立命館大学創思館カンファレンスルーム
言語教育情報研究科20周年記念企画
2023/11/26日

大規模言語モデルと日本語の未来ー方言はどうなるか

田窪行則
京都大学、国立国語研究所名誉教授
国立国語研究所客員教授

はじめに

- ChatGPTのような大規模言語モデルが作る世界で多様性が保証されるかどうか
- 危機言語の記録・維持・再活性化の問題は「低資源言語処理」の問題に帰着する。
- データをどうやって保存するか、増やすか：さまざまな困難。
- 地域でのそれぞれの自主的な解決が必要である：市民科学者の育成。
- 低資源言語処理：原理的には可能だけれど、実際には困難

今日の予定

1．ChatGPTなどの大規模言語モデルと日本の社会
2．日本の言語状況：多言語社会としての日本
3．言語か方言か：相互理解性を測る
4．消滅危機言語の記録・保存・再生（再活性化）の試み：
　　デジタル博物館構想
5．記録保存と言語再活性化の諸問題
6．市民科学としての危機言語記録・再生（再活性化）
7．大規模言語モデルと文化多様性
8．おわりに

２．日本の言語状況：
多言語社会としての日本

- 日本は単一言語・単一文化の国として意識されることが多いが、実際には多言語国家である。
- 先住民語としてのアイヌ語、移民言語（継承言語）としての朝鮮語（韓国語）、ブラジルポルトガル語、などがある。
- 琉球諸語、八丈語などは、相互理解性の観点から「言語」として扱われる場合がある。
- これらはすべて「消滅危機にある」である。
- どれもデータが非常に限定されている。

２．日本の言語状況
消滅危機度：世代間継承指標

- 1.大丈夫（safe）：すべての世代がその言語を話し、次世代への継承が行なわれている。
- 2.危ないかも（vulnerable）：こどもの大部分がその言語を話すが、使用領域が限定されている（家庭など）。
- 3.確実に危ない（definitely endangered）：こどもが家庭でもう母語としてその言語を習得しない
- 4.非常に危ない（severely endangered）：祖父母やお年寄りはその言語を話しているが、父母の世代は理解はできても子供たちに話さず、自分たちの間でも話さない。
- 5.瀕死状態（critically endangered)：一番若い話者が祖父母世代で、しかも使用は部分的で、頻度も多くない。
- 6.消滅(extinct)：　話し手が残っていない

UNESCO Intangible Cultural Heritage - *Endangered languages*

２．日本の言語状況
日本で話されている「言語」

- ユネスコの認定した日本の「言語」の危機度

1.大丈夫:日本語
3.確実に危ない：沖縄語、国頭語、奄美語、宮古語、八丈島語
4.非常に危ない：八重山語、与那国語
5.瀕死：アイヌ語、

- アイヌ語以外は同源であることが証明されている。「日本語」以外はすべて消滅の危機に瀕している。

- 実際には「日本語」も共通語と「関西弁」以外は危ない。実は「関西弁」も危ないかも
- 最近のUNESCOの調査：https://en.wal.unesco.org/countries/japan

