異分野融合によるマルチモーダルコーパス設計 ― 各種アノテーション方法と利用可能性について―

城 綾実　牧野 遼作　坊農 真弓
高梨 克也　佐藤 真一　宮尾 祐介
国立情報学研究所, 総合研究大学院大学, 京都大学
{joh, ryosaku, bono, satoh, yusuke}@.nii.ac.jp,
takanasi@ar.media.kyoto-u.ac.jp

1 はじめに

筆者らは，コミュニケーションを異分野融合型のアプローチで解明していくための土台となるマルチモーダルコーパス作成に取り組んでいる．本稿では，まず，日本科学未来館 (以下，未来館と呼ぶ) の展示フロアにおける科学コミュニケーションの一種である対話を研究するに至った経緯をもとにマルチモーダルコーパス設計の目的を述べる．次に，対話の収録手法について紹介する．筆者らが別稿 [1] で簡単に述べた音声発話および身体動作のアノテーション方法について詳述し，最後にアノテーションを異分野融合研究にどのように活かしていくかについて展望を示す．

2 コーパス設計の背景

2.1 科学コミュニケーターと対話

未来館では，科学者・技術者と一般の人々をつなぐ役割として科学コミュニケーター (Science Communicator；以下 SC と呼ぶ) の積極的な登用・養成を行っている．SC の未来館での職務は，(1) 展示フロアでの解説や実演，(2) 展示やイベントの企画・制作，(3) 科学情報の発信や他の組織とのネットワークづくり，の 3 つに大別できる．特に，(1) に関しては，「知識を伝えるだけでなく，皆さんと共に考えながら話を深めていく」，「正解のない問題に対し，様々な立場の意見を聞くことでみんなが新たな気づきを得る」ことを「未来館スタイル」として打ち出している [1]．

SC は，「未来館スタイル」実現のために，科学技術に関する正確な知識だけでなく，来館者がどのような知識や関心を汲み取って，来館者に合わせた対話[2]を構築する技術も求められる．しかし，現在の勤務体制では，SC は先輩の対話する姿を見ることも，ノウハウを学ぶ機会も乏しい状況である．経験豊富な SC の対話をデータベース化したい SC 側と，老若男女の多様なコミュニケーションのしくみを探求したい筆者らは，対話を構築している実践の解明を目指して，2012 年秋から研究を開始した [2]．国立情報学研究所と未来館とで共同研究計画を結び，未来館内の研究棟を拠点として，インタラクション分析や教育工学の観点から研究を進めている [3][4][5][1][6][7]．

[1] http://www.miraikan.jst.go.jp/online/communication/work.html
[2] 展示フロアにおける科学コミュニケーターと来館者間の科学技術に関する会話をこれ以降，対話と呼ぶ．

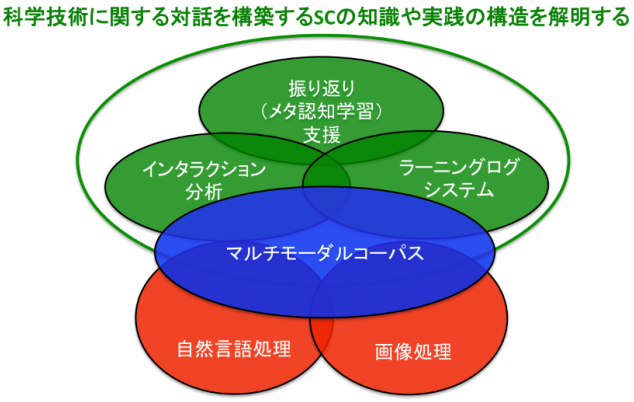

図 1: 本プロジェクト関連分野の見取り図.

2.2 コーパス設計の目的

当初は，会話分析 [8][9] をもとにしたインタラクション分析 [10] と振り返り（メタ認知学習）支援，ラーニングログシステム [4][5] の 3 点から研究を進めていた（図 1 上部）．しかし，インタラクション研究を多角的に進めるためには，自然言語処理や画像処理等のメディア処理技術が将来有用となると考えられる．したがって，これらの複数分野の研究で共通に利用できるデータの構築（マルチモーダルコーパスの設計）を目指すこととした図 1 の中心部）．

マルチモーダルコーパスが完成すれば，個々の研究が進むだけでなく，それぞれの領域の知見や技術を融合させて新たな研究を生み出せる可能性も広がる．また，マルチモーダルコーパスを公開すれば，プロジェクトに参加していない研究領域の研究者（例えば，科学コミュニケーション研究者や話し言葉に関心を持つ言語学者など）も利用することができる．マルチモーダルコーパスの利用者が増え，知見が蓄積されれば，SC と研究者双方にとって有益であるという考えから，筆者らは 2015 年春に一部のデータを公開すべく準備を進めている．

