琉球祖語の再建に向けた比較データ構築用の枠組提案（UniCog）

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

1. Introduction

In order to investigate the history of the Japonic languages, the reconstruction of proto-Ryukyuan (pRk), the most recent common ancestor of all the languages belonging to the Ryukyuan branch of the Japonic languages, is a necessary step.

Detailed reconstructions of the phonological system of pRk have already been proposed (Hattori 1978–1979, Thorpe 1983, Pellard 2009, 2015), and, putting aside some marginal details still in dispute (e.g. Celik (2022a)), the outline of the proposed phonological system has not been challenged so far. However, this situation needs to be put into perspective. First, the availability of new lexical data on Ryukyuan doculects has grown at a dizzying pace. Since the beginning of the 21st century, there has been a steady stream of publications of large dictionaries containing more than 10,000 entries (Miyagi (2003), Izena-jima hogen jiten henshu iinkai (2004), Kiku and Takahashi (2005), Maeara (2011), Tomihama (2013), Kajiku (2020), Tokuyama and Celik (2020), Uechi (2021), Higa and Takaesu (2021), Honda (2021) among others). This string of publications is the result of a shift observed in Ryukyuan lexicography, whereby native speakers took the lead in the compilation of large dictionaries (Aso et al. 2022). Secondly, apart from the appendix in Thorpe (1983) listing 267 reconstructed words, there was until very recently no

* This work has been supported by JSPS grants 22F22305, 21H00353, 20H01259, 19K13174, 18K12390. This article is a revised version of a presentation held by the first author at the 2nd 2021 yearly Conference of Yaponesia (13th of February 2022, online).

extensive reconstruction of pRk's vocabulary.

Needless to say, any proposed reconstruction of pRk should be tested against new data and be updated accordingly. The unprecedented growth in lexical data on Ryukyuan dialects opens up the possibility of a robust testing of the models proposed so far. At the same time, it raises the non-trivial problem of how to integrate this uninterrupted flow of new data into the reconstruction. A solution to this problem would be a data management framework that is able to integrate in a continuous fashion the latest data into the existing model of pRk.

A pioneering answer to this problem is Igarashi's "Nichiryugo Ruibetsugoi (Japonic classified vocabulary)" aka "JR-COGNATES" (Igarashi 2016), a periodically updated list of cognates going back to pRk with their reconstructed tonal class, containing in its latest version (v.7, Igarashi (2019)) more than 1,800 words (mainly nouns). Without any doubt, Igarashi's JR-COGNATES is one of the major achievements in the reconstruction of pRk since Thorpe (1983). It is not, however, without limitations. First, mainly due to copyright clearance issues, it does not make public the dialectal data on which the reconstructions are based. It is thus difficult for example to evaluate the validity of the reconstructed tone classes. Besides, the reconstructed form provided for each cognate is not regarded by the author himself as the result of a careful comparison of all the available comparative evidence but simply as a practical reference for quick comparison of the modern dialects (see the explanatory note in Igarashi (2019)). That is to say, JR-COGNATES does not aim at providing the reconstructed forms of pRk per se, so that these cannot be used as a basis for further discussion. Second, the approach adopted, which could be described as "extractive data-building", has inherent limitations. Lexical data is extracted from each lexicographic source into the database without keeping a co-reference with the data in the source. This design, in which the link with the primary source is lost, not only limits research possibilities but, more importantly, it cannot cope efficiently with dynamic lexical data building, in which the primary source is regularly updated. Note also that because it aims at reconstructing the proto-language, words are only included in the list if they are likely to go back to proto-Japonic, or proto-Ryukyuan.

In the face of this, we devised a framework with a radically different approach, which makes it possible to compile comparative data in a dynamic fashion for the reconstruction of pRk. The approach adopted can be described as "aggregate data-building". Namely, we designed a cognate ID system that can link cognate sets across different lexical data sources, whatever their data structure. Since the link with all primary sources is kept, not only does this design efficiently cope with dynamic lexical building, but it also vouches for more flexible data building. For example, since we link cognates to a common cognate set and not to a proto-form, there is no need to decide whether a particular cognate set goes back to the proto-language or not. This approach also ensures transparency in the reconstruction as we can refer to all the dialectal data with a unique ID. Lastly, this design, which links different lexical sources through cognacy, also opens up many research possibilities other than the reconstruction of pRk.

2. Previous research on the compilation of comparative lexical data and the reconstruction of proto-Ryukyuan

In this section, we review the main previous research dealing with the compilation of comparative data and the lexical reconstruction of proto-Ryukyuan.

2.1 Tashiro

One of the first to adopt a comparative approach to the Ryukyuan languages in Japan is Tashiro Antei. In the first half of the Meiji period, Tashiro did extensive fieldwork in the Ryukyus and documented among others all of the four branches (Amami, Okinawa, Miyako, Yaeyama) of the Ryukyuan languages. In 1888, he redacted the results of his fieldwork into a report of several volumes that he submitted to Tokyo University (Tashiro 1888a). This report is thought to have contained originally word lists of Amami, Okinawa, Miyako, and Yaeyama, although only the word list of Yaeyama is extant today (Celik et al. 2021). Part of the word lists of Okinawa and Miyako was published in the form of an article listing the kinship terms of the two languages (Tashiro 1888b), and another part is found in a manuscript draft (Tashiro 1888c) preserved at the library of National Taiwan University.[1]

It is beyond doubt that Tashiro worked within a comparative approach involving cognacy judgements between the Ryukyuan languages and Japanese. This is not only obvious from the article he wrote about the phylogeny of Yaeyama (Tashiro 1894) but is most evident in the graph contained in the fieldwork report (Tashiro 1888d) that compares the make-up of the lexicon of each of the four Ryukyuan languages (Figure 1). In this graph, he evaluates the proportion of the shared vocabulary between the Ryukyuan languages and Japanese. For example, the lexicon of Miyako is given as 25% Okinawa, 20% Japanese, 8% Yaeyama, 2% Chinese while the rest is specific to Miyako. This sort of calculation can only be executed on the basis of cognacy judgements made between the languages. This means that Tashiro's fieldwork report is the first comparative work on the Ryukyuan languages in Japan. Unfortunately, it was never mentioned nor used in the field of Japanese linguistics.

