領域ウェブサイト
https://research.nii.ac.jp/EmSemi/index.html

オンライン研究会 @ Zoom (2023.02.17)

コーパスの構築・利用と個人情報保護

2022年度科研費・学術変革領域研究(B)

「言語相互行為における手話と身振りを対象とした身体記号学」
(研究代表: 坊農真弓)

「*情報処理技術を活用した文理融合次世代コーパスの構築に基づくモダリティ横断*」
(データ統合班・研究代表: 菊地浩平)

本日の予定

15:00-　　オープニング

15:05 - 15:25　話題提供1　人文学研究での個人情報保護対応の経験から

菊地浩平 (筑波技術大学)

15:25 - 15:45　話題提供2　『日本語日常会話コーパス』構築・公開の経験から

小磯花絵 (国立国語研究所)

(休憩)

16:00 -　　全体ディスカッション

宍戸常寿 (東京大学)，加藤尚徳 (KDDI総合研究所)

話題提供者，本日の参加者のみなさま

16:50 -　　クロージング

本日の研究会趣旨

- コーパスは，収集した言語データを一定の枠組にそって整備し，検索可能な状態にしたもの
- 複数の組織だけでなくディシプリンをまたいでコーパスを構築する，利用するということが増えてきている
- コーパスの構築・利用を考えたとき，私たちは個人情報保護とどのように向き合えばよいのか？
  - 研究倫理としての個人情報保護
  - 個別の事情によるテクニカルな問題
  - 具体的に求められる対応の精査

