アーカイブシステムを兼ねた方言データ可視化の新データベースに関する説明会

セリック・ケナン（国立国語研究所）

令和7年度・第1回
「危機言語の保存と日琉諸語のプロソディー」
合同研究発表会・2025年8月21日・琉球大学

NINJAL
National Institute for Japanese Language and Linguistics

0. 説明会の流れ

- 第一部
  - 「新データベース説明」(https://kikigengo.ninjal.ac.jp/nrdb/)
  - 質疑応答

- 第二部
  - 「同源語接続型枠組み(UniCog)」(https://doi.org/10.15084/0002000156)
  - 質疑応答

1. 新データベース説明

1.1 背景
1.2 問題および課題
1.3 解決法と新DBの基本的理念
1.4 機能概要
1.5 「データセット」という単位
1.6 データ登録の流れ
1.7 （仮）運営方針・リリーススケージュール
1.8 質疑応答

1.1 背景

2018年

2024年

- 『日本の消滅危機言語データベース』（以下「危機DB」）
  - 語彙DB
  - 談話DB
- 危機DBの変遷
  - 2016～2021年度の、国語研基幹型研究PJ「日本の消滅危機言語・方言の記録とドキュメンテーションの作成」（代表：木部伸子）の事業として**内部**で制作され、2016年（？）にリリース
  - 2018年度、既存のシステムに解消困難な不具合などがあったため、**内部**で大きく改変（作り直し）
  - 2021年度、PJホームページー新を決定、DBも含め**外部**で発注
  - 危機DBは2022～2027年度PJ「消滅危機言語の保存研究」（代表：山田真寛）に引き継がれ、機能拡張のため、2022年度と2023年度に同じ業者に発注して、2回の改変を経て現行のDBに至る

1.1 背景

- 現行「危機DB」システム変遷
  - 2022年3月（改新）
    - 『みんなふつ語彙集』（http://doi.org/10.15084/00003679）に対応できるシステム
    - 登録機能（エクセル登録）
  - 2023年3月（改変）
    - メタデータ（出典・話者・調査）追加
    - UniCog ID追加
  - 2024年3月（改変）
    - 音声再生ボタン
    - 例文表示の改変（階層的表示が可能）
    - UniCogの内部実装
    - 他の調整（デザイン・検索範囲など）

1.1 背景

- 現行「危機DB」システム変遷
  - 2022年3月（改新）
    - 『みんなふつ語彙集』（http://doi.org/10.15084/00003679）に対応できるシステム
    - 登録機能（エクセル登録）
  - 2023年3月（改変）
    - メタデータ（出典・調査日時）追加
    - UniCog ID追加
  - 2024年3月（改変）
    - 音声再生ボタン
    - 例文表示の改変（階層化）
    - UniCogの内部実装
    - 他の調整（デザイン・検索範囲など）

- 複雑なデータ構造を持つ**琉和辞典**を（相対的に）簡単に編集できるエクセルファイルでのデータフォーマット
- 構造として5つの関連表から成る
  - 「項目表」（primary表）
  - 「意味表」（「項目」にぶら下がり）

  語彙
  - 「単独音声表」（「項目」にぶら下がり）
  - 「音声資料表」（「項目」にぶら下がり）

  音声
  - 「例文表」（「意味」にぶら下がり）
- …が、編集用には、3つの表としてまとめてある

1.1 背景

- 現行「危機DB」システム仕様
  - 複雑なデータ構造を持つ琉和辞典を登録可能
  - 種々の音声ファイルに対応（単独、アクセント資料、例文）
  - 同源語接続型枠組み（UniCog）を実装（同源語表示）
  - メタデータ登録可能（話者情報・調査情報）
  - 登録機能
    - DB全体の管理者がデータを集約・登録
    - Excelで一括登録
  - バイリンガル：日本語版・英語版
  - スマートフォン（小型スクリーン用デザイン）にも対応

1.1 背景

- データ内容
  - 出典集約型

出典	対象	項目数	単独数	ア資料	例文数	CogID
三樹陽介（2015）	八丈島	1423	1438	0	0	624
松倉昂平（2020）	池田町	567	530	0	241	503
小西いずみ、三樹陽介、吉田雅子（2022）	奈良田	1404	1551	0	1726	1133
平塚雄亮（2019）	甑島里	935	936	0	935	185
小川晋史（2013）	喜界島諸方言	813	827	0	0	577
木部暢子、小川晋史（2013）	喜界島諸方言	1534	1878	0	34	1034
横山晶子（2017）	沖永良部諸方言	992	1019	0	0	681
知名町公民館講座（2022）	沖永良部諸方言	1161	2306	0	0	676
木部暢子（2015）	与論	1370	1396	0	0	352
カルリノ・サルバトーレ（2018）	伊平屋田名	855	896	0	0	297
林由華、セリック・ケナン（2018）	池間西原	384	403	400	0	266
セリック・ケナン（2018）	砂川	570	606	0	118	363
セリック・ケナン、大浦辰夫（2022）	水納島	5780	4432	14443	610	2123
セリック・ケナン、麻生玲子、中澤光平(2023)	波照間	3615	7301	3509	0	2000
山田真寛、中澤光平（2018）	与那国	476	491	0	0	274
『どぅなんむぬい辞典 第二版』	与那国	1925	0	0	0	1071
五十嵐陽介（2016）	JR-COGNATES	1876	0	0	0	1775
松倉昂平（2023）	大野市大字上打波	2198	2623	276	134	1431
セリック・ケナン, 麻生玲子, 中澤光平, 中川奈津子（2021）	八重山語川平	890	1424	9	136	640
	合計	28768	30057	18637	3934	16005

2025年5月1日時点（last更新2023年10月19日）

1.2 問題および課題

- 「危機DB」登録機能
  - コントロールパネルよりエクセルファイルを使ってデータを登録
    - メタデータの各表（「地点」、「出典」等）は一個ずつの登録
    - 語彙データは全データ更新型、かつ、全ての表（「語彙」、「音声」等）を一回で登録
  - 登録ロジック
    - メタデータを登録すると、語彙データが消される（＞もう一度登録）
    - 語彙登録は、**DBをいったん空にした上**、もう一度、全てのデータをDBに登録する＝＞1個のデータセット（出典）を追加する際、DB全体のファイルをまず作成しておく必要がある

＝＞現行の登録ロジックだと、様々な「弊害」が…

1.2 問題および課題

- 「危機DB」登録機能（問題）
  - **個別のデータセット**の登録・更新が不可で、DB全体を毎回登録しないといけない。これを実施するために、個別のエクセルファイルから、**IDを振り直しながら**、全データを含む、語彙データを構成する5つの表（シート）を持つエクセルファイルを作成する必要があるが、実際に困難な作業
  - データが一定量を超えると、サーバに負担がかかり、登録に時間がかかりすぎ、結果、うまく登録できない（既に該当…）
  - Synchronisationの機能がないため、**既存データセットの更新**を行うのに向いていない（更新するにあたり、困難な作業が必要）＜＝最も深刻な問題
  - 既存のデータでも、DBから消す＞DBに登録、という流れなので、**安定したID**がなく、その結果、安定したURLが保証されない。