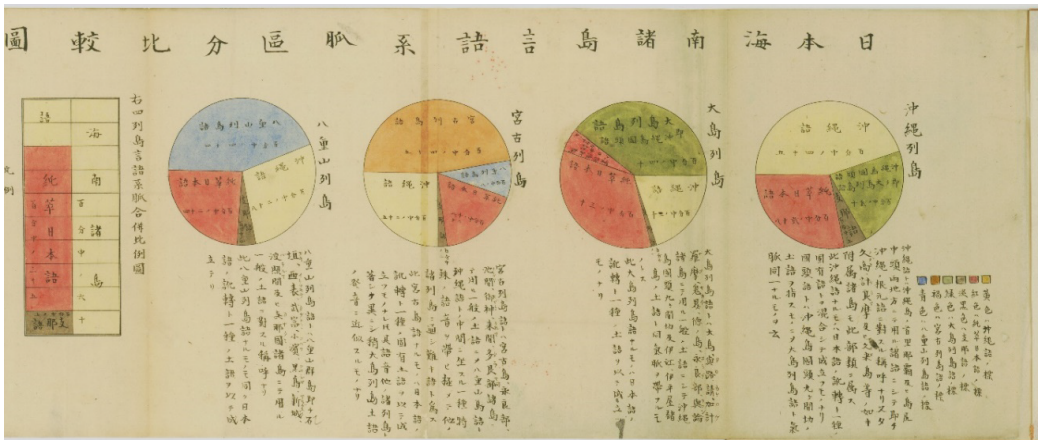

Figure 1. Nihon Kainan Shoto Gengo Keimyaku Kubun Hikaku-zu[2] (from Tashiro 1888d, image cropped)

[1] https://dl.lib.ntu.edu.tw/s/Tashiro/item/714623 (last accessed 2023/07/28).

[2] https://iiif.dl.itc.u-tokyo.ac.jp/repo/s/tashiro/document/b495504c-3487-1d58-210f-ad5f795286a2#?c=0&m=0&s=0&cv=2

2.2 Nevskiy

The next work to include a comparative approach is the dictionary of Miyako compiled mainly in the 1920s by Russian linguist Nikolay Nevskiy. This manuscript, which could not be finished due to the author's untimely death in 1937, is titled "Матерьялы для изучения говора островов Мияко (Materials for the study of the dialect of the Miyako islands)" and contains 5,829 entries of Miyako, principally of the Hirara and Sawada varieties (see Jarosz (2015a) for a detailed description of the materials and Jarosz (2015b) for its full transcription). For many of the entries, Nevskiy lists the cognates found in the other varieties of Japonic, including the contemporary dialects of Japanese. We give an example of the cognate list given for the entry *adu* 'heel' in (1). As can be seen in this example, the search for cognates or related forms can be quite extensive. This is why this manuscript constitutes with its rich annotations of cognates in the other varieties of Japonic a major source on the etymology of Miyako's vocabulary. To give one example, Nevskiy had already identified Yaeyama Hateruma *kanabarï* 'calabash' as a cognate of Miyako Hirara *kanamaz* 'head' (Jarosz 2015b: 244).

(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) aduɯ (Jap. Kumamoto, 肥後 Oita-ken) ado (Jap. Aomori) agɯ̈do (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。
(Rigenśuran増) あくつ 三陸越後の方言きびすをいふ。跟]

2.3 Thorpe

Thorpe (1983) is a systematic reconstruction of pRk phonology, prosody, and morphology based on most of the available data at the time. It also includes an appendix listing 267 cognates sets complete with the comparative data from 41 varieties of Ryukyuan and the reconstructed form in pRk, thus making it a basic reference work for pRk. Tone class is however not reconstructed in this work.

2.4 Hirayama

Teruo Hirayama has conducted extensive fieldwork studies on the basic vocabulary of numerous varieties of the Japonic languages. The results of these studies have been published among others in Hirayama (1983, 1986, 1988), covering 10 varieties of Ryukyuan, and especially in Hirayama et al. (ed.) (1992–1994), which records about 230,000 entries from 72 varieties of Japonic (65 varieties of Japanese and 7 varieties of Ryukyuan, including the 4 varieties whose data is taken from Hirayama (1983)). These works are based on a common lexical questionnaire of approximately 2,400 concepts expressed in Standard Japanese and are organized in the same way. That is, the concepts from the questionnaire serve as entries under which the lexical data from the different doculects are listed. Each concept is typically represented by its pronunciation in hiragana along with its ideogrammatic orthography, such as "かた【肩】" (*kata* 'shoulder'). While each entry contains the translation of the headword in the target doculects, as it was used as such

for lexical elicitation, it is not restricted to it. Indeed, all the words that were elicited when discussing the concept expressed by the headword are also listed under the same entry. This means for example that the names of all sorts of tree species are listed in the entry for 'tree', just like the names of all sorts of baskets are listed in the entry for 'basket'.

These studies are particularly important in many respects. First, they may constitute the greatest source available of primary data on the basic vocabulary of the Japonic languages. Second, the varieties covered are well balanced, and correspond to traditional varieties that may not be spoken anymore. Third, it contains for most of the varieties the information on word prosody. Lastly, the data has been gathered and organized in a standardized format, thus enabling a straightforward comparison of the data. On the other hand, the structure of the data imposes some limitations on cognate based lexical comparison. As explained above, the lexical data is not organized according to cognacy, but according to the source concept through which a form was elicited in the target language. Not only does this make it difficult to know under which entry a particular lexical item would be listed, but it also means that cognacy judgement has first to be applied before lexical comparison. Another, albeit minor, problem, is that the labels of the entries are not always consistent, with some labels lacking, for example, an ideogrammatic orthography.