話題提供1

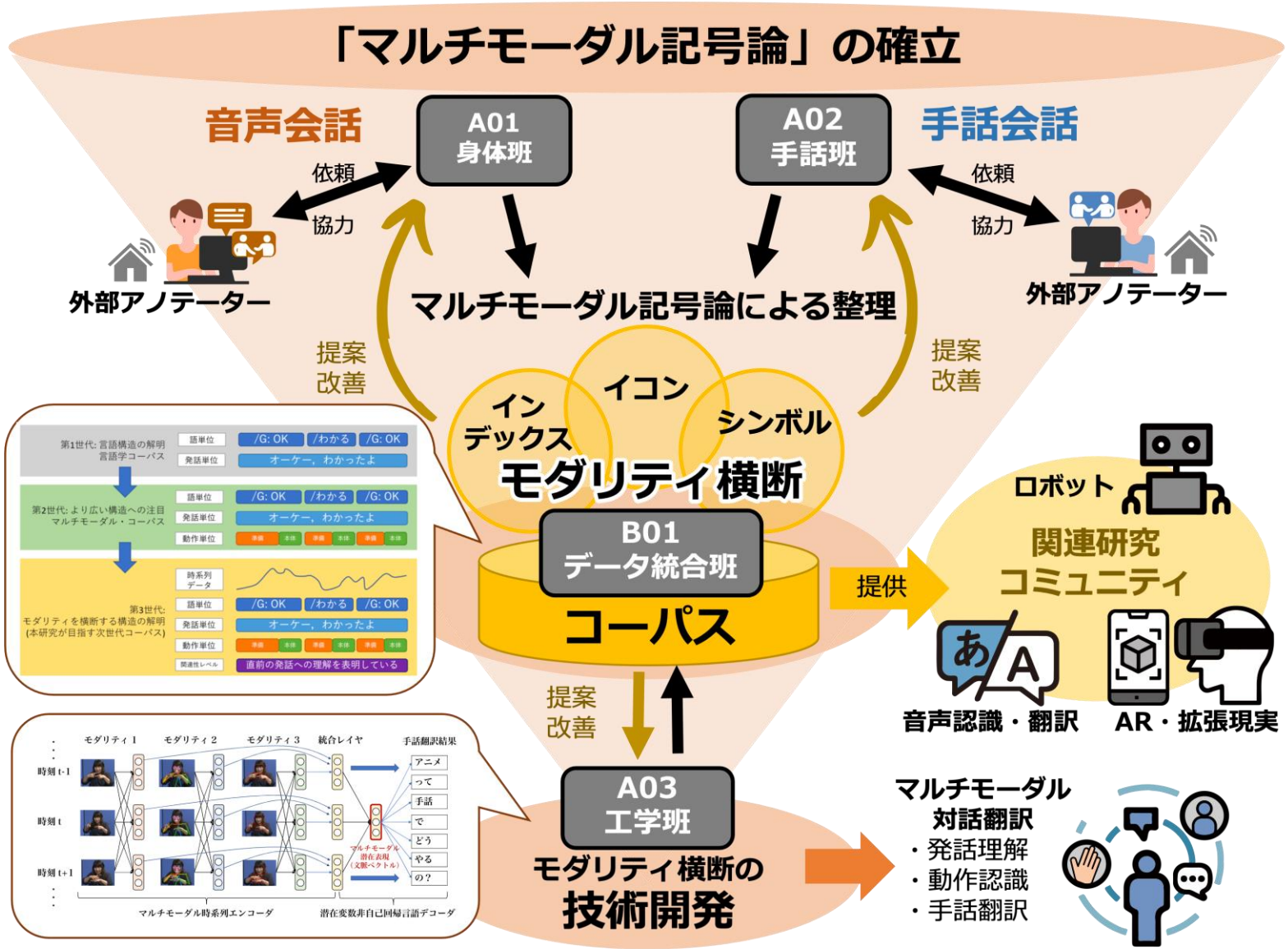
人文学研究での個人情報保護対応の経験から

菊地浩平（筑波技術大学）

私の立ち位置

- 専門: コミュニケーション科学
  - 音声や手話による諸活動，手話通訳者が含まれる諸活動を対象とした研究
  - 相互行為分析の手法やインタビューを用いた研究手法
- データの収集・管理・利用の経験
  - 主に個人または小規模グループでクローズドな扱いのものがメイン
  - 同じデータをいろいろな研究で使う，何年も何十年も使うということもある
  - コーパスの構築に関わるようになったのはだいたい10年ほど前から

この領域の全体図


「マルチモーダル記号論」の確立
音声会話
A01
身体班
A02
手話班
手話会話
依頼
協力
依頼
協力
外部アノテーター
外部アノテーター
マルチモーダル記号論による整理
提案
改善
提案
改善
イコン
インデックス
シンボル
モダリティ横断
B01
データ統合班
コーパス
提供
ロボット
関連研究
コミュニティ
音声認識・翻訳
AR・拡張現実
提案
改善
A03
工学班
モダリティ横断の
技術開発
マルチモーダル
対話翻訳
・発話理解
・動作認識
・手話翻訳

コーパスの3つのフェーズ

集める → 作る → 公開する・利用する

このフェーズに即して私たちの領域での活動を位置づけてみると

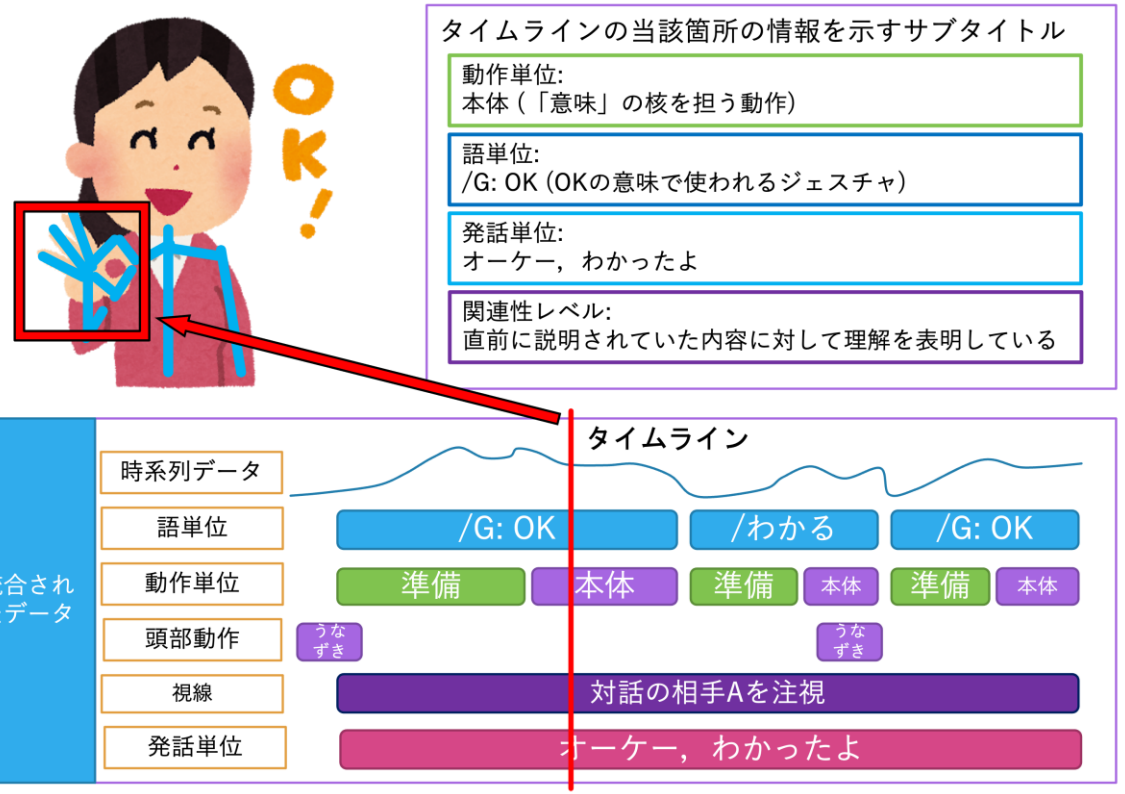
- 新規に次世代コーパスを**作る**にあたって，まず最初にスキームを検討したい
- **集め**られたデータの取り扱いについて，協力者との間で適切な合意ができている既存コーパスがある
- その既存コーパスを議論のたたき台として**使う**+研究プロジェクト内で共有したい