1.2 問題および課題

データ配布の課題

- 本来の発想は入手可能なデータを可視化するに過ぎないが、現状では
  - レポジトリーなどとの連動ができていない
  - 語彙データおよび音声データの配布ができていない

実は、以上の問題は個別データセットの登録・更新が不可能であることと深く関連している。つまり、「データセットごと」と「データの一括登録」とは**相性が悪い**

DB管理者というボトルネック

- 管理者を通さないとデータが登録できない＝＞物理的ボトルネック
- 「この管理者ならいやだ」という場合にも支障＝＞感情的ボトルネック

1.2 問題および課題

- ITの外部発注の難しさ
  - 言語学のデータはspecificity度が極めて高い
    - 既存の枠組の転用と相性が悪い
    - specificであるほど、費用が高額（当然、普通は限られた予算）
  - 指示を正確に伝えることが困難

＝＞複雑なシステムの場合、内部で制作可能なら、内部で制作した方が方断トツといい

1.3解決法と新DBの基本的理念

- （より）理想的な設計として
  - 現行DBと少なとも同じ機能
  - データセットごとの登録・更新（データセットを単位に）
  - 各データセットの更新履歴
  - 各データセットのデータ配布
  - 安定なIDなどのため、Synchronisationシステムの完備
  - 管理者ボトルネックの排除

1.3解決法と新DBの基本的理念

- 基本的な理念
  - 利便性（ユーザーとしてスムーズなデータ閲覧）
  - 透明性（データはいつ誰が更新したかをなど）
  - 再現性（DBの各バージョンを再現可能に）
  - 入手容易性（データはバージョンを問わず常に入手可能）
  - データ利用の明示性（ライセンスの明示など）
  - 自由性（データ作成者に可能な限り自由を）
  - 非属人性（運用が特定の人物を介さない）

- その他
  - バイリンガル版は煩雑さ・コストに全く見合わない＝＞モノリンガル版
  - 構造・コードが可能な限りシンプルに保つ

1.3解決法と新DBの基本的理念

- 『日琉言宝』(https://kikigengo.ninjal.ac.jp/nrdb/)でほぼ全て解決・実装
  - 現行危機DBの機能を全て再現の上、さらなる機能も
  - データセットごとの登録・更新
  - 外部ユーザーの登録機能
  - Synchronisationシステムの完備
  - 自動的バージョニングと更新ログの自動作成
  - データ配布システムの実装

  そのほかに
  - 音声ファイルの音量統一機能
  - 地点の地理的情報の活用

  …

1.3解決法と新DBの基本的理念

- 『日琉言宝』(https://kikigengo.ninjal.ac.jp/nrdb/)でほぼ全て解決・実装
  - 注意点
    - 完全に自作のシステム(費用ゼロ)
    - 国立国語研究所の事業ではなく、**個人の研究PJ**による制作物
    - 試作
    - 現在、「危機言語」PJのサーバーでホストしていただいているが、将来移行する可能性が高い(半永久的にどこに置くかの問題)
  - 参考までに
    - 本格的な自作が可能になっているのは、ChatGPT
    - 命名は、方言は宝であるというポジティブなイメージを積極的に取り入れる

1.4 機能概要

- 閲覧（可視化）機能
  - 方言辞典モード（危機DBと同じ）
    - 方言辞典を閲覧する感じ
    - 横断的検索
  - 語源辞典モード（新規）
    - 語源辞典を閲覧する感じ
    - 日琉諸語の語源辞典が存在しないので、これがおそらく初（？）
  - 「一覧」対「詳細」
    - 検索結果はデーター覧
    - 各項目には詳細ページ

1.4 機能概要

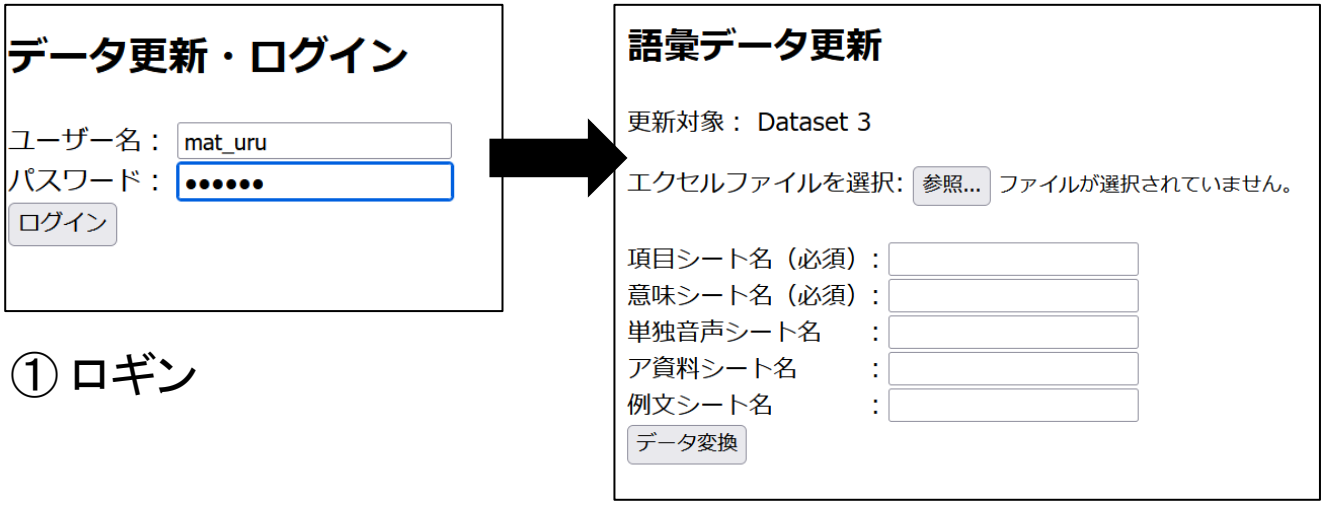
- 語彙データ登録・更新機能（1）
  - 琉和辞典級のデータを登録可能
  - 項目の単独音声の無制限登録
  - 項目の音声資料の無制限登録（構造は例文に準じる）
  - 意味に対する例文の無制限登録（グロスなど実装予定）

