国立国語研究所 異分野融合型共同研究
2024年度 ワークショップ　(2025/02/27)

東北方言昔話のオープンコーパス化

丹治 尚子 (東京大学)、庄司 潤子 (仙台文学館)、 佐藤 照一 (昔話採訪家)、
高道 慎之介 (東京大学/慶應義塾大学)

Takamichi Laboratory
慶應義塾大学 高道研究室

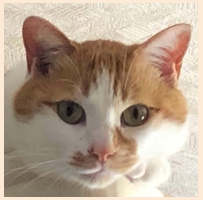
プロジェクトメンバ

目標1：機械学習G　**目標2：コーパス開発G**

研究参加者
(コアメンバ)

佐伯 高明
(東大)

高道 慎之介
(東大)

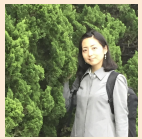


丹治 尚子
(東大)

庄司 潤子
(仙台文学館)

研究協力者

金森・
兵藤・
中田
(東大)



森松
(東大)

佐藤
(昔話採訪家)

田村
(方言ｱﾄﾞﾊﾞｲｻﾞｰ)

佐々木
(仙台)

小幡
(東大)

目次

東北方言コーパスの設計

1.音源
2.コーパス概要
3.作業上の課題
    a.内容確認
    b.書き起こし（基本形）
    c.共通語対訳
    d.話型分類データ
4.今後の目標

音源：

佐々木徳夫

言語を追加

佐々木 徳夫（ささき とくお、1929年（昭和4年）3月19日 - 2010年（平成22年）1月25日）は、日本の民俗学者、昔話収集家。

来歴 [編集]

宮城県中田町（現登米市）出身。旧制宮城県佐沼高等学校、東洋大学文学部哲学科卒業。

社会科担当の高校教師を務めていた1957年より昔話をカセットテープに録音に収集す活動を開始した[1]。集めた昔話は東北地方を中心に1万話を超え、著書は高校教師時代の1966年（昭和41年）11月の自費出版による『酒の三太郎』が最初で[2]、50冊を超える[3]。1992年、昔話の保護に取り組んだ業績が認められ、吉川英治文化賞を受賞した。

後に資料は **仙台文学館** に寄贈

https://ja.wikipedia.org/wiki/%E4%BD%90%E3%80%85%E6%9C%A8%E5%BE%B3%E5%A4%AB

仙台文学館 (宮城県仙台市青葉区北根2丁目7−1)

https://www.sendai-lit.jp/

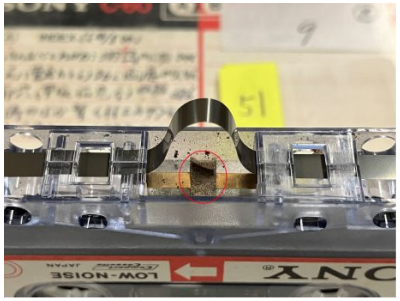
音源：保存音源とディジタル化の進捗

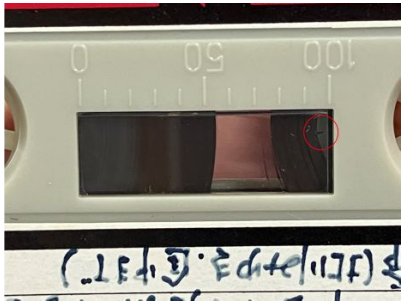
	オープンリール	カセットテープ
本数	88本	171本
ディジタル化	**完了 (2023/02)**	**完了 (2024/01)**
時間数	157時間	283時間
収録時期	1967ー1983	(調査中)
昔話数	2,475	(調査中)
採訪地	58 (区・町単位)	(調査中)
話者数	233 (女: 132, 男: 100)	(調査中)
アノテーション	**一部完了 (後述)**	(検討中)