2.5 Martin

The work of Martin (1987), referred to by Pellard as "a timeless classic" (Pellard forthcoming), remains to this date with its 8,352 entries (4,939 nouns, 2489 verbs, 924 adjectives) the most extensive word list dealing with the reconstruction of proto-forms and tone class in the Japonic languages. Each entry is recorded besides its orthography in Roman alphabet and its meaning in English with the following information: the reconstructed form, the reconstructed tone class, the tone class (when available) found in the Myogisho and the varieties of Tokyo, Kyoto, Kagoshima, Shodon, Shuri and Yonaguni, as well as in some other dialects when needed. Many entries are also provided with etymological and historical notes. This is why it constitutes a major reference work for the etymology and the reconstructed tone class of many Japonic words.

At the same time, it is not without limitations. First, it adopts, as in the case of JR-COGNATES discussed in the introduction, an extractive approach to data-building in which the link with the primary sources is lost. This has the consequence of creating some degree of opacity in the data. For example, the tone class of many varieties of Japonic is given for each entry, but not the word-form, so that it becomes sometimes difficult to guess which word has been deemed cognate with the entry. Intriguing indeed is the tone class A indicated for Shodon, Shuri, and Yonaguni in the entry *sono* 'that ... ' (Martin 1987: 530), knowing that the demonstrative root **so* is not reconstructed for pRk (Kinuhata 2021). This is until one understands that Martin seems to have assumed a cognacy relationship between Japanese *sono* 'that ... ' and pRk **ono* 'that ...'. Another limitation is that, although it lists in a systematic way the cognates in three varieties of Ryukyuan, it is heavily skewed towards Japanese, so that many words distributed only in the Ryukyuan languages are not listed (e.g. pRk *taja C 'strength').

2.6 Matsumori

Akiko Matsumori has to be credited for laying the foundations leading to the creation of the so-called "classified vocabulary" of Ryukyuan (Matsumori 2000a, 2000b, 2012). Classified vocabulary refers to a list of lexical items classified according to the prosodic correspondences

observed between the cognates found in the different varieties, present or past, of one language family. Such a list is basically a reconstruction of the tone class of each word in the proto-language, but it also serves as a lexical questionnaire to gather data on still undocumented varieties. The most famous and widely used list is the one created for Japanese by Kindaichi (Kindaichi 1974). However, as has been shown in Hattori (1958), the prosodic correspondences observed in the Ryukyuan languages do not fit in a straightforward manner those observed in the Japanese varieties, so that Kindaichi's classified vocabulary is not suitable for doing prosodic research on the Ryukyuan languages. Expanding on Hattori's findings through one's own fieldwork results, Matsumori devised a classified vocabulary list specifically designed to reflect the prosodic correspondences observed in the Ryukyuan languages (Matsumori 2000a, 2000b, 2012). This represents a major advance for the reconstruction of the word prosodic system of pRk and for prosodic research on the Ryukyuan languages. It must be noted indeed that Matsumori's framework is at the basis of the flurry of discoveries on the prosodic system of many Southern Ryukyuan varieties, which have been shown to retain the ternary prosodic system of pRk contrary to what was thought before (see for example Matsumori (2010) and Igarashi et al. (2012)).

At the same time, the list proposed in Matsumori (2012) remains with 412 words of somewhat modest size and is limited in the information it provides. It simply consists of the word-forms of three varieties of Northern Ryukyuan (Kikai Akaren, Okinoerabu China, Okinawa Kin) listed by tone class. Neither does it provide a label referring to each cognate set, give the reconstructed form in pRk, nor lists the cognates in the other Ryukyuan languages, thus making it slightly unwieldy when used as a questionnaire.

2.7 Igarashi

Yosuke Igarashi proposed in Igarashi (2016) an extended classified vocabulary list called "Nichiryugo Ruibetsu Goi (Japanese-Ryukyuan classified vocabulary)" aka "JR-COGNATES", which is intended to solve the problems of the lists proposed in the previous literature. Igarashi notes about Kindaichi's list that it suffers from a sample bias, as it is heavily skewed towards the words attested in the prestige variety of Japanese (the variety spoken in Kyoto during the Heian period). First, many words are listed which do not have cognates in the Ryukyuan languages. These words may therefore not go back to proto-Japonic and should maybe not be included in a list with the goal of reconstructing proto-Japonic. Second, many words are absent from the list because they are not attested in the prestige variety, although cognates are found in Ryukyuan and the non-prestige varieties of Japanese. The distribution of these words in Japonic implies that these words are valid candidates to be reconstructed in proto-Japonic, and so should be included in the list. Igarashi also points out that the words whose prosodic correspondences are not deemed regular are excluded on an a priori basis from Kindaichi's list, although the prosodic correspondence is precisely what has to be elucidated.

To resolve the problems pointed above, Igarashi (2016) adopted the following condition to build JR-COGNATES. Namely, the list is constituted by words whose cognates are distributed both in Ryukyuan and in Japanese, whatever the observed prosodic correspondence, and whatever the attestation of cognates in the prestige variety. This condition is to ensure that the words included in the list are likely to go back to proto-Japonic. With the subsequent updates, JR-COGNATES has grown in its 7th version (Igarashi 2019) to a list of almost 1,900 entries (mostly nouns) and is conveniently annotated. It gives among others for each word a cognate

label, its distribution in Japonic, its length expressed in mora, a list of meanings attested in the reflexes, the reconstructed tone class in proto-Ryukyuan and Japanese, and an approximative reconstructed word-form in pRk intended for lexical elicitation. This makes JR-COGNATES the best list available to date for the Ryukyuan languages and has already been used in a number of publications investigating the prosodic correspondences of specific dialects (e.g. Aso and Ogawa (2016), Uwano (2017a), Celik (2020)). It is, however, not without limitations as already discussed in the introduction.

3. Description of the framework

In this section, we describe the basic architecture of the different components of the framework we propose.

3.1 Overall design and components

The framework we propose is designed to be implemented on top of existing lexical data, converting independent lexical data sets into one integrated relational database. To achieve that, we provide the 4 components shown in (2). They are to be implemented as shown in Figure 2.