例：https://kikigengo.ninjal.ac.jp/nrdb/entry.html?id=49044

https://kikigengo.ninjal.ac.jp/nrdb/entry.html?id=49222

1.4 機能概要

- 語彙データ登録・更新機能（2）
  - アカウント制で、外部ユーザーが実施
  - データセットごと、アカウント付与
  - データセットごとの登録・更新
  - エクセルファイルで登録・更新
  - アカウントの主は随意に・随時に自分のデータセットを更新できる



① ロギン

② コントロールパネルで登録・更新

1.4 機能概要

- 語彙データ登録・更新機能（3）
  - 登録・更新はいわゆるSynchronisation with pruning
    - データセットでは、データセットの内部ID、DBではDBの内部ID
    - 外部ID ／DBIDのMappingを記録
  - Mappingが実装されているため、 synchronization with pruning が可能になる
    - DBシステムにないが、個別データセットにある行を追加
    - DBにあるが、個別データセットにない行を削除
    - DB・個別データセットの両方にある行は、比較して、変更なら更新

＝＞安定したリンクが得られる

＝＞登録・更新が極めて速く、サーバーへの負担も少ない

1.4 機能概要

- 語彙データの変更履歴と配布
  - 登録・更新ごと、自動的にバージョニング・変更履歴作成・データ配布
    - データが登録・更新されると
      - データセットおよびDB全体のバージョンアップ
      - 変更履歴の自動作成
      - データバックアップの自動的セーブ・ZIPファイル準備
      - DBのmanifest（DB中味記述）を自動作成
    - 上記3点はindexページで自動的に反映される

＝＞DB全体・各データセットの変更歴が透明

＝＞manifestを基に、DBの**全てのバージョンが再現可能**

- 資料論文化・レポジトリーの段階をバイパスできる

1.4 機能概要

- 同源語接続型枠組実装
  - セリック他（2024）の枠組（後述）を内部に実装
    - 語源辞典モードが可能になる
    - DB内に登録されている**独立したデータセットのデータ**を共通のアノテーションで接続することで、各データの価値を高めていく（相乗効果）

- 同源語ID登録・更新について
  - コントロールパネルからでも可能だが、語彙登録とは別
  - が、一般的に作成者によるアノテーションの見込みが薄いため、同源語IDに限って、DB管理者よりも登録可能（後述の利用規定参照）

1.4 機能概要

- 音声ファイルの音声量の統一
  - アップロードされる音声ファイルの音量はバラバラ
  - boostgainという値を自動的にcomputeしてDBに登録
  - boostgainを基に、再生のときにdynamic的に個別音声ファイルに適用し、音量のおおよその統一感を担保
  - ＝＞ばらばらな音声ファイルでも統一感

- その他
  - 各地点の地理的情報を内部に実装

1.4 機能概要

- 『危機DB』と『言宝』比較

比較変数	危機DB	言宝
作成	発注	自作
コスト	高い	ゼロ
システム複雑さ	複雑（外部LIB有）	シンプル（外部LIB無し）
システム調整	難しい	フルコントロール
機能／丈夫さ	同じ	同じ
登録	一括型	データセット型
Synchronization	無し	有り
登録	管理者のみ	外部ユーザー
Versioning/logs	手動	自動
音量調整	無し	あり

1.5「データセット」という単位

- データセットごとの登録・更新
  - アカウント制で、外部ユーザーによる登録・更新
  - 完全分割化（full compartmentalization）を貫く設計
    - データセットごと、一つのアカウント
    - アカウントのコントロールパネルから、当該データセットしか変更できない
    - 逆に、**コントロールパネルからしか変更できない**
    - バージョニング・変更履歴・データ配布もデータセットごと
    - 音声フォルダーもデータセットごと
  - 利点
    - データセットごとの登録・更新のため、高スピードで登録・更新
    - エラー・誤操作を起こしても、当該データセットにのみ影響
    - データの作成者（あるいはデータセット担当者）に高い自由度を保証

1.5「データセット」という単位

- データセットごとの登録・更新
  - アカウント制で、外部ユーザーによる登録・更新
  - 完全分割化（full compartmentalization）を貫く設計
  - 利点
    - データセットごとの登録・更新のため、高スピードで登録・更新
    - エラー・誤操作を起こしても、当該データセットにのみ影響
    - データの作成者（あるいはデータセット担当者）に高い自由度を保証
      - ◆ データ構造については指定があるが、データ内容については完全に無指定

      ＝＞作成者の自由度を損なわずに健全なデータ構造に基づいたデータ作成を促進

      例えていうと、部屋を借りる感覚

1.7（仮）運営方針・スケージュール

- 現状
  - 制作・管理：セリック
  - ホスト：山田PJサーバー（今後変更ありうる）
  - 試作

- データ登録の流れ
  - User側：アカウント申請（今はセリックまで）※
  - User側：利用規定承諾・データセット関連情報提供※
  - 管理側：アカウント・データセットアップ
  - User側：初回登録、必要に応じて音声ファイルアップ（未実装）

※将来はフォーム申請などを導入して直接のやり取りを可能な限り無くしたい

1.7（仮）運営方針・スケージュール

- 3種類の利用規定
  - アカウント付与に関する利用規定（DB管理側決定）
    - これから議論して決定
    - 同源語アノテーションや横断的なアノテーションの付与には承諾してもらう（ライセンス付与とも関連）
    - インフォーマントの承諾書を更新する必要があるかも？
  - DBの一般的なユーザー（DB管理側決定）
    - これから議論して決定
  - データセットごとの利用規定（作成者側決定）
    - データが配布されているため、データセットごと利用規定を定める必要がある。これに関しては、DB内部にライセンス情報を実装して、ライセンス付与を少なくとも強く推薦する方向

1.7（仮）運営方針・スケージュール

- スケージュール
  - 第一バージョンとして、コーディングは95パーセント終了
  - 今年度中に試作として運営、内部・外部協力者に利用・テストしていただく
  - 今年度中、利用規定などを整備
  - 来年度前半辺り、アカウント申請を完全にオープンにする

- 現行危機DB
  - コストも高い、運用も煩雑のため、更新なしで、今期（2027年度）までとする
  - 既公開データについては、順次に作成者に連絡して移行を提案する（が、移行するかどうか当然**自由で**、**作成者の判断**です）

1.7（仮）運営方針・スケージュール

- データ（とりあえず）
  - セリック一連のデータ（南琉球）
  - 国語研関係のデータ（及位、沖縄語辞典…）
  - その他、内外協力者のデータ（沖縄北部、沖永良部、福井…）
  - 利用・テストに興味があれば、セリックまでご連絡を。

- 今後の拡張
  - 近日実装：例文データベース（複雑な語彙データに比べて単純）
  - 将来実装（？）：データを使った教材の動的作成（DB内のデータを使って、学習コースを組む）