音源について：テープの問題

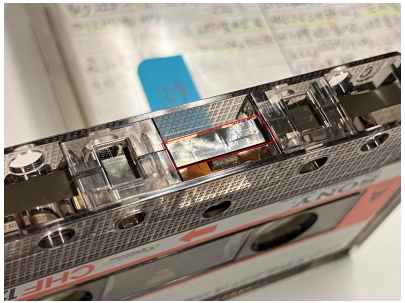
オープンリールのカビ
→ クリーニング作業 (18本)

カセットテープ
パット剥落 (8本), クランプ破損 (12本)





カセットテープ
グリス付着 (1本), カビ発生 (1本)



カセットテープ ケース形状の違い
圧着式の場合はケース破壊が必要

写真提供：(株)プレシード

コーパス概要：話者

話者	性別	生年	住所 (全て宮城県)	話数	時間数	
F001	女	M24 (1981)	仙台市	74	9h15m	★
F002	**女**	**M40 (1907)**	**牡鹿郡女川町**	**23**	**2h25m**	●
F003	**女**	**T13 (1914)**	**遠田郡美里町**	**24**	**1h36m**	●
M001	**男**	**M42 (1909)**	**登米市**	**74**	**9h18m**	★
M002	**男**	**M29 (1896)**	**柴田郡川崎町**	**35**	**2h22m**	●

★：第1弾 (2022) で公開，第2弾 (2025) で追加

●：第2弾 (2025) で公開

太字は著作権者許可済み（享受目的利用可能）

地図：https://www.pref.miyagi.jp/site/access/ken.html

コーパス概要：内容物

- 音声&テキスト
  - 16kHz サンプリング，RIFF WAV 形式
  - 各話書き起こし，共通語対訳(2025年度)

- メタデータ
  - 収録日，各話の開始・終了時刻，
    掲載書籍名, 音声ファイル名,
    話型データ(2025年度)
    - 原題：ラベル記載の題名
    - 改題：書籍掲載の題名

Tohoku folktale corpus (東北地方民話コーパス)

Download / ダウンロード

Click here. [zip 0.5GB]

Restored speech / 復元音声： zip

Description / 内容

This corpus (database) is a digitized and annotated collection of Tohoku region folktales stored on open-reel and analog tapes. These folktales were collected by the folktale collector Tokuo Sasaki (1929-2010), who began collecting them in 1957, mainly in the Tohoku region.
本コーパス（データベース）は，オープンリールテープやアナログテープに保存されていた東北地方民話(昔話)をディジタル化し，アノテーションを付与したものです．この昔話は，昔話採集家の佐々木徳夫（1929-2010)が 1957年より東北地方を中心に収集したものです．

- speech/ ... 16kHz-sampled speech data / 16kHzサンプリングの音声ファイル
- transcript/ ... transcription / 書き起こし
- meta_info/ ... meta information / メタ情報

