NINJAL
National Institute for Japanese Language and Linguistics

『日本語日常会話コーパス』の設計と構築

小磯 花絵
国立国語研究所

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

主要な日本語会話コーパス

コーパス名	規模	概要	音声
千葉大学地図課題コーパス	128会話23時間	地図を用いた課題遂行対話	有
女性のことば・職場編 **男性のことば・職場編**	各21名	職場のフォーマル・インフォーマルな場面の自然談話	無
CALL HOME Japanese	120会話20時間	アメリカ在住日本人と国内の家族・友人との電話会話	有
CallFriend Japanese	31会話	アメリカ在住の日本人同士の電話会話	有
千葉大学3人会話コーパス	12会話2時間	大学生の会話（話題指定）	有
名大会話コーパス	161名100時間	親しい者同士の雑談	無
BTSによる多言語話し言葉コーパス	249会話70時間	友人同士の雑談、教師学生面談会話、電話会話など	一部
さくらコーパス	18会話	大学生の会話（話題指定）	有

主要な日本語会話コーパス

- 話題・収録の状況：
  - ✓ 実験環境
  - ✓ 集まってもらった状況での雑談
  - ✓ 日常生活で実際に行われた会話（少）
- 会話の種類：
  - ✓ 偏った種類の会話・話者（例：電話会話、大学生の雑談、など）
  - ✓ 多様な種類の会話（少）
- 公開データ：
  - ✓ 転記テキストだけ
  - ✓ 転記テキスト ＋　音声データ
  - ✓ 転記テキスト ＋　音声データ　＋　映像データ（少）

『日本語日常会話コーパス』

Corpus of Everyday Japanese Conversation, CEJC

- プロジェクト：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
- 期　間：平成28年4月～平成34年3月（6年間）
- 概　要：日常場面で自発的に生じた会話200時間を収録した日常会話コーパスを構築し、それに基づく分析を通して、日常会話を含む話し言葉の特性を、レジスター･相互行為･経年変化の観点から多角的に解明する。
- 対　象：日常生活の中で自発的に生じたリアルな会話
- 規　模：日常会話 200時間
- 設　計：会話行動調査に基づき話者の属性や場面などの観点から均衡性を考慮（＝多様な会話をバランスよく納めるよう配慮）して設計