1.8 質疑応答

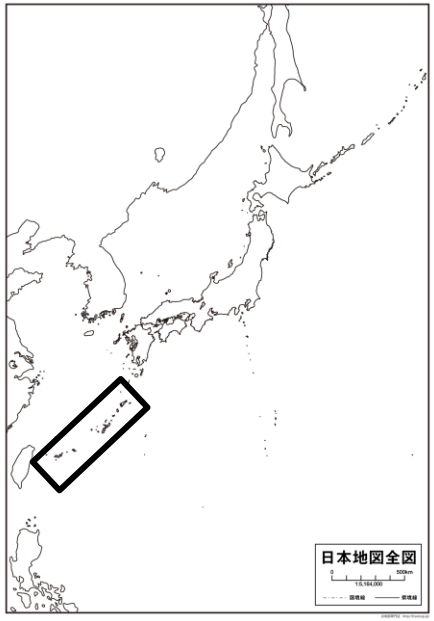
2. 同源語接続型枠組み（UniCog）

Celik Kenan, Kohei Nakazawa, Reiko Aso (2024) UniCog: A Framework Proposal for the Dynamic Compilation of Comparative Data for the Reconstruction of proto-Ryukyuan. *NINJAL Research Papers* 26: 69-97
https://doi.org/10.15084/0002000156

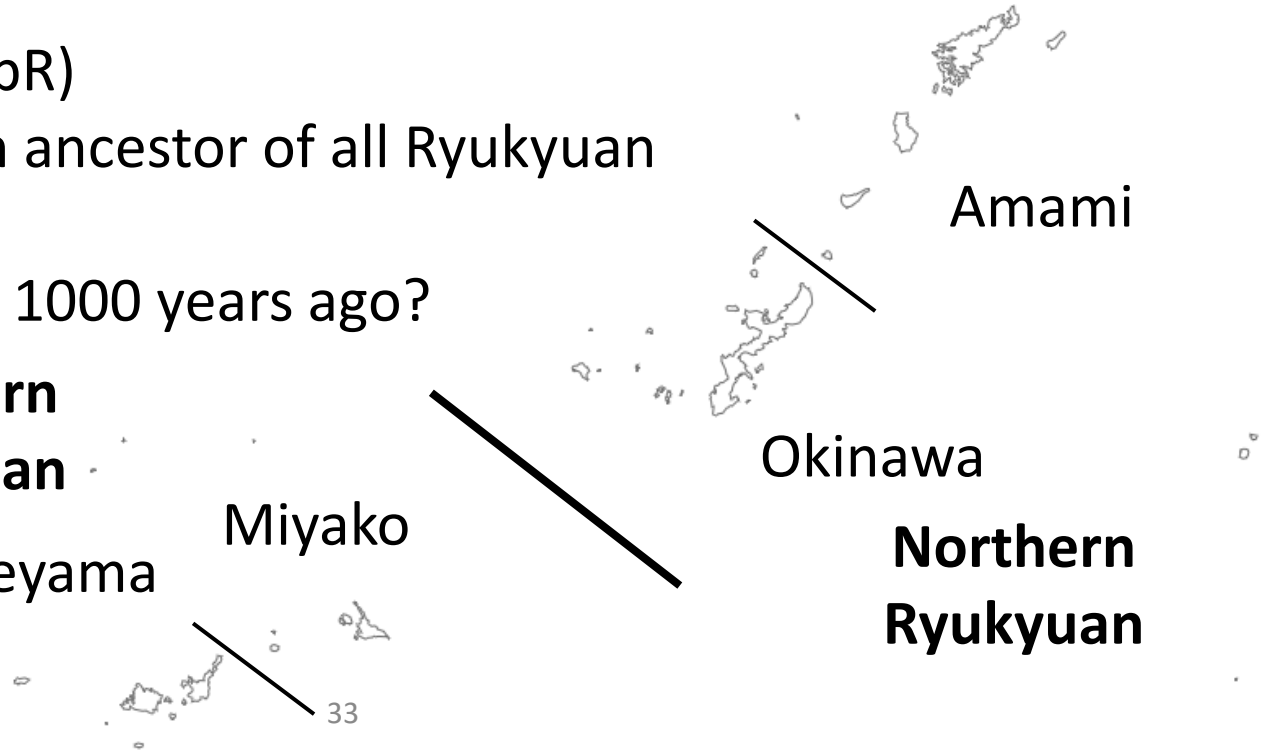
2.1 Introduction
2.2 Previous research
2.3 UniCog: Principles
2.4 UniCog: Applications
2.5 Further possibilities
2.6 質疑応答

2.1 Introduction

- Ryukyuan languages
  - Branch of Japonic languages
  - More than 4 languages
  - High linguistic diversity



- proto-Ryukyuan (pR)
  - most common ancestor of all Ryukyuan languages
  - Shallow: appr. 1000 years ago?



2.1 Introduction

- Reconstruction of pR important task for investigating the history of the Japonic languages

  - Already a somewhat important body of work concerning pR (Hattori 1978-1979; Thorpe 1983; Pellard 2009, 2015)
  - However, **no etymological dictionary of Ryukyuan** (!) (in a sense, “blind” concerning cognates in Ryukyuan)

