NINJAL
National Institute for Japanese Language and Linguistics

コーパスの構築・利用と個人情報保護

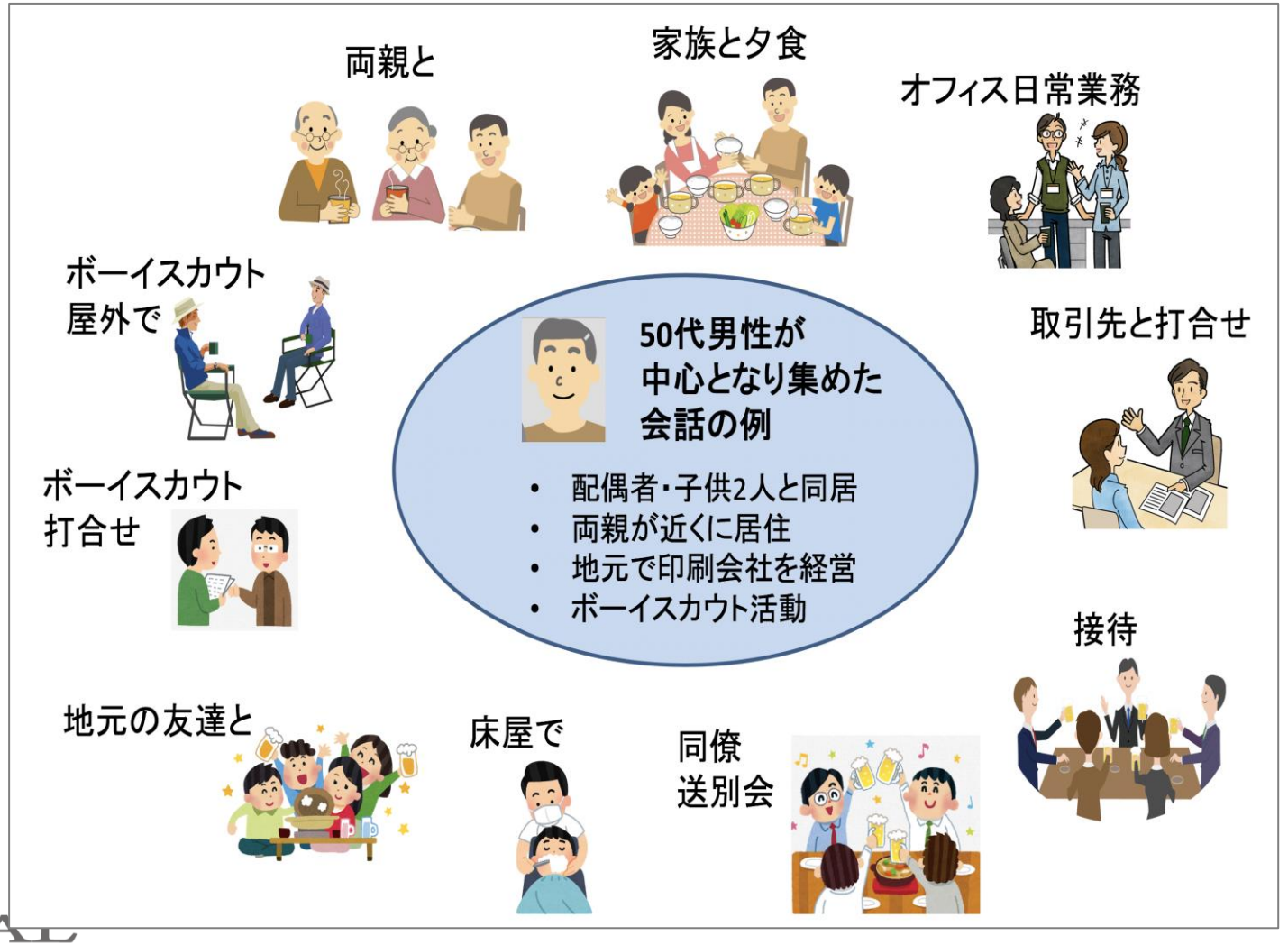
『日本語日常会話コーパス』構築・公開の経験から

小磯 花絵

大学共同利用機関法人 人間文化研究機構
国立国語研究所
National Institute for Japanese Language and Linguistics