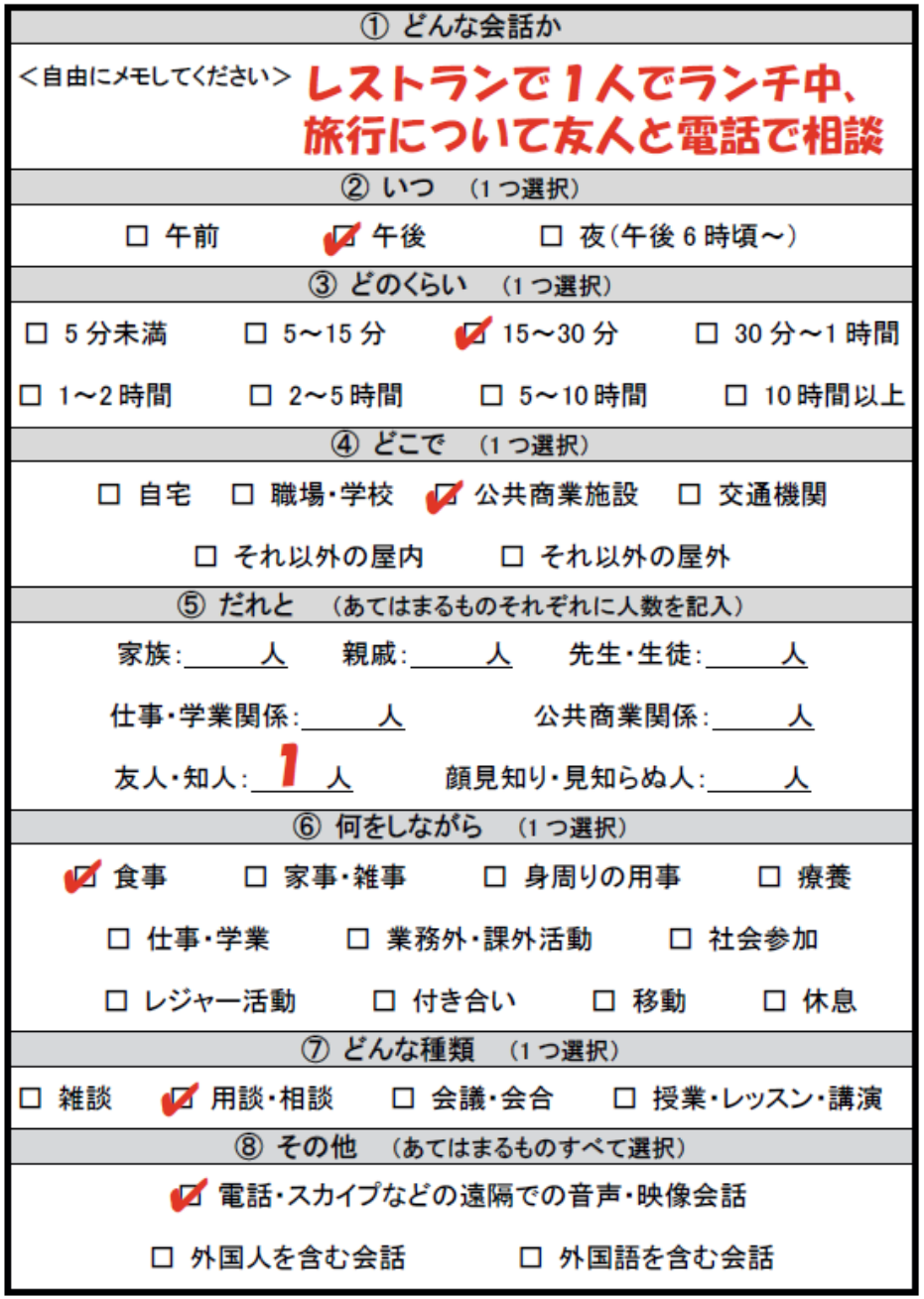
本プロジェクトで目指す会話コーパス

- 多様な場面・話者の会話を記録
  ⇒ ①均衡性を考慮したコーパス設計
- 日常場面で自発的に生じたリアルな活動を記録
  ⇒ ②収録法
- 音声データ・映像データを記録・公開
  ⇒③映像・音声データの収録方法
  ⇒④映像データの公開に向けた法的・倫理的問題
- 形態論情報などの基本的なアノテーションを施す
  ⇒⑤予定しているアノテーションの種類（軽く）

①均衡性を考慮したコーパス設計

会話行動調査

- 目的：日常会話の多様性を明らかにし，それに立脚して多様な日常会話をバランスよく納めたコーパスを設計
- 実施時期：H26.11～H27.2
- 対象：243人
  （年齢・性別バランス）
- 調査日：平日2日・休日1日
  （計3日／1人）
- 総会話数：9272会話



