のそれぞれの開始から 1 分経過時から 2 分経過時までの 1 分間の説明員の視野画像を，ポスター(P)，ポスター以外(n-P)でラベル付けした[6].

表 3. L-mode と I-mode の視野情報

	ポスター(P)	ポスター以外(n-P)
L-mode	71.5%	28.5%
I-mode	31.5%	68.5%

この分析から，L-mode ではこの 1 分間の内，全体の 71.5%の間，視野の中にポスターが入っており，I-mode では全体の 31.5%しか視野にポスターが入っていないことが明らかになった．この結果は，I-mode では説明員は来訪者に向かい合う姿勢で説明を行っていると理解することができる[7].

データ 1 とデータ 2 とは異なる時間帯で同じ説明員が来訪者 1 名に対し，L-mode で説明を行う箇所(データ 3)の 1 分経過時から 2 分経過時までの 1 分間の視線情報を以下に挙げる．

表 4. 1 対多と 1 対 1 の視野情報

	ポスター(P)	ポスター以外(n-P)
L-mode(1 対多)	71.5%	28.5%
L-mode(1 対 1)	55.3%	44.7%

表 4 はデータ 1 とデータ 3 の結果を並べて記載している．これらの比較から，説明員は，来訪者が複数の時は来訪者の集まりに視線を向けず，来訪者が 1 名の場合はその来訪者に比較的視線を向ける傾向があることが分かる．これは，来訪者が 1 名の場合，聞き手としてデザインする相手の候補が複数いるわけではないため，聞き手として相手を特定しやすいということが考えられる．

3.4. 分析から予想できること

ここで 3.1 節に挙げた，なぜ参与できなかったのかという問題に立ち戻ろう．図 2 の事例 1，図 3 の事例 2 共に，E と VA は発話権も数回交換され，E の身体配置も VA に向けられ I-mode が開始されている．VB や VG は，会話を構成している E と VA の結びつきの強さから，2 人の中に入っていくことができなかったのである．もし，E と VA がまだ L-mode を形成していれば VB も VG も Side participant(更には Addressee, Speaker)への昇格が可能であったかもしれない．このように第三者の会話への参与の容易さは，既存の参与者による「インタラクションの深さ」に伴い変化が生じる．3.3 節での分析のように，インタラクションの深さは説明員の発話持続長や視野情報に表れると考えられる．

4. おわりに―今後の可能性

表 3 と 4 の結果が示唆するように，インタラクションの形態は，参与者の人数に左右される部分がある．今後は会話場の形成，L-mode・I-mode の特徴などを詳しく分析し，角ら[3]の収集したセンサ情報などから自動的にそれらを抽出する手法を検討する．このような検討は人の社会における動的な流れや興味の移り変わりなどから生じる，身体や言語の振る舞いを知る手がかりとなるであろう．

謝辞：本研究は通信・放送機構の研究委託により実施したものである． インタラクション・コーパスを用いるにあたり，ATR メディア情報科学研究所第 1 研究室の皆様にご協力いただいた．

参考文献

[1] 山本. 重点領域研究「音声対話」の音声対話コーパス. 人文学と情報処理, No. 12, 63-65. 1996.
[2] 堀内 他. 日本語地図課題対話コーパスの設計と特徴. 人工知能学会誌, Vol.12 No.2 1998.
[3] 角 他.複数センサ群による協調的なインタラクションの記録. インタラクション 2003 (発表予定), 2003.
[4] Kendon,A. *Conducting interaction.* Cambridge University Press. 1990.
[5] 西阪.心と行為―エスノメソドロジーの視点.岩波書店.2001.
[6] Goffman,E. *Frame Analysis.* Northeastern University Press. 1974
[7] Goodwin,C. Audience diversity, participation and interaction. *Text*,6(3): 283-316. 1986.
[8] Clark,H.H. *Using language.* Cambridge University Press 1996.
[9] Clark,H.H.&Carlson,T.B. Hearers and speech acts. *Language,* 58: 332-373. 1982
[10] 高梨.会話連鎖の組織化過程における聞き手デザインの機能. 第 10 回社会言語科学会研究大会予稿集, 2002.

[6] ポスターの一部が視野画像に入っているものも「P」とラベル付けした．ポスターからポスター以外に移る場合，ポスターが完全に視野画像から無くなった時点で「n-P」とし，ポスター以外からポスターに移行する場合，ポスターの一部が視野画像に入るまでを「n-P」とし，入ってからを「P」とラベル付けた．

[7] なお，3.3.1 節での発話持続長の分析と同様，I-mode の会話の中には来訪者の質疑の発話も含まれており，この際説明員は来訪者の方へ体を向けざるを得ない．このような来訪者の話し掛けによる説明員の受動的な身体配置と自分の意志で来訪者の方を向くという説明員の能動的な身体配置の差異は今後分けて分析する必要がある．