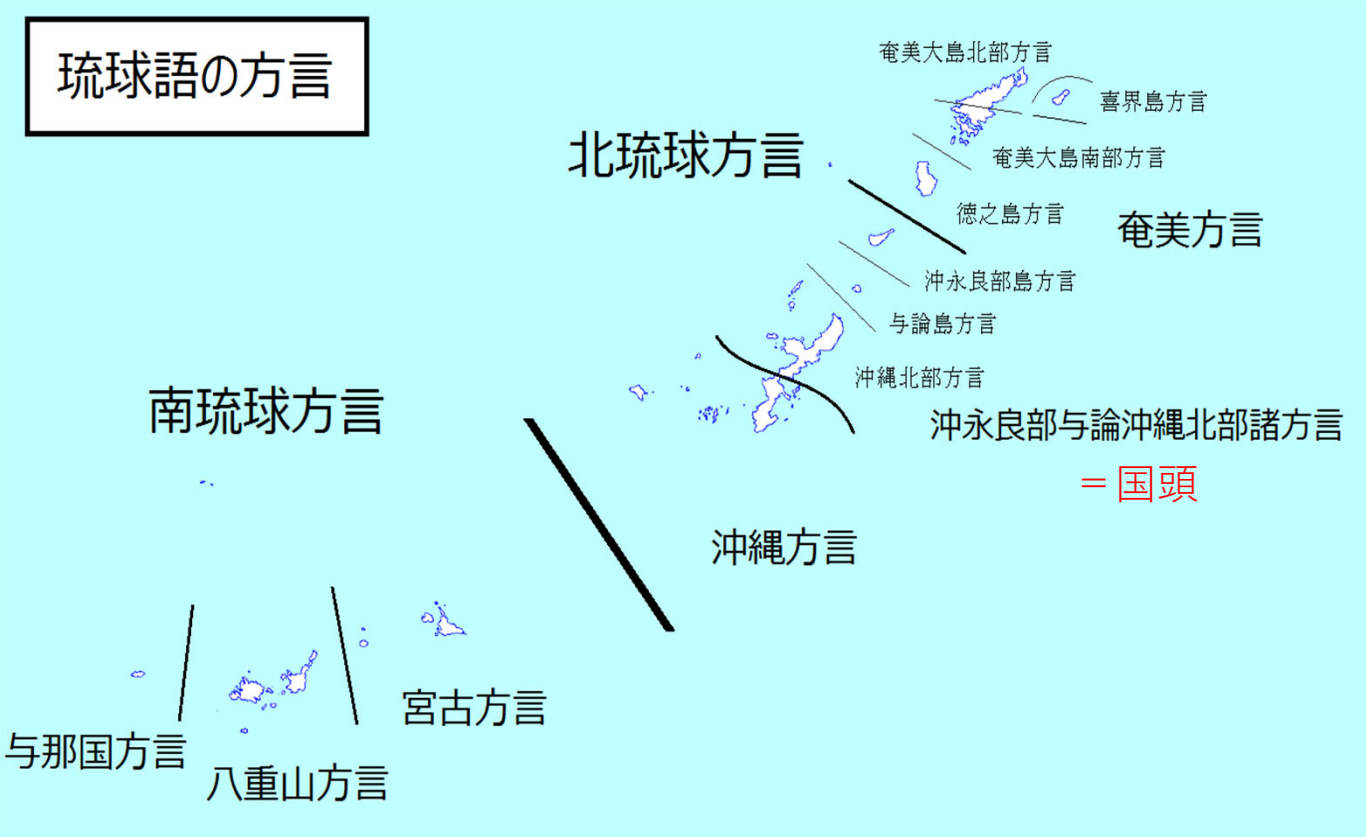
２．方言か言語か：相互理解度を測る

- 相互理解できるのが方言、できないのが言語

　Cf.　Charles Hockett (1958) An idiolect, dialect and language: Chapter 38: *A course in modern linguistics.* The Macmillan Company: New York.

➢Wikipediaの記事（mutual intelligibility)のなかに相互理解できないがしばしば方言とされるもの（ List of mutually unintelligible varieties sometimes considered dialects）として「日本語」に対する「琉球諸語」があがっていた（現在は削除）。

Wikipedia：琉球語の方言



この記事は琉球語を認めて、奄美、国頭、沖縄、
宮古、八重山、与那国は方言とみなす。

３．方言か言語か
言語と方言の区別

- **政治的定義**

「言語とは陸軍と海軍を持つ方言である」

言語学者Max Weinreichが講演中に聴衆の一人が言った言葉とされる。（Wikipedia: A language is a dialect with an army and navy）。

- **相互理解度による定義**（言語学的定義）

お互いに理解できれば方言、理解できなければ方言

（言語学者チャールズ・ホケット Hockett 1958, *A Course in Modern Linguistics*, chapter 38 ）

３．方言か言語か：相互理解度を測る
相互理解度テスト

- UNESCOは相互理解度を測ったわけではない。
- 本当に相互理解が可能かどうかを測ってみる必要がある。

３．方言か言語か：相互理解度を測る

相互理解度テスト作成の試み

言語・方言（地域変種）間の相互理解度がどの程度あるかを測る客観テスト

特徴：

１.方言（地域変種）間の距離を相互理解度で測る。

２.比較言語学による系統関係とは関係ない。

３.借用が多ければ系統が違っていても理解できる部分が増える。

４.各方言が方言でなく別言語とみなすべきかの指標として使える。

５.言語の継承度(=消滅危機度)を測ることができる。

３．方言か言語か：相互理解度を測る
相互理解度テスト　先行研究

ハワイ大O’Grady,W.教授 他：韓国語済州島変種とソウル変種の相互理解度テスト

“済州語は韓国の済州島で昔から話されている言語であるが、しばしば韓国語の「地方なまり」「方言」とされてきた(e.g., King1996, Sohn 1999:74, Yeon 2010). 私たちはこの分類は間違っており、済州語は、独立した言語であり、韓国語族の姉妹語であるとすべきであることを示す (O’Grady 2014　田窪訳)”

方法：済州語と韓国語ソウル方言、プサン方言とは相互理解性がないことを示す。

O’Gradyらの実験　(O’Grady 2014).