① どんな会話か

＜自由にメモしてください＞ レストランで1人でランチ中、旅行について友人と電話で相談

② いつ （1つ選択）

☐ 午前 ☑ 午後 ☐ 夜（午後6時頃～）

③ どのくらい （1つ選択）

☐ 5分未満 ☐ 5～15分 ☑ 15～30分 ☐ 30分～1時間

☐ 1～2時間 ☐ 2～5時間 ☐ 5～10時間 ☐ 10時間以上

④ どこで （1つ選択）

☐ 自宅 ☐ 職場・学校 ☑ 公共商業施設 ☐ 交通機関

☐ それ以外の屋内 ☐ それ以外の屋外

⑤ だれと （あてはまるものそれぞれに人数を記入）

家族：＿＿人 親戚：＿＿人 先生・生徒：＿＿人

仕事・学業関係：＿＿人 公共商業関係：＿＿人

友人・知人：1人 顔見知り・見知らぬ人：＿＿人

⑥ 何をしながら （1つ選択）

☑ 食事 ☐ 家事・雑事 ☐ 身周りの用事 ☐ 療養

☐ 仕事・学業 ☐ 業務外・課外活動 ☐ 社会参加

☐ レジャー活動 ☐ 付き合い ☐ 移動 ☐ 休息

⑦ どんな種類 （1つ選択）

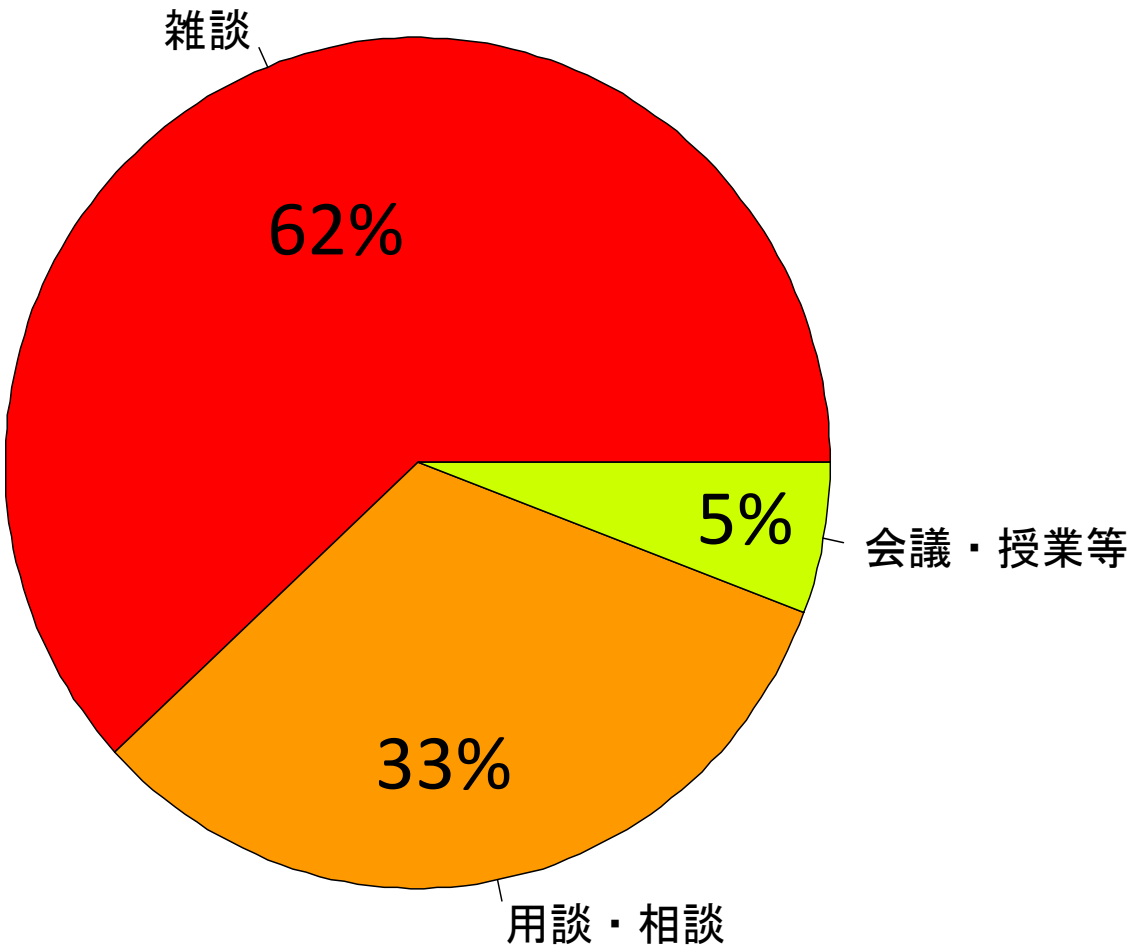
☐ 雑談 ☑ 用談・相談 ☐ 会議・会合 ☐ 授業・レッスン・講演

⑧ その他 （あてはまるものすべて選択）

☑ 電話・スカイプなどの遠隔での音声・映像会話

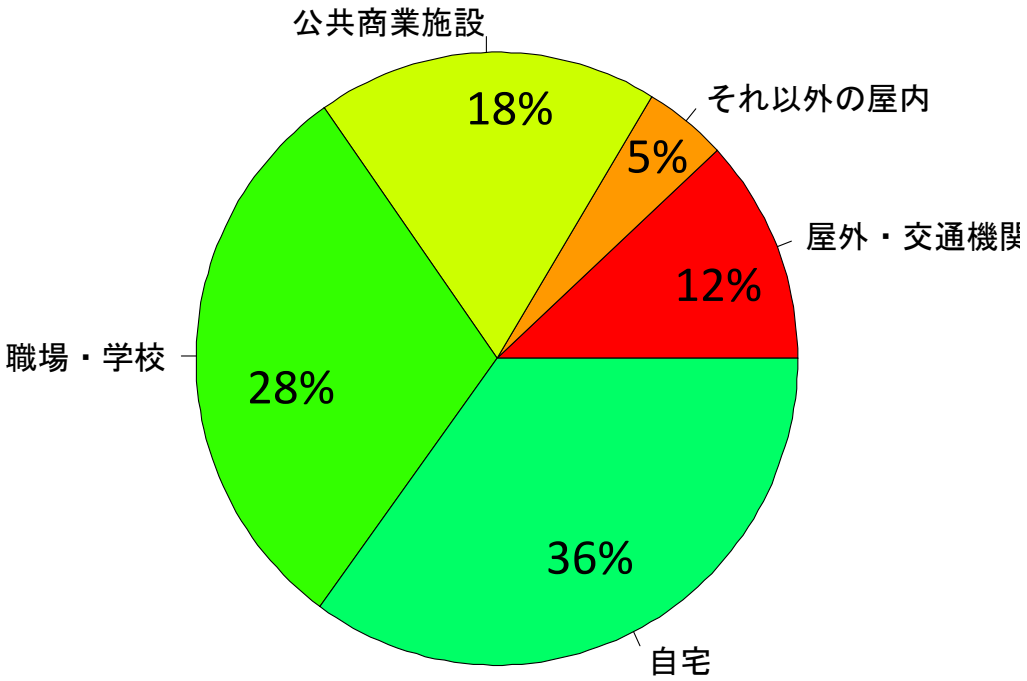