(2) Components
a) cognate IDs
b) cognate table
c) meta-information on sources
d) meta-information on doculects

The 1st component is a list of cognate IDs. These cognate IDs are added onto existing lexical data so that each record in the data is linked to its corresponding cognate via the cognate ID (see section 3.2). The 2nd component is the cognate table containing detailed information about each of the cognate set defined by the cognate IDs. It contains to this date approximately 7,400 entries (see supplementary materials). Cognates constitute the primary key to the whole

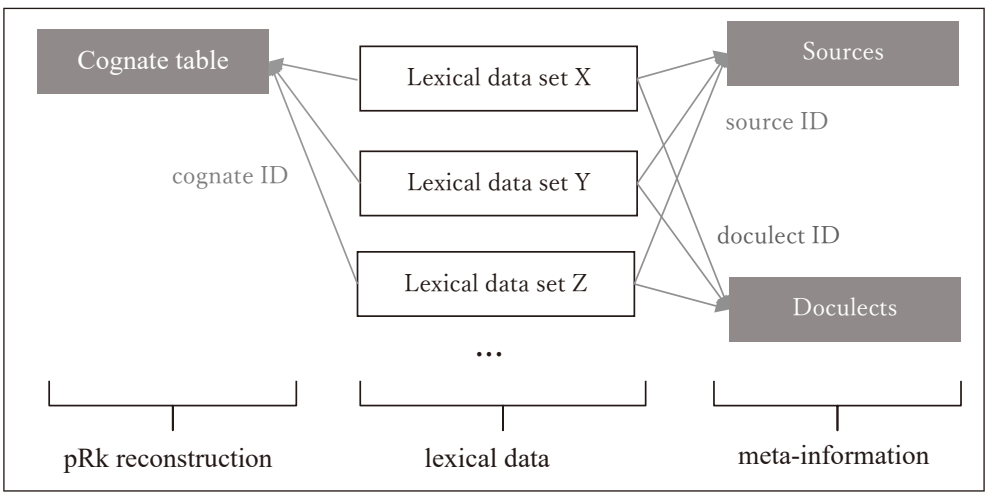

Figure 2. Architecture of UniCog's implementation

database. As noted earlier, not all cognates go back to pRk. This feature makes it possible to build data in a flexible way since there is no need to decide whether a specific cognate set goes back to pRk or not. The 3rd and 4th components correspond respectively to the description of the sources to be integrated into the relational database, and the doculects covered by these sources. They correspond to the meta-information of the data to be contained in the database. Each record from the lexical data sets that are to be integrated has to be annotated with source and doculect IDs. Note that a single source may contain lexical data on several doculects so that it is not possible to link directly source and doculect.

Below we give more details on each component and the system of cognate ID.

3.2 Definition of cognacy and cognate ID system

3.2.1 Definition of cognacy

As explained in section 2.1, cognate set is used as the primary key for the whole database. The definition of "cognacy" adopted is thus central to the implementation of the present framework. In our framework, we consider 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. With this definition, partial cognates, i.e. word-forms that share one or more morphological elements, are assigned to different cognate sets. For example, Shuri *garasi* 'crow' and Taketomi *garaʃi* 'crow' are cognate but not Hirara *garasa* 'crow', because it contains beside the root the suffix *-(j)a*. Likewise, the related verbs **samar-* < **sama-r-* and **same-* < **sama-i-*, both meaning 'to cool off (itr.)' and sharing the same root **sama*, are grouped into two different cognate sets because they are derived from two distinct verbal formants. We also distinguish between the alternative shapes of apophonic nouns, so that for example *ama+kaze* 'rainy wind' and *ame+kaze* 'rainy wind' are assigned to two different cognate sets. The downside of this approach is that the relationship between partial cognates is lost, but we solve this problem by introducing morphological decomposition in the cognate table (see section 3.3).

For practical purposes, we also add part of speech (i.e. the 4 simplified part of speech categories "noun", "verb", "adjectival stem", and "other") as a second criterion for the definition of cognate sets, so that each cognate set is defined as belonging to a single part of speech only. From that criterion, words from different parts of speech are not grouped together even if they are diachronically isomorphic. For example, the Miyako verbal stem *bugari-* 'to get tired' and the adjectival stem *bugari-* 'tired' are assigned to two different cognate sets according to their respective part of speech.

Note that as a consequence of semantic change, the reflexes of a particular cognate set may exhibit widely diverging meanings, as illustrated in (3) (see also the 27 meanings listed for this word in Shogakutosho (1989)).

(3) Meanings attested in the Ryukyuan reflexes of **kamati*
a) 'head': Amami Naze (Hirayama et al. 1967)
b) 'cheek': Miyako Hirara (Hirayama 1983)
c) 'jaw': Yaeyama Taketomi (Maeara 2011)
d) 'doorframe': Okinawa Izena (Izena-jima hogen jiten henshu iinkai 2004)

Based on the definition of cognacy ("isomorphic inheritance") explained above, we can then set up diagnostic criteria like phonological correspondence, tonal correspondence, and semantic

identity in order to motivate cognacy judgements. The difficulty lies in the fact that the diagnostic criteria of cognacy cannot be known aprioristically but only get refined as more data becomes available. This means that at any stage of the research, cognacy judgements may be subject to revision. For example, the words *wata* 'intestine, belly' and *wata* 'cotton padding', being homophonous and of the same reconstructed tonal class, meet the diagnostic criteria of phonological and tonal correspondences. Accordingly, they may be cognate. They are, however, not acknowledged as such in the previous literature (Martin 1987, Sakamoto et al. eds. 1998). This is supposedly because they do not meet the criterion of semantic identity. However, since the meaning of a word undergoes changes over time, adhering to a criterion of strict semantic identity is not tenable for an accurate categorization of reflexes. Suppose for example that the concepts of 'intestine, belly' and 'cotton padding' were to be repeatedly found in the languages of the world expressed by the same word, we could then assume a semantic affinity, and hence a likely path of semantic change between the two concepts and judge *wata* 'intestine, belly' and *wata* 'cotton padding' to be cognate. Still, it remains difficult to decide in an objective manner how much latitude should be allowed in semantic variations. This means that there will always be an intractable residue of arbitrariness in cognacy judgements when implementing the present framework. As in the case of the two words discussed above, words are kept as separate cognate sets in the cognate list we provide as supplementary materials when their cognacy is suspected but not fully warranted by the (admittedly partly arbitrary) criteria at hand.