被験者は簡単な話を済州語で聞いて10の簡単な質問に答える。

済州語母語話者	ソウル	全羅南道	慶尚南道
89.21	12.03	6.00	5.26

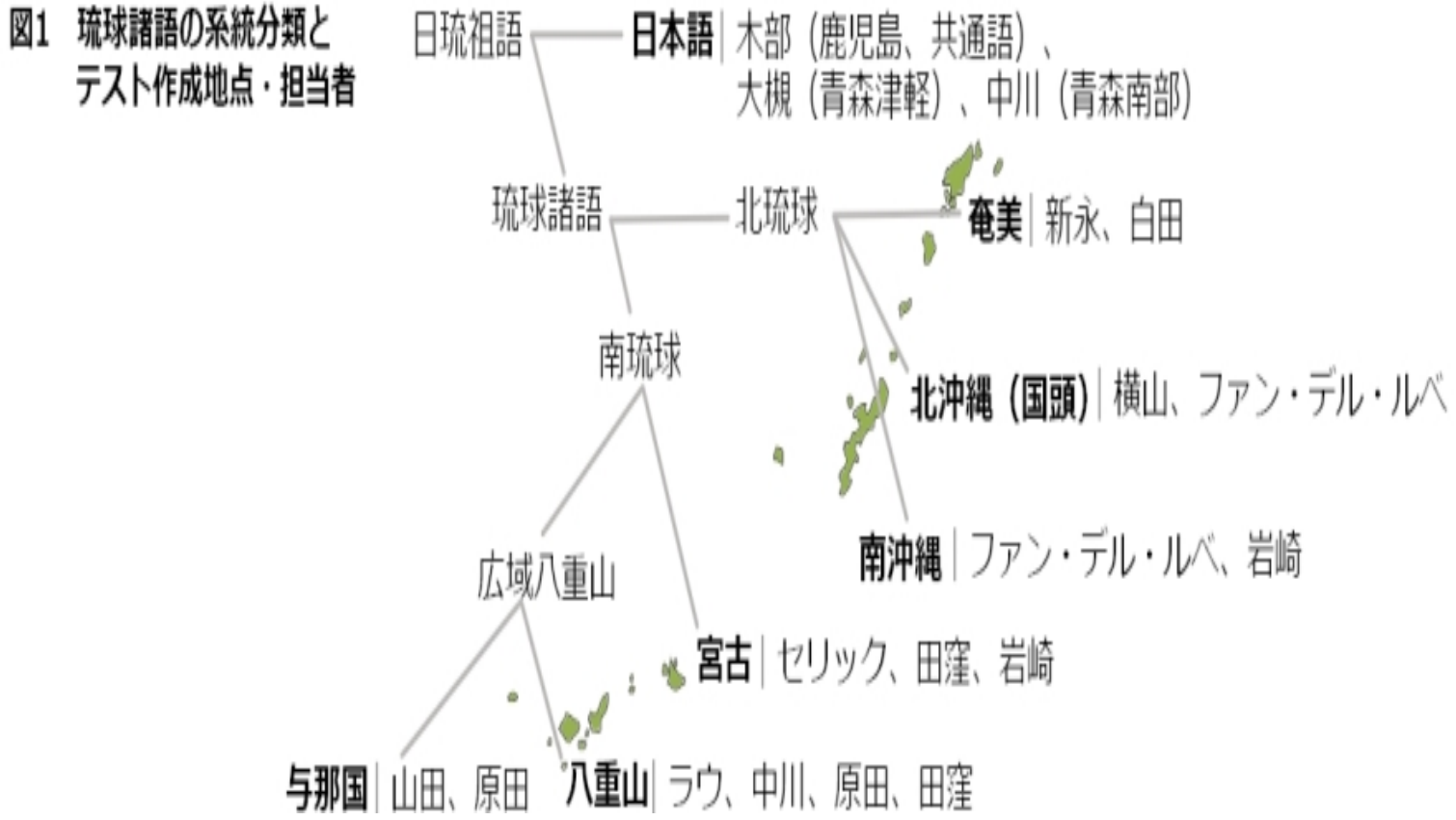
３．方言か言語か：相互理解度を測る
我々の方法

- 彼らのテストを改良してより客観性を高めたものを琉球諸語、本土方言でも作成中（喜界島、奄美、沖永良部、沖縄北部、宮古（4カ所）、八重山（3カ所）、与那国、青森（2か所）、鹿児島）

３．方言か言語か：相互理解度を測る
我々の相互理解度テスト

- トヨタ財団研究助成プログラム：多文化・多言語社会としての日本の理解－消滅危機言語の相互理解性と世代間継承度のための客観的尺度の創出
- 代表　山田真寛（国語研）
- テスト作成　山田、里麻奈美（沖縄国際大学）、田窪