☐ 外国人を含む会話 ☐ 外国語を含む会話

調査結果①

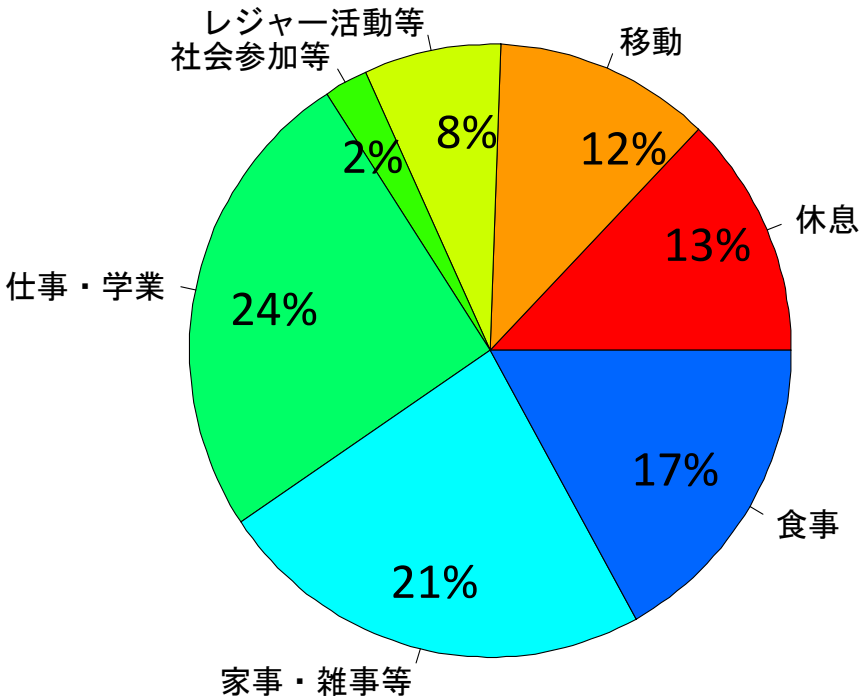


調査結果②

場所


公共商業施設
18%
それ以外の屋内
5%
屋外・交通機関
12%
職場・学校
28%
36%
自宅

活動


レジャー活動等
社会参加等
8%
移動
12%
2%
休息
13%
仕事・学業
24%
17%
食事
21%
家事・雑事等

調査結果に基づくコーパス設計方針

- 会話の形式の構成比を参考に全体の構成比を決定

（例） 雑談:用談相談:会議会合 ＝ 6割 : 3～3.5割 : 0.5～1割

（時間数）6.5～7割 : 2割 : 1～1.5割

- 形式ごとに，場所と活動の観点から構成比の目安を算出

場所／活動		形式								
		雑談			用談・相談			会議・授業等		
		私的	公的	それ以外	私的	公的	それ以外	私的	公的	それ以外
件数	食事・休息	30%	5%	10%	5%		5%	5%		
	仕事・学業	15%	10%	5%	15%	40%	20%	0%	70%	10%
	その他	5%		20%	0%	0%	15%	0%	0%	15%
長さ	食事・休息	30%	5%	20%	10%		5%	5%		
	仕事・学業	10%	5%	5%	10%	50%	15%	0%	70%	10%
	その他	5%		20%	0%	0%	10%	0%	0%	15%