In some other limited cases, and always out of practical reasons, some cognate sets are split into two sets with a different ID. For example, the word for 'monkey (animal)' and the word for 'monkey (zodiac sign)' are clearly cognate and should therefore be grouped into a single cognate set. However, their respective reflexes in the Ryukyuan languages exhibit systematic differences (Table 1), so that we decided to define two cognate sets, one for 'monkey (animal)' and the other for 'monkey (zodiac sign)'.

Table 1. Reflexes of 'monkey (animal)' and 'monkey (zodiac sign)'[3]

Area	Doculect	'monkey (animal)'	'monkey (zodiac sign)'
North Ryukyuan.	Izena	saːru	saru
	Yonamine	saːruː	saru
	Shuri	saːru	saru
South Ryukyuan	Irabu-Nakachi	saru	saɨ
	Tarama	ʃaːru	ʃaɭ
	Ishigaki	sarɨ	sarɨ
	Taketomi	sarakka	saru
	Iriomote-Sonai	sa̤ro	sa̤ru
	Hateruma	saːru	sa̤ri

[3] Data taken from the following sources. Izena: Izena-jima hogen jiten henshu iinkai (2004), Yonamine: Nakasone (1983), Shuri: National Language Research Institute (1963), Irabu-nakachi: Tomihama (2013), Tarama: Tokuyama and Celik (2020), Ishigaki: Miyagi (2003), Taketomi: Maeara (2011), Iriomote-Sonai: Maeo (2002), Hateruma: Celik et al. (2023). Word-forms are shown in standardized orthography.

The implementation of our cognate ID system makes it easy to merge two cognate sets into one while the reverse, that is to say splitting an existing cognate set into two, is quite impractical. That is why we have adopted a somewhat conservative approach when defining cognate sets, so that when in doubt, two separate cognate sets are set up.

3.2.2 Cognate ID system

At this stage of the research, the most accurate way to link each record in the lexical data sets to its respective cognate set is to do it by hand.[4] That is, for each record, a cognacy judgement is made and the relevant cognate ID is inputted accordingly. That is why we need to design a system of cognate IDs which is fully compatible with the practical and efficient implementation of this linking task.

One solution would be to use the cognate labels of JR-COGNATES, which serve as a unique identifier for each cognate set. The labels are composed of the ideogrammatic orthography of the original meaning of the cognate set followed by the katakana orthography of the reconstructed form (for example "頭（アタマ）" for the cognate *atama* 'head'). However, these labels are not practical for the linking task because they are not fully predictable from the lexical data. Indeed, in order to retrieve the cognate label for a particular word-form, one needs to know the meaning that has been deemed original, the ideogram that has been chosen for this meaning, and the form that has been reconstructed. This information cannot however be fully deduced from the lexical data itself, so that one would need to memorize all the IDs to be able to complete the linking task.

For the present framework, we adopted a refined version of the label system used in JR-COGNATES which solves the predictability problem of the IDs. That is, we designed a system of cognate IDs in which any ID can be retrieved based on the lexical data to be linked. The IDs themselves consist of 4 parts separated by a dot: the ideogram of the meaning, the beginning segment of the proto-form, the abbreviation of the part of speech of the proto-form ('n': noun, 'v': verb, 'a': adjectival stem, 'o': other), and a disambiguating number if needed (Figure 3). Concerning the beginning segment of the proto-form, 'u' and 'i' are used for 'back vowel of mid or close height' and 'front vowel of mid or close height' since the reconstruction of vowel height in word-initial position is not always straightforward in pRk (4). Otherwise, it follows the orthography adopted for the reconstructed forms (see section 4)

[4] Although an algorithm and a Python package has been developed to automatically group lexical data which share the same semantic description into cognate sets (List et al. 2018), it turns out to be somewhat unpractical when applied to Japanese lexical data. This is because of the tremendous orthographic variations found across the sources in the semantic description of entries. For example, a word given in a source with the semantic description 'アダン (Pandanus odoratissimus)', is never grouped with its cognate which is given in another source the same description but orthographized as '阿檀'. This wide-ranging variation in orthography observed in Japanese lexicographic sources partly defeats the purpose of the algorithm. The results might however be improved by first applying on the semantic description an automatic morphemic analysis using for example UniDic (Den et al. 2007).

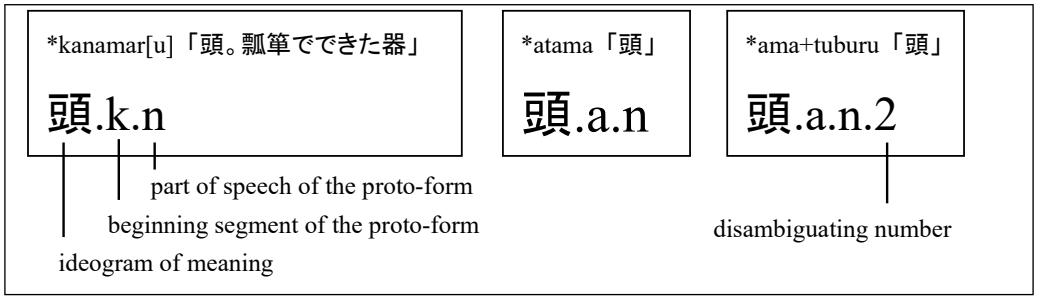

Figure 3. Cognate ID design (cognate IDs for the words meaning 'head')