Terms of use / 使い方

本コーパスは，以下の例外を除き，音声言語の情報解析の用途（商用，非商用を問わない）で利用可能です．

- 公序良俗に反する利用
- 話者，または他者を誹謗中傷したり，信用失墜を意図する内容を含む利用

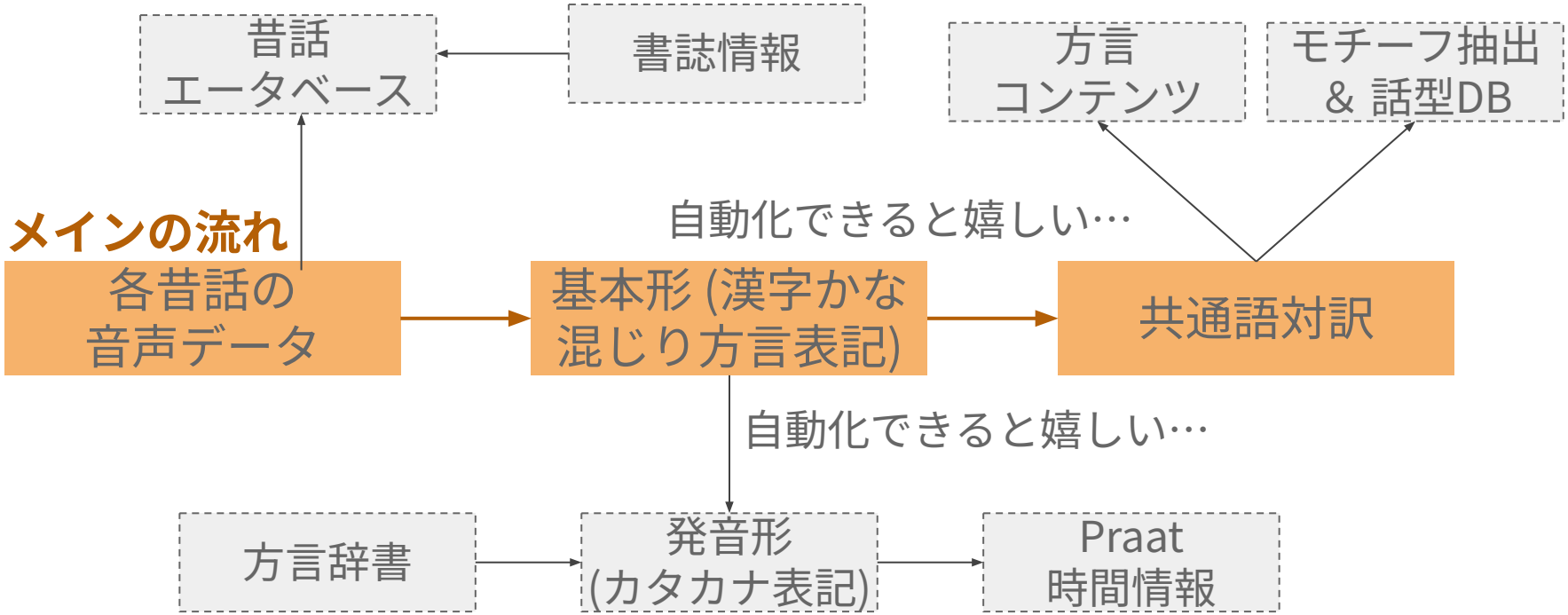
なお，以下の話者については，著作権継承者の許可を頂いたため情報解析以外にも利用可能です．ただし，上記例外の場合には利用できません．

- M001

This corpus may be used for use (including commercial or non-commercial uses) of information analysis of spoken language except in the following cases:

- Any use that violates public order and standards of decency
- Any use that includes contents intended to defame or discredit the speaker or others

https://sites.google.com/site/shinnosuketakamichi/research-topics/tohoku-dialect_corpus?authuser=0


昔話
エータベース
書誌情報
方言
コンテンツ
モチーフ抽出
& 話型DB
メインの流れ
自動化できると嬉しい…
各昔話の
音声データ
基本形 (漢字かな
混じり方言表記)
共通語対訳
自動化できると嬉しい…
方言辞書
発音形
(カタカナ表記)
Praat
時間情報