相互理解度テスト担当者



池間方言（西原地区）

- 例1 

質問：登場人物は何をしていますか。

- 例2 

質問：母さんは何をしていますか。

以下20問まで

沖永良部（上平川方言）

- 例1

  質問：登場人物は何をしていますか。

  例2

  質問：母さんは何をしていますか。

  以下20問まで

相互理解度テスト

- 簡単なストーリー
- 文化的な背景が中立的なものにする
（借用語が多すぎず、内容が簡単に想像できないようにするため）
- 内容に関する簡単な質問をする。
- 客観性を保証するために、他の言語に関しても同じ形式で作成する。
- PinguとPear Storyを使用。

西原で行ったテストの結果（暫定版）

- **話者が西原方言（＝母語）を聞いた場合**

１．65歳以上（N=10）　平均年齢71

19／20点　95%

2.　50代前半（N=5）　平均年齢51

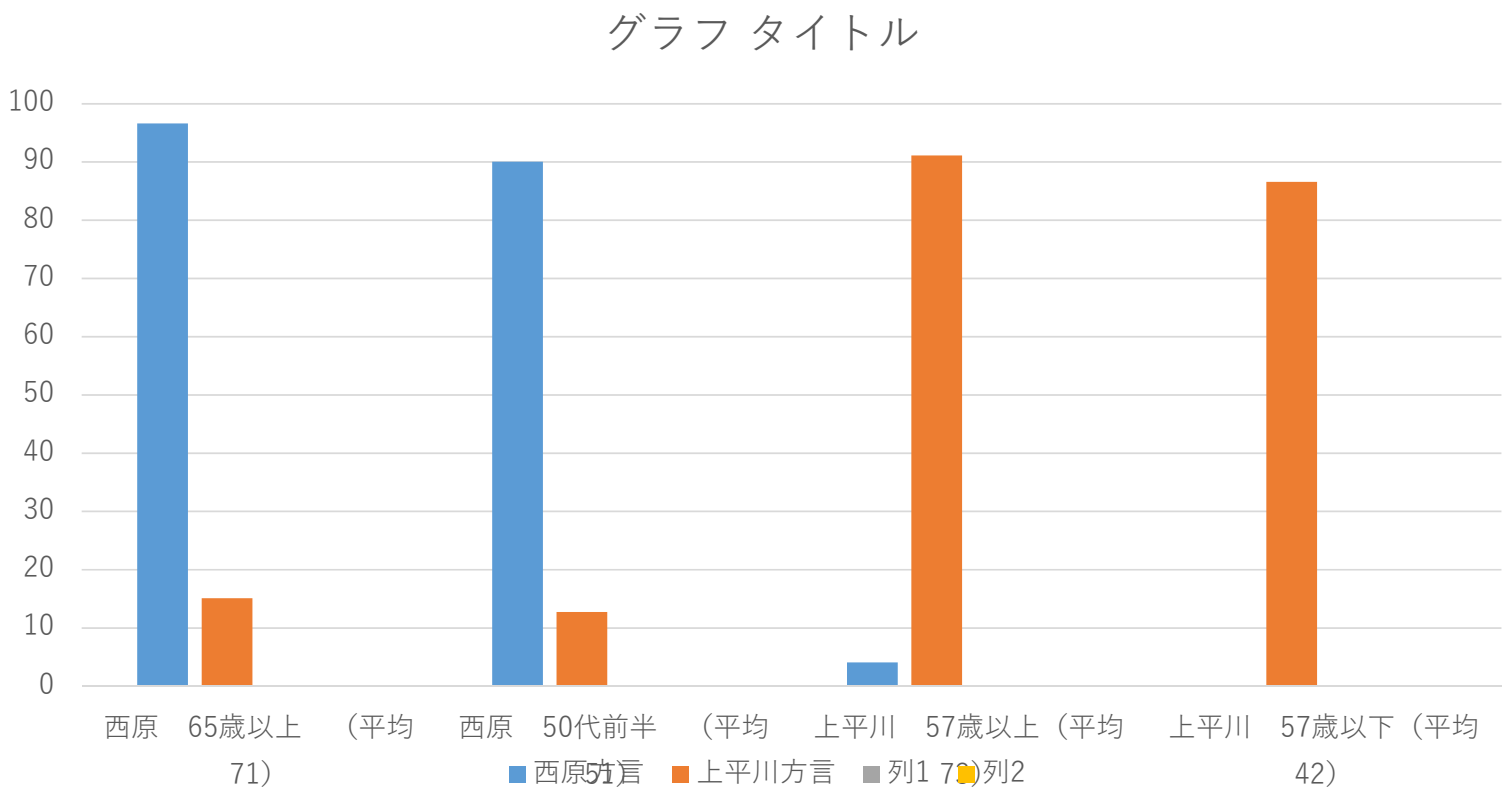
18／20点　90.075%

- **上平川話者が沖永良部方言（＝母語）を聞いた場合**

1. 65歳以上 (N=10)　3／20点≒15%

2. 50代前半（N=5）　2.6／20点≒13%

百分比（%）による理解度


グラフ タイトル
100
90
80
70
60
50
40
30
20
10
0
西原　65歳以上　（平均　71）
西原　50代前半　（平均　51）
上平川　57歳以上（平均　72）
上平川　57歳以下（平均　42）
西原方言
上平川方言
列1
列2