(4) Example of IDs for cognates beginning with mid or close vowel
a) 打.u.v *ut- 'to make hit'
b) 落.u.v *ote- 'to fall'
c) 胞衣.i.n *ira 'placenta'
d) 色.i.n *ero 'color'

We then constructed two sets of IDs[5]: a set of working IDs (作業ID) specifically designed for the linking task, and a set of representative IDs (代表ID). The latter is unique per cognate set and is basically equivalent to JR-COGNATES labels, with the original meaning of the cognate set chosen for the ideogrammatic representation of the ID. For most of the cognates already listed in JR-COGNATES, the original meaning and its ideogrammatic representation have been inherited as such. Working IDs are built in the same way as representative IDs but their meaning part is not limited to the original meaning of the cognate or to one ideogrammatic orthography. It means that for each cognate set, there are one or more working IDs and these stand in a many-to-one relationship to the representative IDs. Working IDs are added to make it possible to complete the linking task without prior knowledge of the original meaning or the chosen orthography. For example, the reflexes of the adjectival stem **kata-* may have the meaning of 'hard', 'strong (of tea)', or 'thick (with plants)'. Additionally, the meaning of 'hard' can be written with two different ideograms. That is why four different working IDs are provided just for this one cognate (5). Introducing these working IDs enables to guess the correct ID whatever the meaning attested in the doculects.

(5) Working IDs for **kata-* 'hard, strong, thick'
a) 堅.k.a.2
b) 固.k.a.2
c) 濃.k.a
d) 密.k.a

This design proved very efficient when doing cognacy judgement and linking lexical records by hand. Provided that the annotator has some notions of pRk's correspondences with the doculect being annotated, this design enables to retrieve the working ID from the lexical data without having to memorize all of the IDs.

[5] We also provide a number ID for a lighter implementation when working with an online database.

3.3 The cognate table

The cognate table contains detailed annotations about each of the cognate set defined by the cognate IDs. Its data structure is described in Table 2 (Nb: Number, INT: Integer, STR: String, obl.: obligatory, (obl.): obligatory but implementation not completed, opt.: optional, PoS: part of speech, pJ: proto-Japanese).

Table 2. Cognate table structure ("CogList" table)

Nb	Column	Type	Status	Description
1	Number ID	INT	obl.	Number ID of the cognate set
2	Representative ID	STR	obl.	Representative ID of the cognate set
3	proto-form	STR	obl.	Reconstructed proto-form
4	pre-proto-form	STR	opt.	If any, pre-proto-form
5–8	Morphological part(s)	STR	(obl.)	Representative ID(s) of the morphological part(s)
9	Simplified PoS	STR	obl.	Simplified PoS of the cognate set
10	PoS	STR	obl.	PoS of the cognate set
11	PoS ID	INT	obl.	PoS ID of the cognate set
12	Mora count	INT	obl.	Mora count of the cognate set
13	Conjugation class	STR	opt.	Conjugation class of verbs
14	Stem-final segment	STR	opt.	Stem-final segment of verbs
15	Semantic category	STR	opt.	Semantic category of the cognate set
16	Lexical statrum	STR	opt.	Lexical stratum of the cognate set
17	Status in pRk	STR	(obl.)	Status ('goes back to pRk or not') of the cognate
18	Cognacy type with Jap.	STR	opt.	Type of cognacy correspondence with Jap.
19	Japanese cognate	STR	opt.	If any, cognate in Japanese
20	Meaning	STR	obl.	Meaning(s) expressed in Modern Japanese
21	Meaning notes	STR	opt.	Notes concerning the meaning of the cognate
22	Etymological notes	STR	opt.	Etymological notes about the cognate
23	Tone class (pRk)	STR	opt.	Reconstructed tone class in pRk
24	Tone class (pJ)	STR	opt.	Reconstructed tone class in proto-Japanese
25	Evidence for proto-form	STR	opt.	Evidence on which is based the proto-form
26	Evidence for tone class	STR	opt.	Evidence on which is based the reconstructed tone class

Since cognate sets are added to the table without a-priori judgment concerning their status in pRk, we must encode for whether each cognate goes back to pRk or not. This is the information contained in column 17 "Status in pRk". This means that some of the reconstructed forms in the table should not be taken as pRk forms but rather as the oldest hypothetical form that can be reconstructed given its distribution in the Ryukyuan languages and this has some implications for the orthography used (see section 4). Note that in the case of the more recent lexical innovations, the reconstructions given may well be completely fictitious. That is why it is imperative that the reconstructed form be always considered together with its status in pRk. In the future, we hope to add the status information for all stages of the proto-language (proto-Sakishima, proto-Miyako, proto-Yaeyama etc.).

In the meaning column, following JR-COGNATES, we provide the list of all the meanings attested in the reflexes, so, for example, the meanings listed for **kamati* discussed above are 'cheek', 'jaw', 'head', 'doorframe'. Because of the way this column is defined, it has to be constantly updated as more and more comparative data gets integrated. When listing meanings, the

meaning deemed original should be given as first, but this feature is not fully implemented yet at this stage.

We have adopted a broad morphological analysis for each cognate. The morphological structure is directly shown in the reconstructed form ('-': suffix boundary, '+': compound boundary, '=': clitic boundary, '~': reduplication), but we have also introduced fields containing the cognate ID of the different morphological parts. This design enables us to encode the relationship between partial cognates.

See the supplementary materials for further information on the values of the different fields. The relationship between working IDs and representative IDs is contained in the table "IDs".

3.4 Meta-information on sources and doculects

As UniCog is designed to be implemented on any kind of lexical data, we have compiled a prospective list of lexical sources on the Ryukyuan languages and a list of all the doculects contained therein (see supplementary data). The sources include published articles, dictionaries, as well as fieldwork notes. For some of these sources, beginning with the authors' fieldwork notes, we have already been inputting the working IDs on an experimental basis, and we were able to verify that the annotation task was both feasible with minimum knowledge, and relatively efficient.