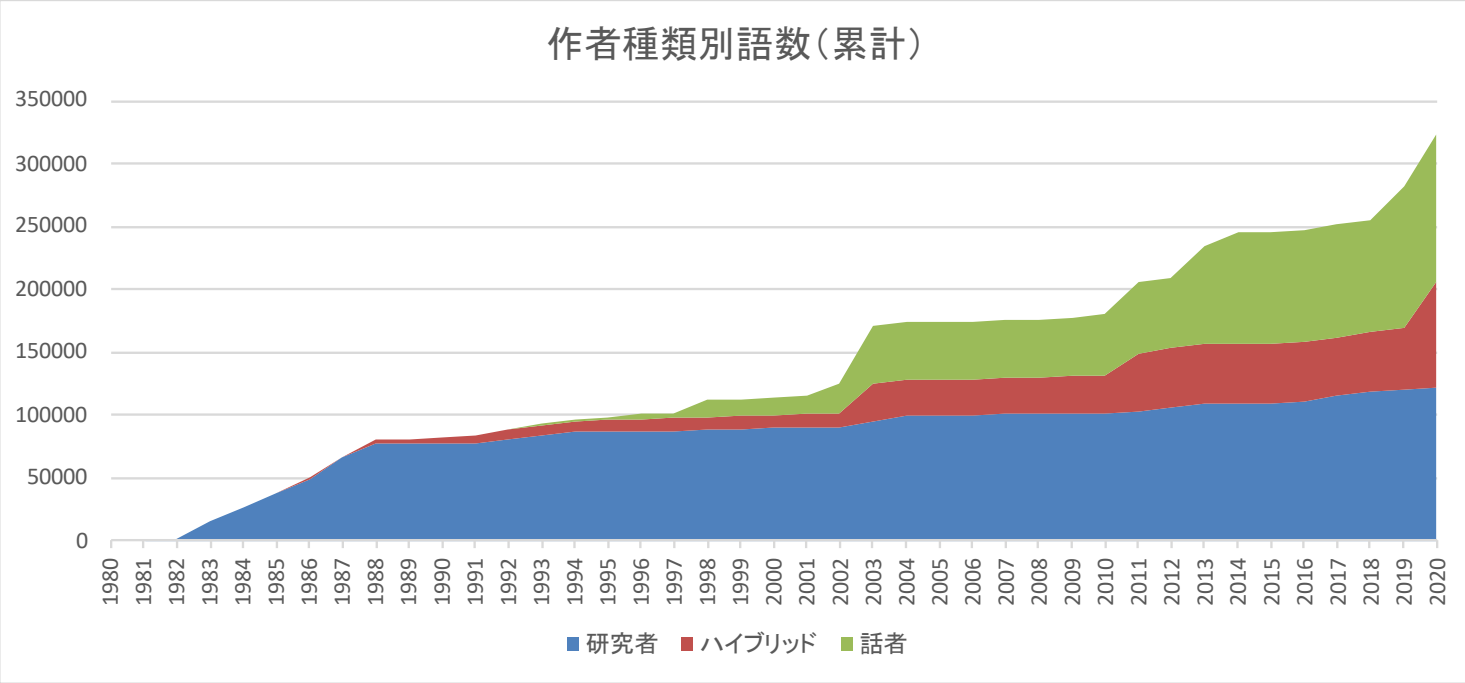
Also,

  - Recently (2010～), enormous stream of new data on Ryukyuan (lexical and grammatical)
  - Results on pR should be tested against this new data and updated accordingly

2.1 Introduction

	Researcher	Hybrid	Speakers	Total
Materials	152	22	18	192
Varieties	57	6	12	58
Words	122797	83292	121630	326795

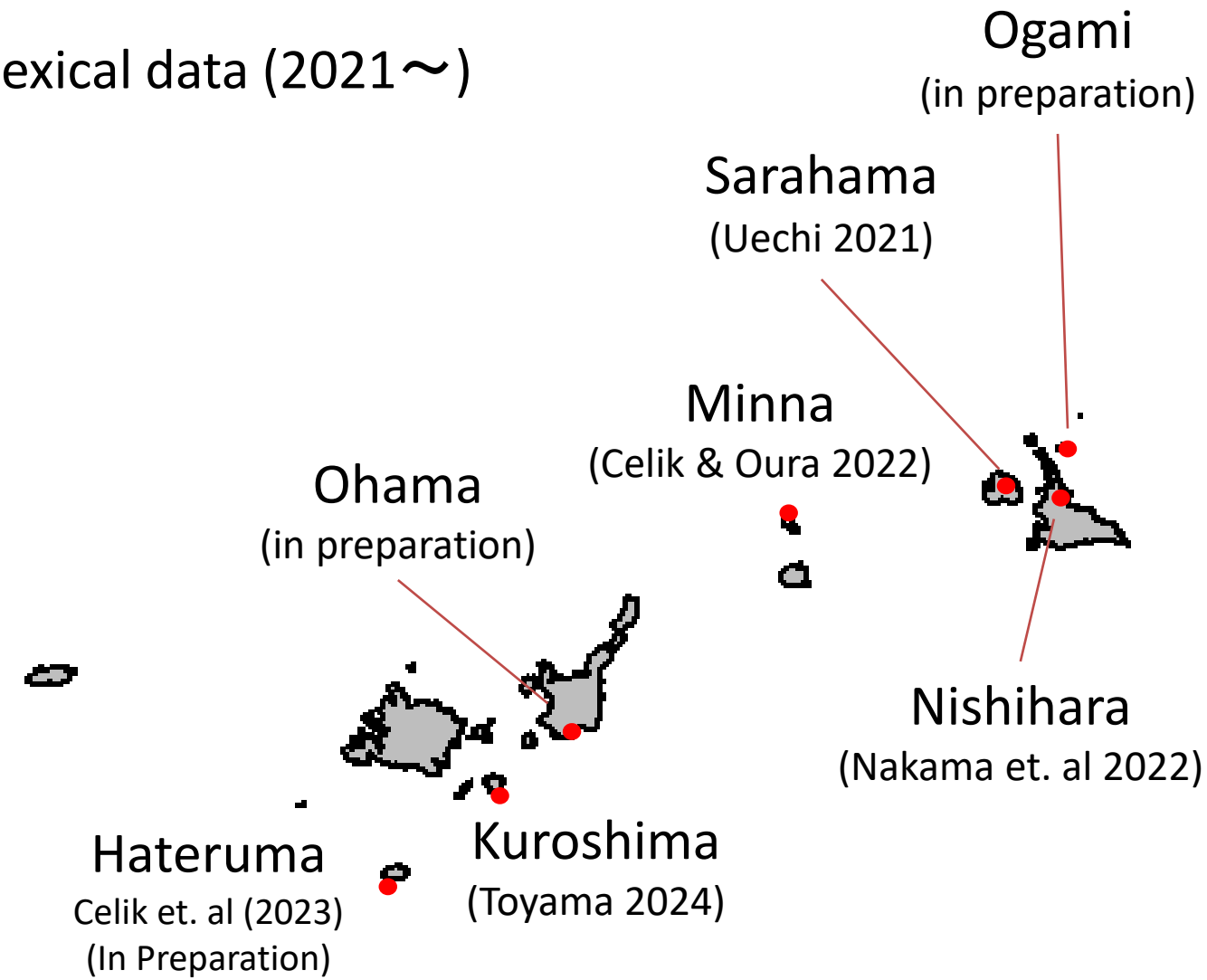
- Stream of new lexical data



Number of words published between 1980 – 2020 for Southern Ryukyuan (Blue: Researcher, Red: Researcher-Speaker, Green: Speaker taken from Aso et al. (2022)

2.1 Introduction

- Stream of new lexical data (2021～)



2.1 Introduction

- Question: how to integrate lexical data for comparative purposes?