⇨ この比率を遵守するのではなく、大きな偏りが生じないよう、一つの指針とする

※調査結果の詳細はプロジェクトHPで公開　検索「日常会話コーパス」

② 収録法
③ 映像･音声データの収録方法

日常会話コーパスの収録方法

■ 個人密着法

- ✓ 性別・年齢などの観点からバランスを考慮して選別された協力者に収録依頼（首都圏在住者、男女×年齢5世代×各4-5人=40-50人、職業偏らないよう配慮）
- ✓ 機材機器等を2-3か月ほど貸し出し，協力者の日常生活で自発的に生じるリアルな会話を記録（1協力者あたり平均約15–18時間収録）
- ✓ コーパス構成比や倫理的問題等を考慮してコーパスに含める会話を選別
  - 1協力者あたり約4-5時間を選別，計160-180時間（目安）

■ 特定場面法

個人密着法では収録の難しい場面

- ✓ 職場での会議・会合
- ✓ 店舗での接客場面、など

NINJAL
National Institute for Japanese Language and Linguistics

収録協力者への依頼事項

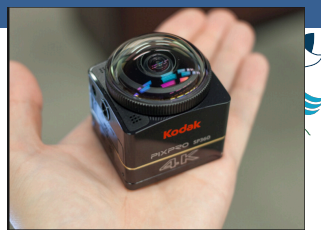
実際の収録に加え、以下についても収録協力者に依頼
（研究者は一切介在せず）

① 会話者に収録の主旨説明
② データ収録・公開に関する承諾書の依頼
③ 会話の属性（会話の日時や参加者、使用機材、配置など）の記録
④ 会話者の属性（性別・年代・職業・出身地など）の収集