こういったケースに対応することを考えると，コーパスの利用者・構築者は具体的に何に注意しなければならないのか

この学術変革領域研究で

作ろうとしているものの特性

- 文理融合次世代コーパス
  ＝モダリティ横断コーパスを作る
- 本研究領域では手話言語や，身振りを含むコミュニケーション場面が対象
- 言語だけでなく非言語要素や身体動作，それらが特徴量として数値化されたデータ，動作がコミュニケーションで果たしている機能についての注釈などが含まれることになる



ローデータが重要

- 手話言語の場合，表情や視線は言語情報の一部を構成しているため匿名化処理ができない
- 音声言語でも周辺言語的要素は重要なデータであり，匿名化できないもの・すると価値が失われるものがある
- 人と人のコミュニケーションにおいて，身振りは重要なコミュニケーションの資源となっている
- 映っているもの全てが研究の対象になりうる

研究協力者との合意

- データを集めるにあたって，研究協力者（データの提供者）との間で，例えば以下のような項目について合意を作る努力はこれまでにもされてきている
- 個人情報が含まれるデータについて
  - どう保管するか　→ 意図しない流出への対応も含めた管理
  - 誰が使ってよいか　→ 研究者本人，研究チーム，研究補助者，第三者
  - 何に使ってよいか　→ 研究 (学会発表，論文投稿，研究会)，教育 (授業)
  - どう使ってよいか　→ 無加工，トレース，モザイク，仮名化 etc.
  - いつまで使ってよいか　→ 研究終了まで，〇年間，期限を定めない etc.
  - 対応が必要な問題が起きた場合どうするか (解決手段・連絡方法の確認など)

公開・非公開にかかわらず

既存のデータやコーパスを利用する・公開するときに気になること

- 個人単位・小規模チーム単位での研究の場合
  - 協力者との合意に基づく研究活動の全体を把握するのはそれほど難しくない
  - データへのアクセスをコントロールするのも難しくない
- 一方でデータへのアクセスを広く可能にする場合は，データを誰にでもアクセスできるようにするために，果たすべき責任に自覚的にならないといけない
  - 研究に限られるとはいえ，不特定多数が、自分が集めた・作った・公開したデータを利用するという事態への対応を，あらかじめ想定しておく必要がある
  - ただし，この対応は時限付きのプロジェクトレベルでは難しいのでは

- データの扱い方の適切さ（個人情報の扱いも含む）は，どう保証するのか？（ディシプリンなのか，法律なのか，あるいはその両方なのか）
- 例えば「合理的な範囲での個人情報の利用目的変更」かどうかは誰が判断するのか？
- 各組織の倫理審査委員会で個人情報保護対応について判断が分かれる場合や，運営の方法が異なる場合，どこで何をどう集約すればよいのか？する必要があるのか？