- Our solution: aggregate data-building model (Celik et. al 2024)
  Unified Cognacy Framework for proto-Ryukyuan

UniCog

- List of 7,400 Ryukyuan cognates
- IDs for connecting lexical data

国立国語研究所論集 (NINJAL Research Papers) 26: 69–97 (2024)
ISSN 2186-1358

UniCog: A Framework Proposal for the Dynamic Compilation of Comparative Data for the Reconstruction of proto-Ryukyuan

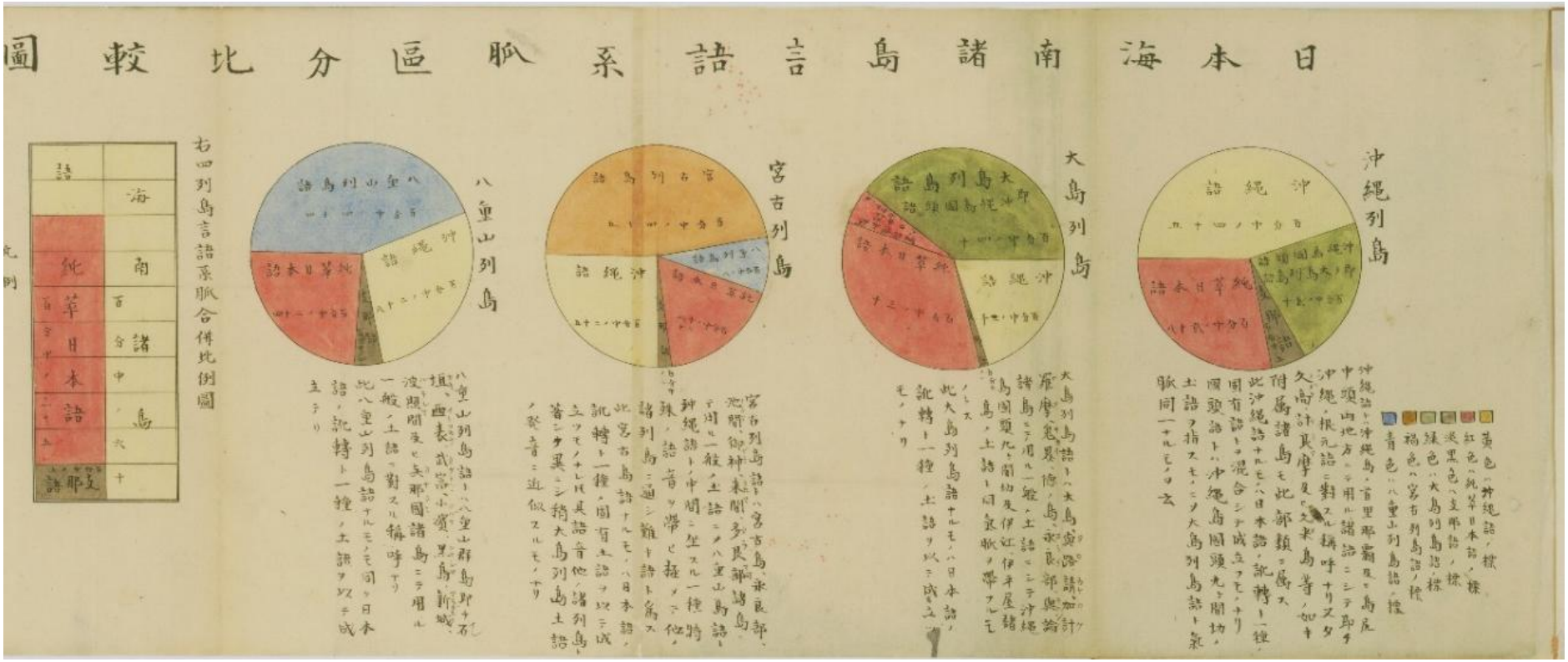
CELIK Kenan[a] NAKAZAWA Kohei[b] ASO Reiko[c]
[a]Research Department, NINJAL
[b]Shinshu University
[c]Meio University

Abstract
In this paper, we propose a framework which includes an approximately 7,400-word cognate list for the dynamic compilation of comparative data with the goal of reconstructing proto-Ryukyuan. The core concept of this framework, which we call UniCog (**Uni**fied **Cog**nacy Framework for proto-Ryukyuan), is to provide a cognate ID system to link all the existing lexicographic data of the Ryukyuan languages. We then show what can actually be achieved with this framework in terms of dynamic compilation of comparative data. Lastly, we propose a standardized orthography for all Ryukyuan dialects for the specific aim of comparing and aligning the word-forms between the doculects. We hope that the introduction of this framework will make some contribution to the field of historical linguistics of Japonic languages.*

Keywords: proto-Ryukyuan, lexicography, historical linguistics, cognate ID

2.2 Previous research

- Tashiro (Meiji): the first to adopt a comparative approach to the Ryukyuan languages in Japan



2.2 Previous research

- Nevskiy (pre-war): centered on Miyako but comparative, unfinished

(1) Cognates (and related forms) given for the entry *adu* 'heel' (from Jarosz (2015b: 5), see Jarosz (2015a) for the legend of the abbreviations)
[(Ya) adu (Rk) aduu (Jap. Kumamoto, 肥後 Oita-ken) ado (Jap. Aomori) agüdo (Sado) akuto (ヤラ) (イト) aru (アラ) atu (カサ) (ヤマト) kado (イリ) kadu (ナセ，カサ，ヤマト，コニ，イス，スミ，サネ，ヒヨ) ado (キカ，トク，イセ，ヨロ，ナゴ，シユ，ナハ，クロ，ハテ，ヨナ) adu (エラ) a:du
(Wakunkan) あくと 三議一統に見ゆ きびすをいへり 今も東国はきもいひ又あどともいふ。
(Rigensuran増) あくつ 三陸越後の方言きびすをいふ。跟]

2.2 Previous research

- Thorpe (1983): outstanding work, but few etyma reconstructed (～300)
- Hirayama (1983-1994): Compilation of basic vocabulary of 72 Japonic varieties (inc. 7 Ryukyuan varieties)
- Martin (1987): Extensive word list, however heavily focused on Japanese central language
- Matsumori (2000): “Classified vocabulary” for pR
- Igarashi (2016): Classified proto-forms of pR (～1800 in v.7)

2.2 Previous research

- Igarashi (2016) is the closest to what we need, however ...