『日本語日常会話コーパス』

日常生活の多様な場面・話者による会話200時間を映像を含めて公開



調査協力者40名を中心に多様な場面・話者との会話を2016～2021年に収録

コーパスに含まれるデータ

	有償版		「中納言」
	全体	コア(20 時間)	
映像データ	○	○	×
音声データ	○	○	検索箇所前後の再生のみ
転記テキスト	○	○	×
形態論情報（短単位情報）	○	○	○
形態論情報（長単位情報）	○	○	○
談話行為情報	×	○	×
韻律情報	×	○	×
係り受け情報	×	○	×
会話・話者に関するメタ情報	○	○	△（備考情報等除く）

大規模な日常会話コーパス構築に向けた準備プロジェクト（2014～2015）

① 均衡性を考慮した会話コーパスの設計

- 会話行動調査（250～300人）の実施
- 調査に基づくコーパスの設計

② 種々の日常会話を収録するための方法論の整備

- 収録機器（レコーダーなど）の選定・改良
- 収録データの加工技術（背景音低減、映像加工など）の検討
- 映像データの収録・公開に伴う倫理的・法的問題の検討

③ 日常会話を適切・効率的に転記するための方法論の整備

- 転記基準の策定
- 転記作業量の試算

- 方針：会話参加者の顔にボカシをかけずに公開
- 対応：会話参加者には、事前に、映像を記録すること、
  顔にボカシをかけずに映像を公開することの承諾を得る

国語研究所と**契約を交わした者に限定して公開**する場合、
会話参加者の顔にぼかしなどの加工は加えません。
研究者など限られた利用にとどまります。

会話参加者の顔以外に問題となるケースを
収録したデータに基づき洗い出し分類

映像データ公開時の問題の整理

- **公表著作物の写り込み**
  - テレビの画面や音楽
  - 書籍やパンフレット
  - 公開のwebサイト画面、など

- **非公表著作物の写り込み**
  - 打合せの内部資料
  - 個人で撮影した写真、など

- **上記以外で個人情報に関わるもの**
  - 収録・公開の同意を得ていない第3者*の顔
  - 個人を特定しうるもの（例：氏名・住所・カードナンバー）
  - スケジュール帳など

- **その他の問題となりうる行為・話題の例**
  - 車のスピード違反、タバコのポイ捨て
  - 未成年の飲酒に関する話題、など

著作物の「写り込み」等に係る規定

プライバシー権
肖像権
個人情報保護法
など

その他

* 同意を得ていない第3者＝同意書を交わしていない人（店員・お客・他の旅行客など）

個人情報等に関連する法律・権利

- 判例上成立（憲法13条幸福追究権がベース）
  - プライバシー権：私生活をみだりに公開されない権利
  - 肖像権：自己の容貌等をみだりに撮影・公表されない権利
  - パブリシティ権：芸能人の肖像など経済的価値を保護する権利
- 個人情報保護法
  - 個人情報の適切かつ効果的な活用などその有用性に配慮しつつ、個人の権利利益を保護することを目的とする法律

肖像権との関係

以下の各要件を総合的に見て肖像利用の受忍限度（肖像権侵害とみなすか否か）を判断

- 被撮影者の社会的地位：公人・芸能人の方が一般人より受忍限度は低い（肖像権侵害と判断されにくい）
- 被撮影者の活動内容：一般的な活動(低) ←→ センシティブな活動(高)
- 撮影の場所：一般に公開された場所(低) ←→ 私的・閉鎖的な空間(高)
- 撮影の目的：公共性が認められる場合(低)
- 撮影の態様：一般的な方法(低) ←→ 隠し取り(高)
- 撮影の必要性：目的との関係において必要性・必然性が低い場合(高)

『日本語日常会話コーパス』の場合

- 公の場所において(「撮影の場所」)
- 普通の行動をしているところを(「被撮影者の活動内容」)
- 研究教育目的である日常会話コーパス構築という公共性の高い目的のために(「撮影の目的」)
- 隠し撮りなどではなく通常の方法で(「撮影の様態」)
- 日常生活における会話の記録のために必要となる範囲(「撮影の必要性」)

を収録するものについては、**肖像権の非侵害に傾きやすい**

※ 研究教育利用・商業利用(統計利用のみ)に限定し、その利用目的のもとで国語研究所と契約した人にのみ提供
(インターネットやSNSなど，拡散可能性の高い公開方法ではない)

対応の例

ボカシ対象外の例

- 公的な場（店舗・役所・公道など）で一般的な行為をしている第3者の写り込み
- 公道を走っている車のナンバープレート
- 一般に公開されているブログなどの写り込み

ボカシ等の対象の例 (いずれも認識可能な程度のサイズ・鮮明さの場合)