⇒各種個人情報を扱うなど重い責任を伴うため協力者は成人に限定

※参考：一連の調査協力に対し謝礼は12万円
（調査終了時にヒアリングも実施）

収録に用いる機材の例



Kodak PIXPRO SP360 4K

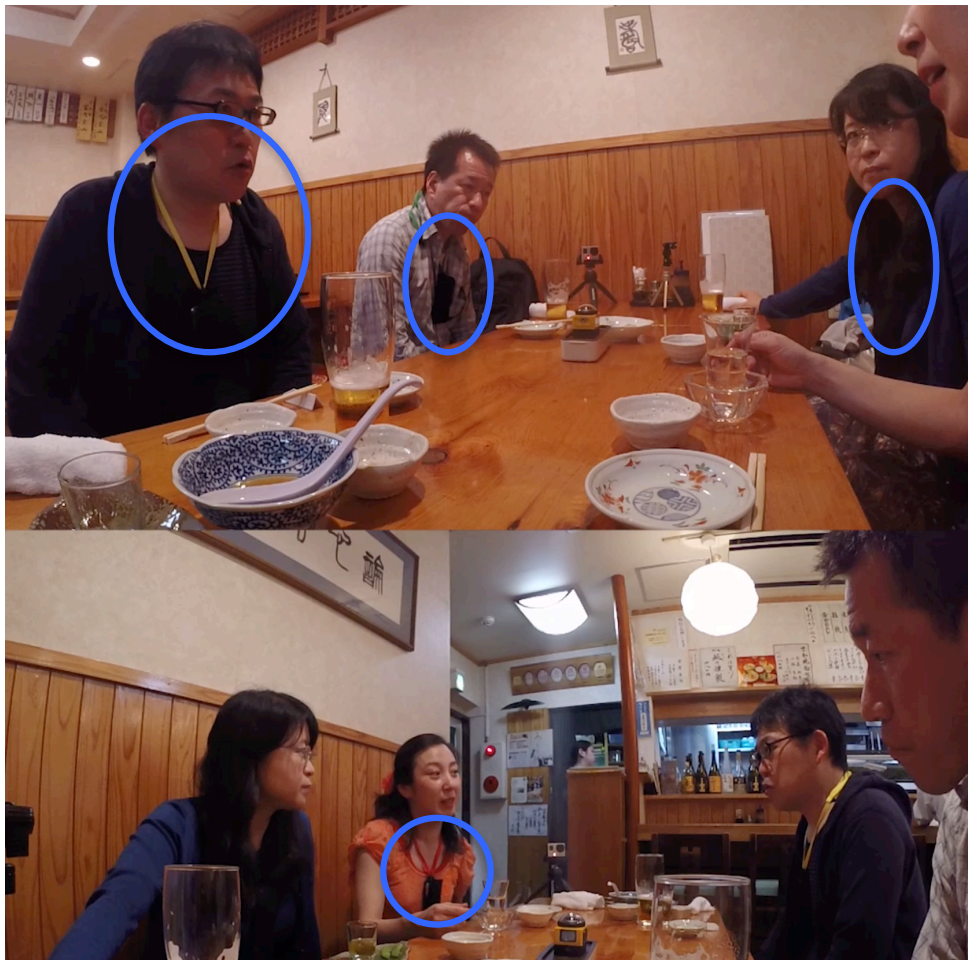
会話者の中心に360度撮影可能なカメラを配置

GoPro Hero3+

会話を俯瞰的に記録するカメラを1～2台配置

音声データの収録機器：屋内の場合

- 会話者ごとにICレコーダーを首から下げたフォルダーに入れて装着



Sony ICD-SX734

あご下
約15cm

NINJAL
National Institute for Japanese Language and Linguistics

④映像データの公開に向けた法的・倫理的問題

会話収録・公開に関する同意の手続き

協力者が実施（研究者は一切介在せず）

- 会話収録の前に行うこと
  - 調査の趣旨・公開方法の説明
  - 会話収録・公開に関する同意書への署名

  ※全ての会話者の同意が得られた場合のみ収録
- 会話収録の後に行うこと
  - 会話中で公開を望まない箇所の聴取
  - 同意撤回書（オプトアウト）の説明

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。

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(かめい)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。

NINJAL
National Institute for Japanese Language and Linguistics

同意書に記したデータ公開方法

◇ 映像

研究教育利用・商業利用（統計情報利用）の条件に同意し
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。

今回の議論
の対象

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。
会話全体ではなく短いシーンごとの公開に留めます。

同意書に記したデータ公開方法

◇音声・テキスト

個人情報（お名前・所属組織名・住所・電話番号）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
- 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	公開用に加工したテキストの例
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	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

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・車ナンバー）
  - 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）

著作物等の写り込み

著作物の「写り込み」等に係る規定（30条の2）

（付随対象著作物の利用）

第三十条の二 写真の撮影、録音又は録画（以下この項において「写真の撮影等」という。）の方法によつて著作物を創作するに当たつて、当該著作物（以下この条において「写真等著作物」という。）に係る写真の撮影等の対象とする事物又は音から分離することが困難であるため付随して対象となる事物又は音に係る他の著作物（当該写真等著作物における軽微な構成部分となるものに限る。以下この条において「付随対象著作物」という。）は、当該創作に伴つて複製又は翻案することができる。ただし、当該付随対象著作物の種類及び用途並びに当該複製又は翻案の態様に照らし著作権者の利益を不当に害することとなる場合は、この限りでない。

２ 前項の規定により複製又は翻案された付随対象著作物は、同項に規定する写真等著作物の利用に伴つて利用することができる。ただし、当該付随対象著作物の種類及び用途並びに当該利用の態様に照らし著作権者の利益を不当に害することとなる場合は、この限りでない。

分離困難性

軽微な構成部分
※1～2割程度（あくまで目安）

著作権者の利益を不当に害しない

NINJAL
National Institute for Japanese Language and Linguistics

公表著作物の対応

- 写り込みの範囲と考えられるものは、ボカシなどの処理はしない

  ⇒ 大半が写り込みの範囲と判断（知財を専門とする弁護士に相談）

- 写り込みの範囲以上と考えられるものは、ボカシ処理あるいはデータの対象外とする

  ⇒ 本の読み聞かせ、など

肖像権・個人情報などが関わる問題

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（特定個人を識別できる情報、氏名・住所・経歴・画像・音声など）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