4. Phonological and prosodic model of proto-forms

In this section, we briefly describe the phonological model adopted for the reconstruction of the word-forms of pRk (corresponding to column 3 and column 23 of the cognate table). We mainly follow the reconstruction of Hattori (1978–1979), Thorpe (1983), Pellard (forthcoming) with the exception of the few points explained below. Note that the reconstruction of proto-forms is merely a working hypothesis, never a definite answer. The phonological model we adopt here is specifically designed to investigate the comparative data of the Ryukyuan languages and has at this stage of the research no other purpose. That is why we have decided to adopt by design highly conservative, and likely controversial, hypotheses on some facets of pRk's phonological model.

4.1 Vowels

We adopt the five-vowel model (*a, *i, *u, *e, *o) for our reconstruction (Table 3). This system is similar to Modern Japanese and (Northern) Ryukyuan dialects, but does not directly correspond diachronically, like for instance **medu* 'water' and **kusori* 'medicine' (Hattori 1976, 1978–1979, Thorpe 1983, Pellard 2008, 2013, 2015, forthcoming). It is quite possible that the raising of *e and *o had already begun at the stage of pRk, at least phonetically, but we do not explicitly mark this feature.

Table 3. Basic vowels of proto-Ryukyuan

	front		back
high	*i		*u
mid	*e		*o
low		*a	

4.2 Consonants

As elucidated in the previous literature, we posit the following thirteen consonants: *p, *b, *m, *t, *d, *s, *z, *r, *n, *k, *g, *w, *j for pRk (Thorpe 1983, Pellard forthcoming) (Table 4).

Table 4. Consonants of proto-Ryukyuan

	labial	coronal	dorsal
plosive	*p, *b	*t, *d	*k, *g
fricative		*s, *z	
liquid		*r	
nasal	*m	*n	
approximant	*w	*j	

Semivowels *w and *j may form a cluster with the other consonants (clusters are not included in Table 2). Following Thorpe, we postulate the phonological cluster *tj in pRk (Thorpe 1983: 51–52), like in the word for 'tea' reconstructed as **tja*. It is difficult to assess the exact phonetic realization of *tj at the level of pRk and we remain agnostic as to whether it was realized as [tja] or had already become an affricate as [tʃa] (we are also agnostic about the phonetic realization of *ti, *tu, *di, *du). However, since the reflex of *tj corresponds to an affricate in most present languages, we decided to adopt the orthography *cj for representing that sound.

We also need to introduce the phonological clusters *sj and *zj for the reconstructed forms, although it is a matter of debate whether they already existed in pRk. The former, *sj, is found for example in **sj[o]* 'master'. The latter, *zj, has to be introduced because of the distinction between *z and *zj found in some Northern Ryukyuan languages such as the Shuri dialect spoken by noblemen (e.g. *dzaa* 'seat; post', *adza* 'nevus' versus *dʒaa* 'snake', *dʒootuu* < **zjautou* 'excellent' (National Language Research Institute 1963)). Since most of the words for which *zj is reconstructed belong to the Sino-Japanese vocabulary stratum, they may well be borrowings postdating pRk. Notwithstanding, *zj is still needed because we provide a reconstructed form for all the cognate sets, including those not going back to pRk.

In addition, since we adopt the distinction among yotsu-gana as discussed in section 4.3.2, it is necessary to reconstruct *dj as well. *dj is basically only reconstructed in words belonging to the Sino-Japanese stratum, and the source for its reconstruction is based almost entirely on the Old Japanese On-yomi. When no cognate is found in Japanese, such as the word **ma[d]j[o]ni* 'together', the reconstructed sound is represented by [d], meaning either *z or *d (see section 4.4 on the use of brackets).

As *cj, *sj, *zj, *dj can be analyzed as a complex cluster composed of a consonant and a glide, we do not include them in Table 4.

We believe that there is convincing evidence for pre-nasalization of voiced plosives in proto-Ryukyuan (Martin 1987:21–22, Nakazawa and Yokoyama 2021a, 2021b), but we leave out this feature unmarked in our orthographic representation since it bears no relevance to the basic phonological contrasts of pRk.

4.3 Differences with previous literature

Although we follow the previous literature (see above) for the bulk of the phonological model of pRk, there are slight differences that we discuss below. For the reasons explained above, all these

differences involve the adoption of hypotheses more conservative than the previous literature.

4.3.1 Reconstruction of intervocalic *p

Most reconstructions of pRk assume that proto-Japonic intervocalic *p had already lenited, partially or fully, in pRk. The evidence for that claim is however not completely straightforward and we can point out to inconsistencies in the treatment of intervocalic *p. For example, Thorpe (1983) reconstructs **utawi* 'song' with a partially lenited *p but posits **masio* < ***masipo* 'salt' and **wototoi* < ***wototopi* 'the day before yesterday' with fully lenited *p (that is without any trace of *p). Moreover, in the cases where intervocalic *p is retained in the modern dialects, Thorpe mechanically reconstructs a geminate pronunciation *Qp for pRk. The facts are however not as straightforward as it seems. For example, the two adjectival stems for 'soft' (below 'soft 1' and 'soft 2') do not show the same reflexes between them although they share the same root (Table 5). In Miyako, 'soft 1' retains intervocalic *p while 'soft 2' does not. In contrast, intervocalic *p of 'soft 2' is retained in Northern Ryukyuan dialects. What value should we then reconstruct for these two stems? As Thorpe points out, the retention of intervocalic *p is mostly observed in adjectival roots (Thorpe 1983: 61), in which emphatic forms (hence the geminate) may have contributed to the retention of *p. It is however not limited to them. We can indeed point out to Miyako *sipis-* 'get deaf' whose root is cognate with Middle Japanese *sifi-* (癈ひ) 'lose sensation or function of body organs' or even to pRk **apa* 'reef stonefish' whose intervocalic *p is retained in all the modern dialects (emphatic pronunciation of 'reed stonefish' does not seem very likely).