B	H	I	J	K	L	M	N	O	P	Q	R	S	V	W
琉語類	拍	田一	語類	日本	琉球	品詞	同源語ラベル	漢字	カナ	同源性	異語	品詞	意味	琉球祖語形(調査の参
1693	2	I	IA	I	A	形	赤い(アカイ)	赤い	アカイ	日琉		形容詞	赤い(アカイ)	aka
1694	2	I	IA	I	A	形	浅い(アサイ)	浅い	アサイ	日琉		形容詞	浅い(アサイ)	asa-
1695	4		oA	o	A	形	惜しい(アタラシ	惜しい	アタラシイ	日琉		形容詞	勿体ない(モッタ	atarasi-
1696	2	I	IA	I	A	形	厚い(アツイ)	厚い	アツイ	日琉		形容詞	厚い(アツイ)	atu-
1698	2	I	IA	I	A	形	甘い(アマイ)	甘い	アマイ	日琉		形容詞	甘い(アマイ)	ama-
1699	3	I	IA	I	A	形	怪しい(アヤシ	怪しい	アヤシイ	日琉		形容詞	怪しい(アヤシイ)	ajasi-
1700	2	I	IA	I	A	形	荒い(アライ)	荒い	アライ	日琉		形容詞	荒い(アライ)／粗	ara-
1705	3	I	IA	I	A	形	卑しい(イヤシ	卑しい	イヤシイ	日琉		形容詞	卑しい(イヤシイ)	ijasi-
1706	2	I	IA	I	A	形	薄い(ウスイ)	薄い	ウスイ	日琉		形容詞	薄い(ウスイ)	usu-
1711	2	I	IA	I	A	形	遅い(オソイ)	遅い	オソイ	日琉		形容詞	遅い(オソイ)	oso-
1714	2	I	IA	I	A	形	重い(オモイ)	重い	オモイ	日琉		形容詞	重い(オモイ)	obu-
1715	2	I	IA	I	A	形	堅い(カタイ)	堅い	カタイ	日琉		形容詞	固い(カタイ)／濃	kata-
1716	3	I	IA	I	A	形	悲しい(カナシ	悲しい	カナシイ	日琉		形容詞	悲しい(カナシイ)	kanasi-

"JR-COGNATES"

- Somewhat limited size (e.g. very few verbs)
- No comparative data ⇒ difficult to assess the validity of the reconstruction
- Extractive data-building model (link with source lost)

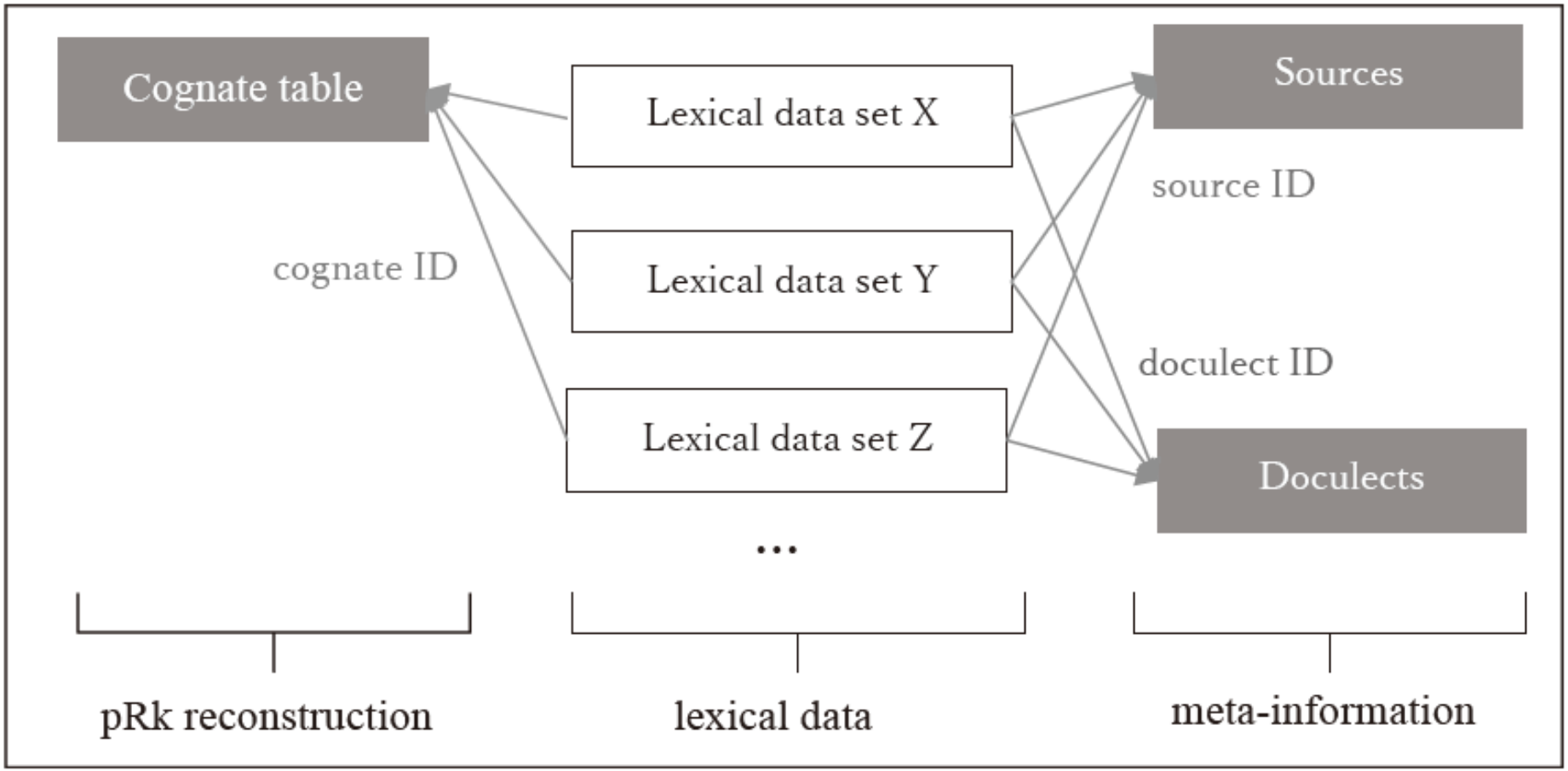
2.3 UniCog: Principles

- The idea behind UniCog
  - aggregate data-building: always keep link with the source
    ⇒ **Connect through cognacy independent lexical resources**

  - Cahier des charges (requirement):
    - ✓ 'light' framework for annotating existing lexical data
    - ✓ Annotate without the need for reconstruction
    - ✓ Devise a system of cognate annotation efficient to implement

2.3 UniCog: Principles

- Overall design and components



2.3 UniCog: Principles

- Overall design and components (cognate table)