- 一般に人の出入りが自由ではない場（小中学校など）やセンシティブな場所（病院など）での人の写り込み
- 社会的に見て保護されやすい対象（乳幼児、児童、障害者など）の場合は特に配慮
- 問題となる行為をしている人の写り込み
- 自宅に停めている車のナンバープレート・自宅の表札
- 個人の手帳・内部資料・プライベートな写真など

個人情報保護法

個人情報:「生存する個人に関する情報であって，当該情報に含まれる氏名，生年月日その他の記述等により特定の個人を識別することができるもの（他の情報と容易に照合することができ，それにより特定の個人を識別することができることとなるものを含む）」

3つの義務

① 利用目的の特定・通知

② 安全管理措置義務

③ 第三者提供の制限

利用目的の特定・通知

会話参加者の方への説明文書

1. 研究課題名： 大規模日常会話コーパスに基づく話し言葉の多角的研究

2. 研究の目的・意義

この研究では、日常の会話場面で私たちがどのような言葉遣いや振る舞いをしているかを調べるために、会話の映像と音声を記録し、研究に適した形で整えた上で、会話のデータベースとして公開します。こうしたデータベースは、たとえば次のようなことを調べるのに役立ちます。

- 場面や相手によって言葉の使用はどのように変わるか
- 書き言葉と比べて会話の言葉はどのような特徴があるか
- 昔の会話の言葉と比べて言葉の使い方や声の抑揚などはどのように変化しているか
- 視線やジェスチャーなどの非言語行動はどのような役割を果たしているか

日常場面での営みの多くは会話を通じてなされます。そのため、会話における言葉遣いや振る舞いの特徴を調べることは、言葉の理解だけでなく、会話を通したコミュニケーションや対人関係などの理解につながります。また、昔の会話を見聞きして当時の生活や文化を知ることができるように、こうしたデータベースは、後世の人々が 21 世紀初頭の日本の生活や文化を知るための貴重な記録となります。

利用目的の特定・通知

国立国語研究所
日常会話 調査研究

■ どんな研究？

国語研究所では、さまざまな会話場面の映像と音声を収録し、日常の会話場面で私たちがどのような言葉遣いや振る舞いをしているかを調べています。収録した映像・音声データは、会話のデータベースとして整備・公開することを予定しています。

■ 何が分かるの？

会話のデータベースがあると、たとえば次のようなことを調べることができます。

- 場面や相手によって言葉の使用はどのように変わる？
- 書き言葉と比べて会話の言葉はどのような特徴がある？
- 昔の言葉と比べて言葉の使い方や声の抑揚などはどのように変化している？
- 視線やジェスチャーなどの非言語行動はどのような役割を果たしている？

■ 何の役に立つの？

- 会話の言葉遣いや振る舞いの特徴を調べることは、単に話し言葉のしくみの理解にとどまらず、会話を通したコミュニケーションや対人関係などの理解につながります。
- 場面や相手による言葉の使い方の違いなどが分かると、外国人が日本語を学ぶときの手助けになります。

- 今、人間と会話をするロボットが出てきました。私たち人間の日常会話を数多く納めたデータベースは、人間と会話するロボットの性能向上につながります。

- 昔の会話を見聞きして当時の生活や文化を知ることができるように、後世の人々が21世紀初頭の日本人の生活や文化を知るための貴重な記録となります。

◆◆◆◆◆◆◆◆◆◆ お問い合わせ先 ◆◆◆◆◆◆◆◆◆◆

この調査に関してご質問やご相談などありましたら、下記までお問い合わせください。

担当：国立国語研究所　日常会話コーパスグループ
メールアドレス：
電話：
代表 小磯花絵

プロジェクトのWebサイト
http://pj.ninjal.ac.jp/conversation/corpus.html

NINJAL
National Institute for Japanese Language and Linguistics

個人情報保護法

個人情報:「生存する個人に関する情報であって，当該情報に含まれる氏名，生年月日その他の記述等により特定の個人を識別することができるもの（他の情報と容易に照合することができ，それにより特定の個人を識別することができることとなるものを含む）」

3つの義務

① 利用目的の特定・通知

② 安全管理措置義務

③ 第三者提供の制限

収録・公開に関する同意書

私は、XX氏から、「大規模日常会話コーパスに基づく話し言葉の多角的研究」に関する説明を受け、この調査への参加、および、この調査において記録された私の映像・音声・文字化資料、研究用情報、フェイスシートに記入した情報、会話状況情報等の公開について、以下の条件のもとに同意します。