3 収録方法

本節では，映像・音声データの収録について説明する．詳細は [1] を参照されたい．

未来館内で SC と来館者の対話場面の収録は，国立

表 1: 公開予定のデータセット

データ	形式	備考
音声	wav	業者編集済．映像に格納された音声と同一．
映像	mov, mp4	業者編集済．ファイル形式以外は同一．
動作情報	独自形式	RGB 映像, Depth 映像 (RGB と同期), 骨格情報を提供予定．
音声発話アノテーション	eaf	年度内作業終了予定．eaf は ELAN で用いられる形式．
身体動作アノテーション	eaf	現在作業中．eaf は ELAN で用いられる形式．

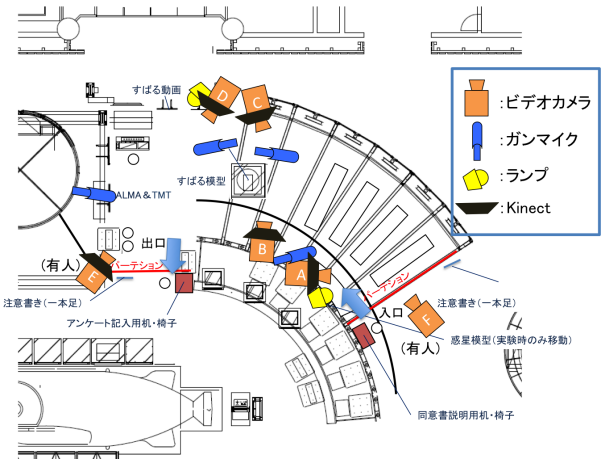

図 2: 収録機材の配置図（[6] より転載）．

情報学研究所倫理委員会からの許可を得たものである．研究者が収録準備を進める一方で未来館側は，多くの SC が「ベテラン」と考える SC（現役，OB/OG）に収録への協力を依頼した．収録に適した「空間のひろがり」「巨大望遠鏡で宇宙の謎に挑む」というふたつの展示フロアを，収録時間のみパーティションで区切り収録スペースを確保した．その内部に 5 台のビデオカメラ（A〜E），ビデオカメラと同じ位置に Kinect を，さらに 4 本の指向性ガンマイクを設置した (図 2)．

収録直前，SC と来館者に対して説明した後，同意書に署名を求めた．SC と来館者は，3 本の無線マイクを胸元に装着する．対話の様子は，前述した固定のビデオカメラに加え，ビデオカメラを持った撮影者（F）が来館者の様子を撮影した．

収録後，来館者と SC にはアンケート形式で回答を求めた．SC にはアンケートに加え，対話終了後に収録した対話の映像をもとに，自身の対話を振り返ってもらい，その様子も 30 分から 1 時間程度収録した．そして，対話に関するログ入力 [4][5] を求めた．

4 対話データ

当該展示フロアを収録のために部分的に区切ることが可能な時間は，1 日につき 1 時間程度であった．全 10 日の収録日を通して，ベテラン SC2 人に 1 日 30 分ずつ，それぞれ 3 組の来館者に対して，できるだけ普段通りに対話をしてもらい，その様子を撮影した．1 組当たりの対話時間や展示物の紹介方法については特に定めなかった．上記の収録により，対話数 35，総対話時間 8 時間 17 分 22 秒，平均対話時間 14 分 30 秒の raw データを得た．

収録した映像と音声は業者によって編集された．編集済みの音声と映像，後述するアノテーションや Kinect から取得したデータを公開する予定である（表 1）．

5 アノテーション

本コーパスでは，データから明らかな情報や，データ公開時に多くの研究者が利用しやすいであろう諸特徴をアノテーションとして付加することを試みている．具体的には，映像と音声を確認しながら，ELAN[3] (Max Planck Institute for Psycholinguistics) を使用してアノテーションを行っている．本節では，音声言語と身体動作について，何をアノテーションの対象とし，どのような手順でアノテーションしているのかを紹介する．

5.1 音声発話アノテーション

本コーパスでは，SC と来館者の発話，咳，笑い声を転記対象とする．まず，ELAN 上で音声発話の開始・終了を同定し，注釈（tier）として切り出す．IPU(Inter-Pausal Unit) の考え方 [13] に準じて，発話内に無音区間が 100msec 以上存在する場合は当該の無音区間直前で一旦区切り，次の再開位置から新たな範囲指定を行う．次に，注釈内にできる限り聞こえた通りに転記する．笑い声，音の延び，聞き取りにくい音声等は，表 2 に従って転記する．