沖永良部で行ったテストの結果（暫定版）

- **上平川方言（母語）聞いた場合**

57歳以上　（N＝１１）　平均年齢73歳

18／20点＝90%

57歳以下-36歳以上（N＝１２）　平均年齢42歳

17/20点≒85%

- **池間方言を聞いた場合**

57歳以上　（N＝１１）

0.8/20点＝4%

沖縄県の北と南の二つの「方言」

- 沖永良部→池間：

  全く通じない

- 池間→沖永良部：

  ほとんど通じない！

（10%程度だと、木とか自転車という単語を聞いて想像できる人ならとれる。）

なんで池間の話者のほうがすこし沖永良部のことばを聞けるのか：

沖縄に住んだことがあるひとが何人かいた。

- 琉球の言葉はすべて本土の人から見れば別の言語とみるべき
- 琉球の中でも南と北では全く通じない：別の言語とすべき
- 南琉球の諸言語、北琉球の諸言語でも通じない可能性は高い：
- 青森の言葉も東京の人には通じない可能性大：別の言語か

相互理解度は連続的

- 二つの言葉の差で測る。

　A：　{沖永良部と沖縄（うちなーぐち）} ＜ {池間と沖永良部}

　B:　{池間と宮良} ＜ {池間と沖永良部}

　C：？ {津軽と東京} ＜ {池間と宮良}

これらの差は連続的で相互理解度が何パーセントから別の言語かは恣意的にしか決められない。

相互理解性テストの問題点と改善策

- 琉球諸語の相互理解性テスト

Yamada, M. et al. 2020. Experimental Study of Inter-language and Inter-generational Intelligibility: Methodology and Case Studies of Ryukyuan Languages. In: *Japanese/ Korean Linguistics. 26,* 249-260.

- 問題点：心理言語学的厳密差を保証するために時間と手間がかかりすぎ。調査者・非調査者ともに疲労困憊。

O’ Grady氏たちの改善策：穴埋めテストを使う。

Sejung Yang, et al. The Status of Jejueo: Endangered Language or Disappearing Dialect? (Draft, University of Jeju)

以下の論文に基づいてCloze testを行う。

- Charlotte Gooskens, Vincent J. van Heuven, Measuring cross-linguistic intelligibility in the Germanic, Romance and Slavic language groups, *Speech Communication, Volume 89*, 2017, Pages 25-36.
- Charlotte Gooskens, Vincent J. van Heuven, Jelena Golubović, Anja Schüppert, Femke Swarte & Stefanie Voigt (2018) Mutual intelligibility between closely related languages in Europe, *International Journal of Multilingualism*, 15:2, 169-193, DOI: 10.1080/14790718.2017.1350185

The test used by Gooskens and her colleagues consisted of four oral passages, versions of which were prepared for the various languages in the study. The passages, each approximately **200 words in length**, were at the B1 level of difficulty (roughly "intermediate") in the Common European Framework of Reference for Languages prepared by the Council of Europe (https://www.coe.int/en/web/common-europeanframework-reference-languages).

Each narrative contained **12 gaps**, signaled by a beep of one second, with 30 ms of silence on each side. At each beep, participants had to select **a word to fill the gap from among 12 items in a written list** that appeared at the top of their computer screen, arranged 4 in three columns of four words each. Here is an example test for English. (The narrative itself was presented orally and did not appear on the screen.)

Word list from which participants choose items to fill gaps in the narrative:

heart good stop
exercise dangerous carrying
muscles easier help
advantage true start

The oral narrative:

Getting enough exercise is part of a healthy lifestyle. Along with jogging and swimming, riding a bike is one of the best all-round forms of exercise. It can __ to increase your strength and energy. Also it gives you more efficient muscles and a stronger __ . But increasing your strength is not the only __ of riding a bike. You're not __ the weight of your body on your feet. That's why riding a bike is a __ form of exercise for people with painful feet or backs. However, with all forms of exercise it's important to __ slowly and build up gently. Doing too much too quickly can damage __ that aren't used to working. If you have any doubts about taking up riding a bike for health reasons, talk to your doctor. Ideally you should be riding a bike at least two or three times a week. For the exercise to be doing you good, you should get a little out of breath. Don't worry if you begin to lose your breath, it could be __. This is simply not __. Shortness of breath shows that the __ is having the right effect. However, if you find you are in pain then you should __ and take a rest. After a while it will get __ .