テープの内容確認

問題点

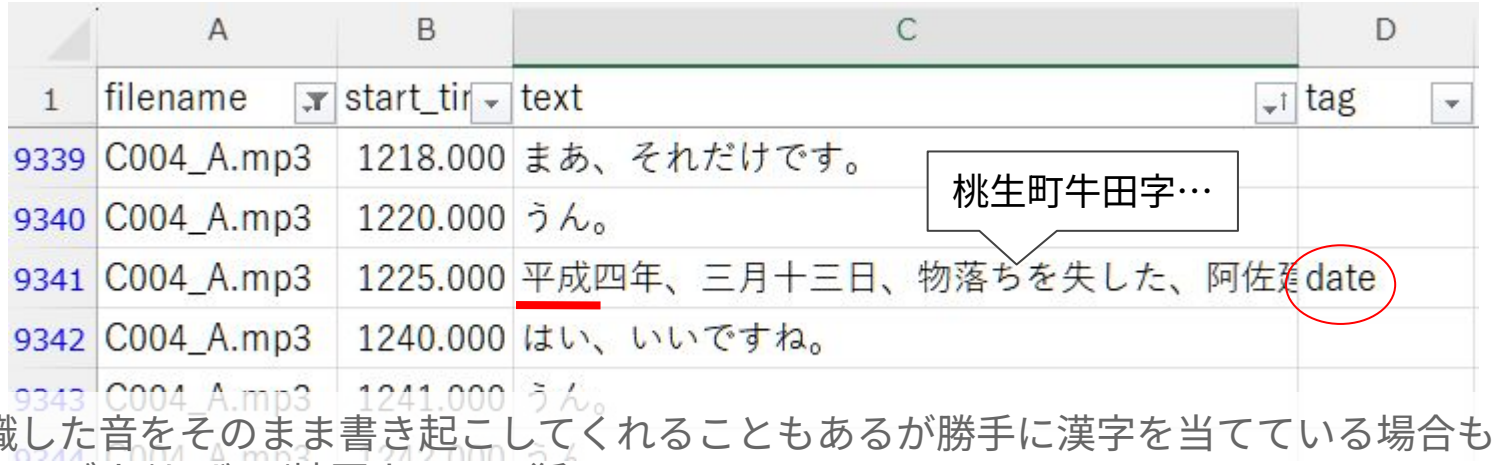
- 1本のテープに複数の話者の語りが収録されているものがある
- ラベルの記載が不十分

対応策

- 音声認識によるテキスト化→分割用タグ付け→メタデータファイル
  - Whisperを使用（デモ：https://huggingface.co/spaces/openai/whisper）
  - 認識結果からタグ付与を一部自動化
  - 音声と同期させ、波形を見ながら話の切れ目を探す→題名タグ付け

テープの内容確認：認識結果例

- date：元号…採録日、住所、氏名、生年月日
- start：語り始めの語…むかし/むがし/むがす/あったずもな　etc.
- end：語り終わりの語…どんどはれ/えんつこもんつこさげだ/なるほどね　etc.



	A	B	C	D
1	filename	start_tir	text	tag
9339	C004_A.mp3	1218.000	まあ、それだけです。	
9340	C004_A.mp3	1220.000	うん。	
9341	C004_A.mp3	1225.000	平成四年、三月十三日、物落ちを失した、阿佐美	date
9342	C004_A.mp3	1240.000	はい、いいですね。	
9343	C004_A.mp3	1241.000	うん。	

認識した音をそのまま書き起こしてくれることもあるが勝手に漢字を当てている場合も…
例：チヅネ/ちずに/地図人　→　狐

書き起こし（基本形）：凡例

- 漢字・平仮名混じり
- フィラーも書き起こす
- 漢字の前には ｜ を入力し、漢字の後の［］に読みを平仮名で入れる
  - 例）｜一軒［いっけん］の、　｜貴様［くさん］
- 漢字が続いている場合は単語ごとに「｜」を入れる。
  - 例）｜大変［たいへん］｜繁盛［はんじょう］
    (四字熟語の場合は ｜四字熟語［よじじゅくご］)
- 聞き取り困難な部分は音の数だけ ＊を入れる。
  - 例）｜食［た］べ＊の
- 表記に迷う場合は （｜）(かっこ内に ｜ で区切って入れる)
  - 例）客：「きゃく」か「きゃぐ」か迷う→｜客［きゃ（く｜ぐ）］
- 語中の濁音の前鼻音は全角の「ン」を入れる。
  - 例）やンど（宿）　まンど（窓）
- 歌は始めと終わりに ෴ を入力する
- 話と無関係な発話は（）に入れる（聞き手の相づち、聞き手への確認など）
- かっこ、｜、＊などの記号はすべて全角にする
- 息継ぎの有るところに読点，文の切れ目と思われるところに句点を入れる
- 母音が続く場合も長音符号を使わず書き起こす。文末・語末で母音の引き延ばしの場合は長音符号を使う
  - 例）｜大［おお］きな　それー　なあー