5.2 身体動作セグメンテーション

身体動作は音声発話よりも開始・終了の同定が難しく，また，研究目的によって記法が多様なため，確立され，多用される記法がほとんどないのが実情である．

本コーパスでは，以下に示すセグメンテーションによって注釈すべき区間を定めてから，アノテーションを行っている．最初に，ELAN のテンプレートを作成し，該当する注釈層に 1 秒刻みの幅を設ける（図 3）．これを本稿ではウィンドウと呼ぶ．ウィンドウは，セグメンテーション作業の土台となる．

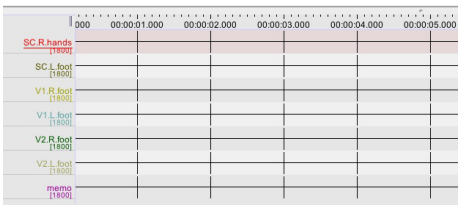

図 3: セグメンテーションの基礎となるウィンドウ．

次に，ウィンドウにタグを付与する．以下では，現在作業を進めている，(1) 手や腕の動き，(2) 展示物間の移動の 2 種類の身体動作について紹介する．

[3] http://tla.mpi.nl/tools/tla-tools/elan/

表 2: 音声発話アノテーション使用記号一覧

記号	意味
,	発話の区切りに付与．明らかに続きが発せられそうな場合．
.	発話の区切りに付与．
?	質問の末尾に付与．
;	セミコロン前に聞こえた通りの表記，後に正規表現を記す．
(h)	笑い声の部分に付与．
ー	長音に付与．
|	100mcec 以上の沈黙がある箇所に付与
( )	聞き取りにくい音声を示す．
[ ]	複数の聞き取りの可能性がある場合，[で; ね] のように示す．

(1) 手や腕の動きについては，ELAN 上で映像を目視し，右手と左手それぞれについてウィンドウ内に 4 種類のタグを付与する（表 3）．4 種類のタグを付与した後，動き（moving）についてはさらに別のタグを付与する（表 4）．

表 3: 手の動きに関するセグメンテーションのタグ一覧

タグ	状態
moving	手や腕が一部分でも動いている
notmoving	手や腕が全く動いていない
missing	動きの有無が判断できない
unclear	上記のいずれのタグも付与できない

表 4: moving に関するタグ一覧

タグ	状態
自律	自律的に動いている
他律	手や腕以外の動きに付随して動いている
不明瞭	上記のいずれのタグも付与できない

(2) 展示物間の移動には，SC と来館者のうち，誰か一人でも他の展示物への移動を行っている場合は “transfer” というタグを，全員がひとつの展示物の前にいる場合は “exhibit” のタグを付与する．

5.3 身体動作アノテーション

本コーパスでは，身体動作に自然言語文のアノテーションを付与する．しかし，インタラクションにおける身体動作はそれ自体を自然言語で記述することが困難である．たとえば,「展示物を指す」という行為は,「前方に向けて腕を伸ばす」とも「注目すべき対象を指示する」とも記述することができる．そのため，数ある行為の中から，まずは何を転記対象とし，どのような粒度でアノテーションを付与していくかを決めた上で作業を進める．

現時点では，前述のセグメンテーションにより注釈区間を定めた後，当該行為（例：指差し，移動など）を，簡潔な動詞 (例：指す，下がる，など) で表記する．これを上位層とみなす．次に，上位層を構成していると考えられる動作を，誰もが直観的に理解可能な粒度（例：右手を伸ばす，一歩後ろに引く）で，下位層に記述する（図 4）．

図 4: 身体動作アノテーション例

6 異分野融合研究の展望

映像と音声だけでなく，三次元情報やアノテーションが加わることで，どのような利益があるのか．本稿では，インタラクション分析を主軸に，自然言語処理，画像処理それぞれの領域との融合研究の展望を簡単に紹介する．

自然言語処理の分野では，多くの場合，新聞記事や小説などから集められるテキストコーパスが利用されてきた．近年，自然言語処理の分野でも，テキストや音声のみならず，画像を解析してテキストを生成する研究 [11] や，言語獲得において非言語情報を利用する研究 [12] など，マルチモーダルデータを利用した研究が行われている．

本データの利用の一例としては，音声発話アノテーションや身体動作アノテーションで与えたテキストを，映像・音声データから自動生成するという研究が考えられる．例えば，身体動作を表すテキストが自動生成することができれば,「指差し」などの特定の動作やその状況を検索したり，映像中の様々なイベントを要約して提示したりすることができるようになる．現在は，インタラクション研究者は全ての映像を網羅的に見て分析を行っているが，この作業の一部を，自然言