相互理解性テストを継承度のスケールに使う

- 消滅危機言語の危機度の算定

  消滅危機言語の危機度＝世代間継承度

- 相互理解性テストを世代間継承度を測るのに使う。
- 継承度が低くなければ、上げることが可能。

消滅危機度：世代間継承指標

- 1.大丈夫（safe）：すべての世代がその言語を話し、次世代への継承が行なわれている。＞80点以上
- 2.危ないかも（vulnerable）：こどもの大部分がその言語を話すが、使用領域が限定されている（家庭など）。＞65点以上80点以下
- 3.確実に危ない（definitely endangered）：こどもが家庭でもう母語としてその言語を習得しない　＞50点以上65点以下
- 4.非常に危ない（severely endangered）：祖父母やお年寄りはその言語を話しているが、父母の世代は理解はできても子供たちに話さず、自分たちの間でも話さない。＞20点以上35点以下
- 5.瀕死状態（critically endangered)：一番若い話者が祖父母世代で、しかも使用は部分的で、頻度も多くない。　＞10点以上20点以上
- 6.消滅(extinct)：　話し手が残っていない　＞10点以下

言語再活性化への道が開ける

- ３５歳でもかなりわかる人がいる！
- 65%以上の理解度があれば、たくさん聞けばすぐに90%以上の理解度を達成できる。
- たくさんコンテンツを作って聞いてもらえば危機言語の再活性化への道が開ける！