私の名前、学校や会社など私が所属する組織の名称、自宅・所属組織の住所・電話番号~~などの個人情報~~

① 国立国語研究所が定める研究教育利用・商業利用（統計情報利用）の条件に同意し契約を交わした者に対して公開する際には、データに以下の処理をほどこす。

- 私の名前、学校や会社など私が所属する組織の名称、自宅・所属組織の住所・電話番号の音声が聞こえないように加工し、文字化資料においても、仮名や伏せ字に置き換えるなどの処理をする。
- 私が公開を望まない箇所の音声・文字化資料も同様に加工する。

② 第1項以外の方法での公開については、上記に加え、顔の一部にぼかしを加えるなど、私個人が特定できないように映像を加工する。また、会話全体ではなく短いシーンごとの映像・音声の公開に留める。

同意書の範囲外の個人情報等の扱い

「会話者の名前，所属組織名，自宅・所属組織の住所・電話番号」および本人が公開を希望しない箇所以外はそのまま公開？

パスポート番号や保険証番号など上記に類するものや、発話内容・まわりの映像等から自宅の場所が特定できるものなども対応

ヒアリングで更に意志を確認

■ どうやってデータを公開するの？

◇映像

国語研究所と契約を交わした者に限定して公開する場合、
会話参加者の顔にぼかしなどの加工は加えません。
研究者など限られた利用にとどまります。

上記以外の方法で公開する場合、
左の図のように、全ての会話参加者の顔にぼかし処理を施します。日本語学習者や学校の先生などの幅広い利用が見込まれます。

なお、会話全体ではなく短いシーンごとの映像・音声の公開に留めます。

◇音声・テキスト

個人情報（お名前・所属組織名など）やご本人が公開を望まない箇所は、次のように加工します。

- 音声：聞こえないように加工（ホワイトノイズ処理）
- 文字化テキスト：仮名（かめい）や伏せ字（××）に置き換える

仮名に置き換えた場合の例
音声は聞こえないように加工

実際の会話の例

あっ 山本さん 醤油 そこの棚に置いてあるんだけど ちょっと取ってくれる？ あと ついでに 太田さんに醤油皿もお願い

公開用に加工したテキストの例

あっ 横山さん 醤油 そこの棚に置いてあるんだけど ちょっと取ってくれる？ あと ついでに 鈴木さんに醤油皿もお願い

コーパスの公開

- 利用許諾契約書

（禁止事項）第8条 乙は、日常会話コーパスの利用にあたり、以下に定める行為をしてはならない。

① 日常会話コーパスの全部又は一部を、本目的のために必要な範囲を超えて複製または改変すること。

② 本目的のためであるか否かに関わらず、日常会話コーパスの全部もしくは一部、又は日常会話コーパス等の類似物を譲渡、貸与、販売、配布、上映、公衆送信または刊行すること。

<略>

⑥ 日常会話コーパス及びサンプルデータを用いて甲または第三者の名誉等を毀損し、あるいはその他の権利を侵害すること。

⑦ 甲が予め伏字にした情報を復元・公表すること。

<略>

⑪ 記録された話者情報以外の話者に関する情報を公開すること。また、それを利用することによって他の利用者が日常会話コーパスによる記録以外の話者情報を取得することのできる情報を公開すること。なお、第１２条に定める研究成果の公表に付随するものであってもこれを認めない。

利用ガイドラインの作成･配布

- 利用目的の範囲･利用方法
- データの管理
- 研究成果の公表
- 研究成果の報告
- 授業や演習等での利用について

https://www2.ninjal.ac.jp/conversation/cejc/guideline.html

まとめ

- それぞれ関連する法律や権利のもとで適切に対応

公表著作物	著作物の写り込み規定（30条の2）、など
個人情報に関わるもの	プライバシー権、肖像権、個人情報保護法、など

- 同意書を交わした人は、その同意書に記された条件が優先される

⇒ 同意書の文言は非常に重要

⇒ 同意取得方法の適切性

⇒ オプトアウトの機会を設ける

- 法律や同意条件だけ守っていればよいということではない
- 公開後の適切な扱いを守るために最善を尽す
- 対応の方針を定め、ガイドラインとしてまとめて公開

関連研究者との情報共有

小磯花絵・伝康晴

「『日本語日常会話コーパス』データ公開方針：法的・倫理的な観点からの検討を踏まえて」

『国立国語研究所論集』第15号, pp.75-89, 2018.7.

http://doi.org/10.15084/00001518