語処理技術で効率化することが期待される．

画像処理との融合については，Kinect により得られた三次元映像ならびに骨格情報と，これまでジェスチャー研究で行なわれてきた空間における人の身体配置の変化 [14][7] を結びつけた研究が考えられる．これまで，Kendon が提唱する身体配置の変化とその場で行なわれている対話内における活動（説明の開始や終了，次の展示物に移動する等）との関連は，目視による分析が行われてきた．このような人間の作業からなる分析を三次元映像により自動解析できれば，ジェスチャー研究において，これまで以上に大量のデータを用いて特定の現象を集めて分析することが可能になると考えられる．

7 おわりに

本稿では，日本科学未来館の展示フロアにおける科学コミュニケーションを対象としたマルチモーダルコーパス設計の概要とその利用可能性について紹介した．本コーパスは 2015 年の春にデータの一部を公開する予定である．公開後も，マルチモーダルコーパスに新たなアノテーションを追加したり，ユーザーによるフィードバックシステムを構築し，マルチモーダルコーパスがより発展可能なしくみを構築予定である．

謝辞

日本科学未来館および収録に参加してくださった皆様，らくだスタジオ，アノテーション作業者の阿部廣二さん，河村美雪さんに感謝する．本研究は，JST さきがけ，国立情報学研究所グランドチャレンジ「ロボットは井戸端会議に入れるか」，学融合推進センター学融合研究事業「科学技術コミュニケーションの実践知理解に基づくディスカッション型教育メソッドの開発」，および科学研究費助成事業 (学術研究助成基金助成金) 挑戦的萌芽研究「知識伝達インタフェースとしての科学コミュニケーターの活動実践の理解と支援」の助成による．

参考文献

[1] 城綾実，牧野遼作，坊農真弓，高梨克也，佐藤真一，宮尾祐介: 異分野融合によるマルチモーダルコーパス作成-展示フロアにおける科学コミュニケーションに着目して-, *SIG-SLUD-B401*, 71, pp. 7–12 (2014).

[2] 坊農真弓, 高梨克也, 緒方広明, 大崎章弘, 落合裕美, 森田由子: 知識共創インタフェースとしての科学コミュニケーター：日本科学未来館におけるインタラクション分析, ヒューマンインタフェース学会論文誌, No.15, Vol.4, pp. 375–388 (2013)

[3] Mayumi Bono, Hiroaki Ogata, Katsuya Takanashi, and Ayami Joh: The practice of showing ‘ who I am ’: A multimodal analysis of encounters between science communicator and visitors at science museum, In *Universal Access in Human-Computer Interaction*, Vol. 8514, pp. 650–661 (2014)

[4] 緒方広明, 毛利考佑, 坊農真弓, 城綾実, 高梨克也, 大崎章弘, 落合裕美, 森田由子: ラーニングログシステムを用いた実践知の共有・活用支援における Learning Analytics の役割: 日本語学習と科学コミュニケーションを例として, 日本教育工学会第 *29* 回全国大会論文集, pp. 67–70 (2013)

[5] Ogata, H., Mouri, K., Bono, M., Joh, A., Takanashi, K., Osaki, A., Ochiai, H., Morita, Y. : Analysis of ubiquitous learning logs in the context of science communications in a museum, 4th International Workshop on Technology-Transformed Learning: Going Beyond the One-to-One Model?, In *Conjunction with ICCE 2013*, pp. 74–79 (2013)

[6] 城綾実, 坊農真弓, 高梨克也: 科学館における「対話」の構築: 相互行為分析から見た「知ってる？」の使用, 認知科学, 22(1) (2015 年 3 月発行予定)

[7] 牧野遼作，坊農真弓，古山宣洋: フィールドにおける語り分析のための身体の空間陣形：科学コミュニケーターの展示物解説行動における立ち位置の分析, 認知科学, 22(1) (2015 年 3 月発行予定)

[8] Schegloff, E. A.: *Sequence Organization in Interaction*, Cambridge: Cambridge University Press (2007)

[9] Sidnell, J. and Stivers, T. (eds.): *The Handbook on Conversation Analysis*, Wiley-Blackwell (2012)

[10] 坊農真弓, 高梨克也（編）／人工知能学会編: 多人数インタラクションの分析手法, オーム社 (2007)

[11] Rashtchian, C., Young, P., Hodosh, M., and Hockenmaier, J.: Collecting image annotations using amazon ’s mechanical turk, In Workshop on Creating Speech and Language Data with Amazon ’s Mechanical Turk. (2010)

[12] Johnson, M., Demuth, K., and Frank, M.: Exploiting social information in grounded language learning via grammatical reductions, In Proc. ACL 2012.

[13] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., Den, Y.: An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogues. *Language and Speech*, 41, pp. 295–321 (1998)

[14] Kendon, A.: Spatial organization in social encounters: The F-formation system, In *Conducting Interaction: Patterns of Behavior in Focused encounters*, Cambridge University Press, pp. 209–237 (1990)