肖像利用の受忍限度

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)

肖像権や個人情報などに関わるもの

ポイントの整理

- 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)

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）」

  → 同意書において個人情報の範囲を「会話者の名前，所属組織名，自宅・所属組織の住所・電話番号」と特定し、かつ、「それが分からないようにデータを加工」するとしている

- 個人情報取得時における利用目的・公開の有無の伝達義務

  → 同意書に「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」と明記

肖像権や個人情報などに関わるものの対応

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

映像データ公開の問題 まとめ

- 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
- 同意書を交わしていない第3者や著作物の写り込みについては、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

- 対応の方針を定め、ガイドラインとしてまとめて公開

第3者の音声の写り込み

- 対象の会話に一時的に参加する場合

  例） 店員の注文行動

  知人が一時的に会話に参加

- 対象の会話とは独立に生じる音声がレコーダーに写り込む場合

  例） 隣の席の客の声など

問題となりうる行為・話題

- 行為の例
  - 車のスピード違反、シートベルト未装着
- 話題の例
  - 未成年の飲酒に関する話題
  - 部屋の又貸しに関する話題

音声の写り込みの対応　①

収録対象の会話に一時的に参加する場合：

- 公的な場での社会的な行動とみなせる場合（店員とのやりとりなど），
  あるいは，一般の人で挨拶程度の軽い会話の場合
  → 当該の会話部分も転記する。ただし第3者の顔はボカす。

- 一般の人で踏み込んだ内容の場合
  → 後日，データ公開に関する同意をとる。
  同意がとれない場合はその部分を公開の対象外とする。

音声の写り込みの対応　②

対象の会話とは独立に行われる音声が写り込む場合：

- 発話内容が明瞭に聞きとれ，かつ，私的な内容の場合
  → 当該の音声データ範囲を公開対象外とする

- 上記以外（不明瞭，あるいは，私的な内容ではない）
  → 公開対象とする

その他の問題

その他の問題となりうる行為・言動

公開の妥当性や当事者に与える損害の大小などを考え個別に判断

例) 70代男性による未成年の飲酒喫煙話

- 「自分（70代男性）は高校生の頃から飲酒や喫煙をしていた」
- 「それで高校生の孫にも飲ませてみた」

⑤ アノテーション（軽く）

コーパスの構成

収録　600時間～

※ 個人密着法：調査協力者4–50人 × 平均15時間
特定場面法：未定（上記収録状況を見て検討）

本公開対象　200時間

※ 個人密着法：4–50人 × 4–5時間
特定場面法：未定

［人手作成］　転記テキスト・発話単位
［自動付与］　形態論情報（短単位・長単位）・文節・係り受け

モニター公開対象　50時間

※ 個人密着法：20人 × 2.5時間

コア 20時間

［人手修正］ 形態論情報（短単位・長単位）・文節・係り受け
［人手付与］ 談話行為・韻律情報

アノテーション

アノテーション	概要	全体 200H	コア 20H
転記テキスト	CSJ・千葉大3人会話コーパスの基準を参考に整備	人手	人手
発話単位情報	統語的・談話的・相互行為的な観点から定義した『長い発話単位』に準拠	人手	人手
形態論情報（短単位・長単位）	BCCWJの単語・品詞設計に準ずる。日常会話に頻出する口語表現の扱い要検討。全体自動、コアは人手修正	自動	人手
係り受け情報	発話単位を範囲に文節間の係り受け情報を付与	自動	人手
談話行為情報	国際標準化規格 ISO24617-2 を基づき日常会話用に整備した基準に基づき付与	ー	人手
韻律情報	CSJ構築時に整備したアノテーションスキーム X-JToBI を簡略化した体系に準拠して付与	ー	人手

コーパスの規模

時間（目標値）	200時間
語数（推定値）	200万語
会話数（推定値）	400会話
延べ話者数（推定値）	1200人
異なり話者数（推定値）	600人

1万語／1時間

※独話中心のCSJ
1.1万語/1時間

公開スケジュール（宣伝）

2018年度（秋頃）　50時間　モニター公開

2021年度末　　200時間　本公開

※2018,3,19 シンポジウム「日常会話コーパス」（国語研）にて
　モニター公開データのデモンストレーションを実施予定