書き起こし（基本形）：問題点

- 鼻濁音が反映されていない
  - 第2弾で話者F001の鼻濁音表記を追加

- 判断に迷う音の表記をどちらかに統一すべきか
  - [し]？or[す]？→［し］と［す］は区別がない→[す]とする
  - [ひ]？or[し]？→近い音になっているだけ→［ひ｜し］とする

- 発音形(カタカナ表記)の作成に向けて
  - COJADSの形式に合わせるにはタグが不足している
  - 分かち書きを自動化できるか ー 方言辞書が必要
    - 語彙集を電子ファイル化

共通語対訳：

- 基本形から共通語対訳を作成する
  - 国会図書館デジタルコレクション（2022.10〜）を活用

- 道具・食物・地名など、現代にないものの聞き取り・書き起こし・対訳が難しい→注釈が必要
  - 参考文献：リストを作成
  - 本文中の注釈：必要最小限に止める
    - 読める/読めない、わかる/わからないに個人差がある→どこまで付与するか
    - 機械学習での利用には注釈や読み仮名はない方がよい⇔対訳を読んで理解するためには注釈や読み仮名は必要
    - 別途用語集を作成する

共通語対訳：凡例

- できるかぎり意訳はしない。（意訳タグが必要か）
- 助詞がないと理解しにくい場合は適宜補い、[ ]　に入れる。
- 意味が通りにくい場合は、語句を補い、[ ]に入れる。
- 注釈が必要な場合は［＝注釈］とする。
  - 例：唐櫃［＝からど。足のついた長持。］
- フィラーは、音の数だけ×を入れる（言葉にならないもののみ。「あの」「その」とはっきり聞こえるものは書き出す）。
- 言い間違い、言い淀みなど共通語訳ができない部分は文字数分×を入れる。
- 歌や節回しのある文などは〽の記号を前後につける。漢字を当てることが出来る場合は漢字を当てる。
- どうしても訳せない語はそのままにする。

共通語対訳：問題点

方言？ or 共通語？

- 良く知られている方言/東京方言でも使う語 は、和英辞書に見出し語がなければ共通語形にした
  - 例：めんこい→かわいい　おっかない→恐ろしい/怖い　たまげた→驚いた/びっくりした
- 共通語と同じ形をしているが、使い方が違う語がある？
  - 例：「もはや海にして」…あと少しで海、というところで事切れていた、という描写。現代語の「今となっては。もう/早くも。すでに」とは意味が異なる
  →用例が少なく、訳を決めきれない
- 「はあ」の扱い…文末詞？　副詞？　感動詞？
  - 感動詞と捉えて「ほら」の訳語を当てた

半沢 康, “ 東北地方におけるハーの伝播と変化,” 方言の研究, vol. 4, pp.159-180, 2018

共通語対訳：問題点

- 「なぜこう訳すか」を説明できない場合がある

  例：「助けられらいんちゃ」→「手伝ってもらいましょうよ」と訳した
  （仙）通常は［助けで貰わいんちゃ］と使うが、意味は同じと考える
  →「～らいん」＝「～なさい」　なので、「手伝ってもらいなさいよ」

  ・対面での調査と異なり、発話者に質問できない

  ・同じ用例がないか？⇒話型による分類情報が必要

話型分類

- 昔話の話型分類データを付与
  - メタデータに追加
  - 関敬吾ほか編『日本昔話大成』第11巻の話型分類番号を利用

- 共通語対訳→モチーフ・テーマ抽出→話型分類
  - 同じタイプの話を取り出して比較＝話型別コーパス（2025〜)
  - 地図×話型分類×音声

今後の課題（2025年度目標）

- アノテーションの拡充
  - 第2弾コーパスの公開
  - 話型別コーパスの作成
- メタデータの整備
  - 書誌情報とリンクした佐々木徳夫採集昔話データベース
- 権利処理
  - 話者ご遺族宛て郵便：届けば100%承諾
  - 郵便が届かない場合の対応
- アウトリーチ活動
  - 仙台市内の小学校にて昔話と方言の楽しさを味わうイベントを企画