番号ID	代表ID	琉祖形	日祖形	琉	日	第一	第二	第三	簡	品	拍	活	末	分	語	琉	対	対応語	相当語
157	味見.a.n	amame		1	-	味見.a.n			n	名	1	3				転				味見（アジミ）
158	宿借.a.n	amamV		1	-	宿借.a.n			n	名	1	3			動					宿借（ヤドカリ）
159	余.a.v	amar-		1	-	余.a.v			v	動	3	3	1	r					余る（アマル）	余る（アマル）
160	余.a.v.2	amar-as-		1	-	余.a.v	使役.a.o		v	動	3	4	1	s					余らせる（アマラセル	余らせる（アマラセル）
161	余.a.n	amar-i		1	-	余.a.n			n	名	1	3				転			余り（アマリ）	余り（アマリ）
162	余.a.v.3	amas-		1	-	余.a.v.3			v	動	3	3	1	s					余す（アマス）	余す（アマス）
163	甘さ.a.n	ama-sa		1	-	甘.a.a	名詞化.s.o		n	名	1	3							甘さ（アマサ）	甘さ（アマサ）
164	髪.a.n.2	ama[d][u]		1	-	髪.a.n.2			n	名	1	3			体					髪（カミ）
165	飴.a.n	ame		1	-	飴.a.n			n	名	1	2					o		飴（アメ）	飴（アメ）
166	雨.a.n	ame		1	-	雨.a.n			n	名	1	2					o		雨（アメ）	雨（アメ）
167	浴.a.v	ame-		1	-	浴.a.v			v	動	3	2	2	me					浴びる（アビル）	浴びる（アビル）
168	雨風.a.n	ame+kaze		1	-	雨.a.n	風.k.n		n	名	1	4							雨風（アメカゼ）	雨風（アメカゼ）。嵐（アラシ）
169	雨雲.a.n	ame+kumo		1	-	雨.a.n	雲.k.n		n	名	1	4						広	雨雲（アマグモ）	雨雲（アマグモ）
170	虹.a.n	ame+nomi-ja(?)		1	-	雨.a.n	飲.n.v	屋.j.o	n	名	1	5						広	雨（アメ）+飲み（ノ	虹（ニジ）
171	雨垂.a.n.	ame+tari		1	-	雨.a.n	垂.t.n		n	名	1	4						広	雨垂れ（アマダレ）	雨垂れ（アマダレ）。軒（ノキ）
172	浴.a.v.2	ames-		1	-	浴.a.v.2			v	動	3	3	1	s					浴びせる（アビセル）	浴びせる（アビセル）。浴びさせ

- ✓ Any word whose **cognates** are distributed in more than two Ryukyuan varieties
- ✓ Various relevant annotations for each cognate set

2.3 UniCog: Principles

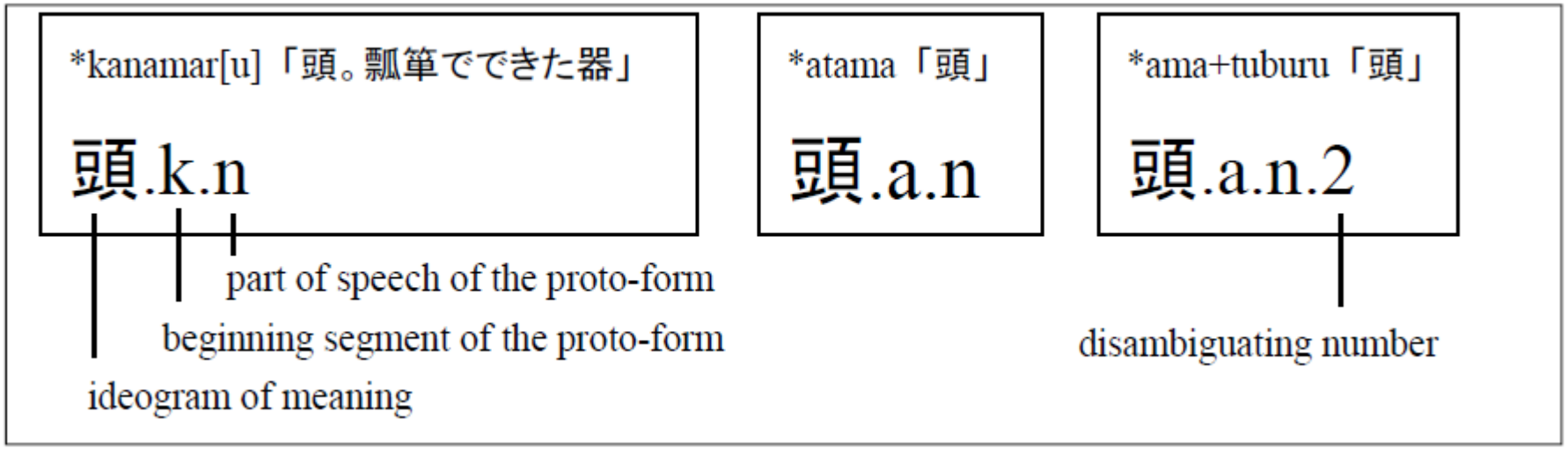
- **Cognate set** is the **primary key** for the whole DB
  ⇒ Importance of the definition of cognacy

  "Cognate word-forms which have inherited exactly the same morphological component(s) from the proto-language. That is, word-forms are cognate if and only if they are **diachronically isomorphic**" (= strict cognacy)

  - Shuri *garasi* 'crow', Taketomi *garaʃi* 'crow' cognate, but not Hirara *garasa* 'crow' < *garasu-(j)a
  - *samar- < **sama-r- and *same- < **sama-i- 'to cool off (itr.)' both contain the root *sama but are are grouped into two different cognate sets because of the difference in verbal formant

2.3 UniCog: Principles

- Cognate ID system for efficient annotation of lexical sources
  - not possible to learn all IDs for annotating
  - ID should be retrievable with minimal knowledge
    - ID = Ideogram of meaning + first segment + PoS



2.3 UniCog: Principles

- Cognate ID system for efficient annotation of lexical sources

Simplex forms

- *amar- ‘remain (itr.)’ 余.a.v
- *pata ‘flag’ 旗.p.n
- *pije- ‘cold’ 寒.p.a
- *taja ‘strength’ .....

Complex forms (combine the ideograms)

- *asa+mono ‘breakfast’ 朝物.a.n
- *miti+sipo ‘high tide’ 満潮.m.n
- *iri+moko ‘adopted son-in-law’ ……

2.3 UniCog: Principles

- Cognate ID system for efficient annotation of lexical sources
  - ‘Representative’ IDs (unique) and ‘working’ IDs (plural)

Some practical problems for ID design:

- A cognate may have: several meanings
- A ‘meaning’ may have: several ideograms

Working IDs are designed to solve this problem
*kata- ‘hard, strong, thick’ (Representative ID 固.k.a)
Working IDs: 密.k.a
堅.k.a.2
濃.k.a

2.3 UniCog: Principles

- How to annotate

2.4 UniCog: Applications

- Lexical sources annotated with UniCog's IDs
  - Dynamic compilation of etymological dictionary of Ryukyuan
  - Sophisticated searches: 日琉言宝
  - Automatic compilation of lexical questionnaires for closely related languages by "subtracting" lexical resources

2.5 Further possibilities

- Traditional glossing not suited(?) as morphological annotation of pluri-dialectal corpus (as far as I tried...)
  - different label for the ‘same’ word when semantic change is involved

- UniCog’s IDs could be used as morphological annotations replacing glossing for annotating corpus of Ryukyuan languages.

2.6 質疑応答