４．消滅危機言語の記録・保存・再生の試み：デジタル博物館構想

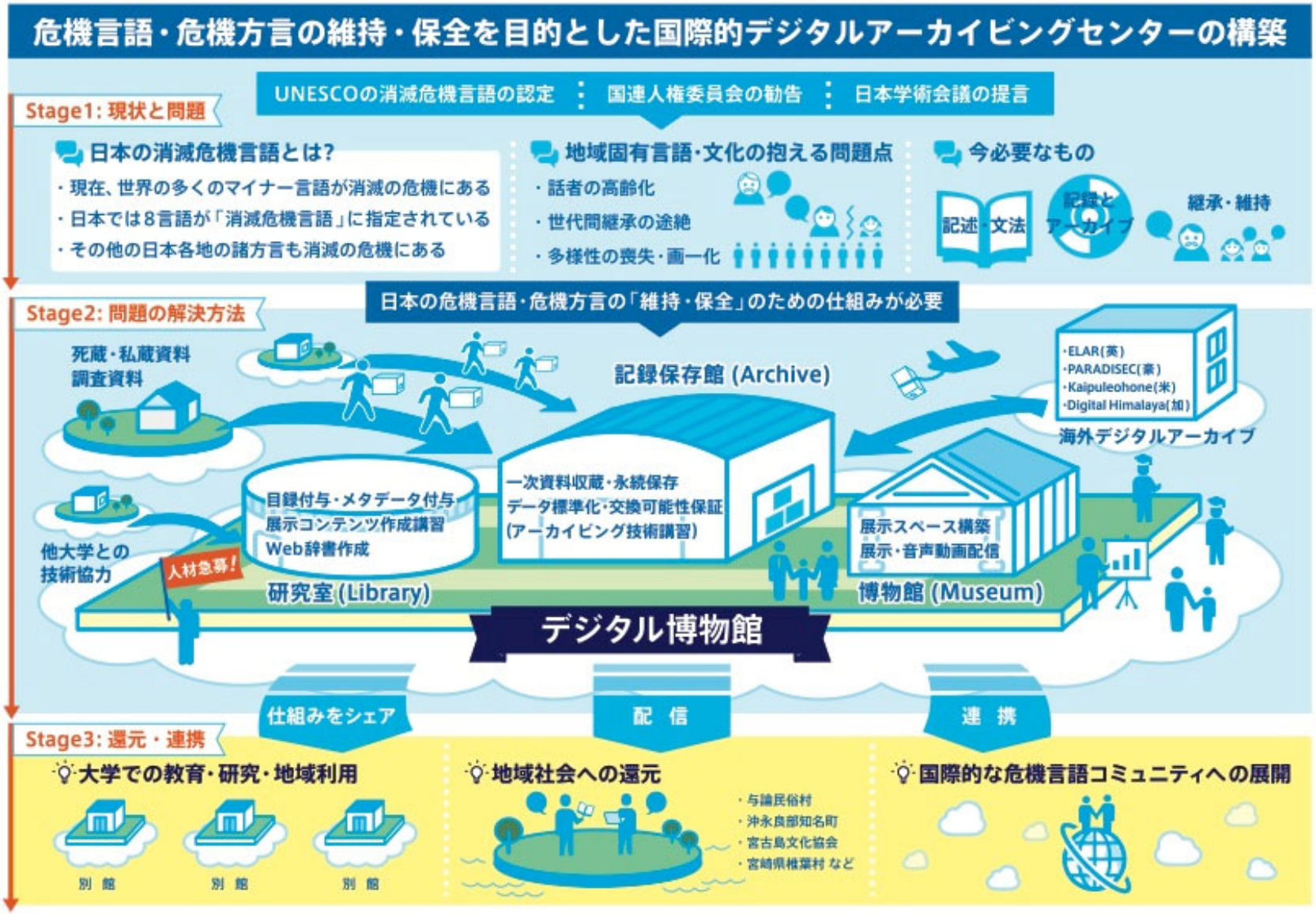
- 2008年ごろから構想

The Digital Museum project for the documentation of Ikema Ryukyuan. Takubo. Y et al. The 1st ICLDC. The University of Hawaii, Manoa. March 13th、2009

- 2013年プロトタイプを部分的に公開：デジタル博物館

Constructing a digital museum with a large-scale archive for endangered languages. Takubo, Y. et al. 4th ICLDC, Honolulu Hawaii, February 25-March 1, 2015.

- 2018年概算要求（次ページ）
- 2021年ことばのミュージアム


危機言語・危機方言の維持・保全を目的とした国際的デジタルアーカイビングセンターの構築
UNESCOの消滅危機言語の認定
国連人権委員会の勧告
日本学術会議の提言
Stage1: 現状と問題
日本の消滅危機言語とは?
・現在、世界の多くのマイナー言語が消滅の危機にある
・日本では8言語が「消滅危機言語」に指定されている
・その他の日本各地の諸方言も消滅の危機にある
地域固有言語・文化の抱える問題点
・話者の高齢化
・世代間継承の途絶
・多様性の喪失・画一化
今必要なもの
記述・文法
記録とアーカイブ
継承・維持
Stage2: 問題の解決方法
日本の危機言語・危機方言の「維持・保全」のための仕組みが必要
死蔵・私蔵資料
調査資料
記録保存館 (Archive)
・ELAR(英)
・PARADISEC(豪)
・Kaipuleohone(米)
・Digital Himalaya(加)
海外デジタルアーカイブ
他大学との
技術協力
人材急募!
目録付与・メタデータ付与
展示コンテンツ作成講習
Web辞書作成
研究室 (Library)
一次資料収蔵・永続保存
データ標準化・交換可能性保証
(アーカイビング技術講習)
展示スペース構築
展示・音声動画配信
博物館 (Museum)
デジタル博物館
仕組みをシェア
配 信
連 携
Stage3: 還元・連携
大学での教育・研究・地域利用
別 館
別 館
別 館
地域社会への還元
・与論民俗村
・沖永良部知名町
・宮古島文化協会
・宮崎県椎葉村 など
国際的な危機言語コミュニティへの展開

４．消滅危機言語の記録・保存・再活性化の試み：
デジタル博物館による記録保存、継承保存、再活性化

- 物⇔　技術⇔　言語

<----------:------->

Tangible　　intangible

物：実物、レプリカ、三次元写真
技術（ものづくり、踊り、歌、など）：秘伝書、振りの記録、楽譜、動画・音声、
言語：担い手がいなくなれば消える
　　音声・映像＋書き起こし

- **物、技術、言語をセットで記録保存する。**

４．消滅危機言語の記録・保存・再活性化の試み：
動画・音声の記録と保存

- 動画、音声のメタデータ

https://docs.google.com/spreadsheets/d/14poz4dnEnzqf7Omh9XZcXtA2A8nz7J3MyUAoKxtxDRA/edit#gid=0

- 書き起こし、サブタイトル
  interlinear glosses (一部自動化が可能)
  https://www.eva.mpg.de/lingua/resources/glossing-rules.php

これまでは公開は前提としていない。
博士論文のフィールドデータは公開されていない。
オープンサイエンスの検証のためのデータ公開と
再活性化コンテンツのためのデータ公開は異なる。

５．　記録保存と言語再活性化の諸問題

- フィールドワークによる記録保存とは

Boasの三点セット｛文法、辞書、テキスト｝
　最近はこれに「動画、音声」が加わる。

言語再活性化という観点からは3点セットでは足りない
　　教科書、電子辞書、読本、動画コンテンツ

- これまでフィールドノートや生の資料などが公開されることは（ほとんど）なかった。

- オープンデータ、オープンサイエンスの立場から
  記述データの典拠の明示化
  プライバシーの問題　データの利用に関する制約
  アーカイビンクの問題
  メタデータの書き方
- 実際の例

動画・音声データの管理

１．編集なしの原データ（雑音や不要な部分がある）
２．記録・保存のための最小限加工したデータ
３．コンテンツのための編集加工データ（変換が必要）

公開するためには１，３が必要となる。一つの動画に対して複数の動画コンテンツができる。

１を公開のためだけにストリーミングすることができるか：原理的にはできるが困難。

再活性化のためには

- 再活性化のためには動画・音声コンテンツの**公開**が必要となる。

- デジタル化と公開の問題

  所有者、記録者がアナログデータを私蔵して死蔵する：

  　資産として考える

  データ利用の許可だけでなく動画内容のプライバシー

  保護の問題

６．　市民科学としての危機言語再活性化

- 「市民科学者の育成－地域の中に記録者を一」
- 「しまむに（北琉球沖永良部語）の言語復興活動」
- 「沖永良部島の事例報告 －家族を対象にした取り組み－」（プレゼンをリンクしています）

しまむにサロン

自分たちのしまのことばを自分たちの手で次の世代に伝える「市民科学者」の育成

— 知名町中央公民館講座 —

話せる人も、話せるようになりたい人も

しまむにを流暢に話せる人、話す練習をしたい人、話せるようになりたい人、研究している人が参加しています。島生まれ育ちの人も、島外出身の人もいます。

豊かな方言差を持つしまむに

誰かの心と結びついた地域ごとのしまむにを大切に、一人ひとりの「しまぬ むに（故郷の ことば）」を残すために。参加者の字のことば一つひとつが対象です。

文法の解説のために紹介するときはもちろん、ことばの調査をするときは**必ず「どこの字のことばか」と合わせて記録**しています。特定の字のことばを特別扱いしたり、「沖永良部標準語」をつくったりはしません。豊かな方言差を持つ沖永良部語をまるごと次の世代に継承することを目指しています。

講座は知名町と国立国語研究所が締結した**沖永良部語継承保存のための連携協定**にもとづき、2019年度から毎年月1回2時間開講されています。

講座の目標

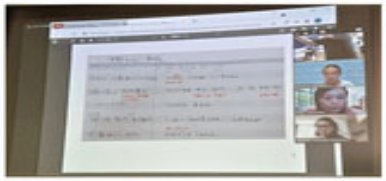
しまむにを教える力・習う力・記録する力を、学習編と実践編をとおして身に着けることを目指しています。実践実習では毎年度、辞書データ（音声付きの語彙と用例）や動詞活用などをテーマに参加者の方言ごとの調査収集を行ってきました。

1．教える力・習う力

「話せること」と「教えられること」は同じではありません。人に教えてもらうときはもちろん、若い人や子供たちにしまむにを教えるときに役立つ、**しまむにのしくみ（文法）**を学習します。

2．記録する力

外部の研究者に任せっきりにせず、一人ひとりが、**じぶんたちの手でじぶんたちのことばを次の世代に伝えるために**必要な、言語を記録する力を身に着けます。



文法の理解も、ことばを記録する力も、しまむにを話せる・教えられるようになるための近道です。

3．消滅危機言語を継承する方法

他の地域で消滅危機言語について研究している人のゲストレクチャーを受けて、えらぶでの言語継承への応用方法を考えていきます。これまでに藤田ラウンド幸世さん（国際基督教大学｜バイリンガル子育て・教育）、半嶺まどかさん（名桜大学｜生活の中でしまのことばを継承する方法）、生島常範さん（喜界島言語文化保存会｜喜界島での取り組み）、冨岡裕さん（神田外国語大学｜タイの少数言語の現状）にお話ししていただきました。

つづきはこちら

音声付き語彙集や**教材**、展示している**会話集**を、インターネット上で公開しています。 で検索してみてください。

しまむにサロン関係のお知らせを配信している**公式ラインアカウント**もあります！検索してともだち登録してみてください。

横山・山田の沖永良部での実践

AI、大規模言語モデルで再活性化を容易にする工夫

- 集まれ動物の森
- Metaverse
- 生成AIによる画像生成

- 大規模言語モデル
- ChatGPT
- Vrew
- Whisper

7. おわりに

- まだ市民科学としての敷居が高すぎる。
- 使いやすいテンプレートやマニュアルの作成
  - 辞書作り、絵本づくり、動画、アニメーション
- メタデータ付与の自動化（グロス、書き起こし、字幕など）

おまけ１

- 音声認識の実力

- 石臼１
- 石臼２
- 石臼３

おまけ２：ChatGPT４と危機言語再生に関する対話

- 大規模言語モデルと言語多様性について

ChatGPT3.5

ChatGPT4

- グロッサリーの付け方

ChatGPT3.5

ChatPT4

８．結論

- **現在の日本語の言語状況は文化多様性と逆の方向危機言語の記録・維持・再活性化の問題は「低資源言語処理」の問題に帰着する。**
- **ChatGPTのような大規模言語モデルは原理的には低資源言語処理に使える。実際に使うためには専門知識と調整が必要。**