Table 5. Reflexes of 'soft 1' and 'soft 2' (-: no cognate found)[6]

Doculect	'soft 1'	'soft 2'
Ie	-	japara:sa
Shuri	-	jaɸarasaɴ
Hirara	japa-	ja:ra-
Tarama	japa-	ja:ra-
Ishigaki	jaɸa-	ja:ra- ~ jaɸara
Yonaguni	-	daraɴ

In the light of these caveats concerning the lenition of intervocalic *p at the level of pRk, we adopt a purposefully conservative hypothesis, namely that *p was retained at the level of pRk. Since it is not always possible to decide between intervocalic *p or *w using just the data on Ryukyuan, we use when available the data on Japanese as an external criterion.

4.3.2 Reconstruction of yotsu-gana *zu, *du, *zi, *di

No dialect of Ryukyuan is known to fully retain the proto-Japonic distinction called "yotsu-gana" in Japanese and facetiously referred to as "the kana gang of four" (Martin 1987). This distinction involves *zu vs. *du and *zi vs. *di. There are however two pieces of evidence, one indirect and one direct, perhaps for reconstructing the distinction in pRk. If we follow Igarashi's new phylogeny of

[6] Data from the following sources. Ie: Oshio (2009), Shuri: National Language Research Institute (1963), Hirara: Jarosz (2015b), Tarama: Tokuyama and Celik (2020), Ishigaki: Miyagi (2003), Yonaguni: Yonaguni hōgen jiten henshū iinkai (2021).

the Japonic languages (Igarashi 2021), in which Ryukyu and Kyushu dialects form a single clade, it must be the case that proto-Kyushu-Ryukyuan retained the distinction of *zu, *du, *zi, *di since the distinction is found in many dialects of Kyushu (Oto (1695), Kyushu Hogen gakkai (1969), Sugimura (2001) among others). pRk cannot be equated with proto-Kyushu-Ryukyuan but the two languages must have been very close, so that the retention of the four-way kana distinction in pRk is a plausible hypothesis. The other piece of evidence regards a puzzling alternation found in Miyako. In many dialects of Miyako, 'take off (one's clothes)' is expressed by the verb *paddz-*, cognate with Shuri *hadʒijuɴ* and which is built on the same root as Japanese *hazure-* < *fadure-* 'come off' (note the reversed transitivity of *paddz-* and *hazure-* and the difference in conjugation class). In some other dialects of Miyako, however, we find the related form *paduk-* < **padok-*. The two forms *paddz-* and *paduk-* are clearly related and the latter is clearly an innovation. Indeed, cognates of *paddz-* are widely distributed in both Northern and Southern Ryukyuan dialects, enabling to safely reconstruct the word in pRk. Meanwhile, **padok-* is restricted to a few dialects of Miyako and must therefore be the result of a lexical innovation postdating the split of Miyako from the other Ryukyuan languages. We cannot, however, explain the form **padok-* if we stick to standard model of pRk according to which *du and *zu had already merged. That is, if, following the standard model, we reconstruct **pazur-* in pMk, we would expect the form **pazok-*, which is not attested. To explain **padok-*, we must reconstruct a plosive (i.d. **padur-*) at the level of proto-Miyako and assume that the merger of *du and *zu is a later innovation.

Admittedly, these are thin pieces of evidence. Still, we decided to adopt, again on purpose, the (very) conservative hypothesis according to which *zu, *du, *zi, *di were distinguished in pRk. We believe that adopting this hypothesis might help uncover facts that would otherwise be left overlooked if an innovative, and hence less informative hypothesis were adopted in the first place. Of course, as in the case of intervocalic *p, we are forced to use Japanese data as an external criterion to decide between *zu vs. *du and *zi vs. *di.

4.4 Indeterminacy of segments in the reconstruction

It is often the case that the comparative data is not informative enough to determine the full segmental make-up of the proto-form. For example, if a form beginning with [ɸ] ~ [f] is only distributed in Southern Ryukyuan (e.g. cognates of Tarama *fuʃe:k-* 'be luxuriant'), there is no telling from the phonological correspondence whether the consonant goes back to *k or *p. To handle this kind of indeterminacy, we devised a simple orthographic system that encodes the suboptimal informative level of the comparative data (see Table 6).

We show some examples of the use of each of these symbols in Table 7.

Table 6. Symbols for the indeterminate status of reconstructed segments

Symbol	Meaning
( )	present or absent
C	Consonant of unknown value
V	Vowel of unknown value
[u]	high vowel: *i or *u
[e]	front vowel: *i or *e
[o]	back rounded vowel: *u or *o
[d]	voiced coronal: *d or *z
[p]	labial consonant: *p and *w
[w]	labial consonant: *w or *b
[k]	voiceless velar of bilabial plosive:*k or *p

Table 7. Example of indeterminate reconstructions[7]

Reconstruction	Meaning	Notes
*ku(C)a	'child'	see Celik (2022a)
*t[u]na-	'young'	cf. Ishigaki *tsɨnasaːɴ*, found only in Yaeyama so that 1st syllable may be either *tu or *ti.
*p[e]p[e]za	'goat'	Height of reflex vowel in 1st and 2nd syllables incongruent across dialects.
mo[d][u]re-	'become presbyopic'	cf. Tarama *muddʒi*-. Found only in Miyako so that 2nd syllable may be either of *du, *di, *zu, *zi.
*tukana[p]-	'raise'	Since no Japanese cognate is found, we cannot decide between *p or *w.
*[w]ik-	'win (in a bet)'	cf. Tarama *bɨk*-. Found only in Miyako so that first syllable may be either *wi or *bi.

4.5 Prosodic system

Concerning the tonal system of pRk, we follow the ternary tonal class model first laid out by Hattori and later expanded by Matsumori (Hattori 1958, Matsumori 2000a, 2000b, 2012). That is, we postulate the three tonal classes A, B, and C in pRk. The symbols used for the reconstruction of the proto-tone are shown in Table 8. We adopted the symbols first introduced by JR-COGNATES (with the exception of N) and added some further symbols we deemed necessary.

[7] The examples here are given as illustrations. It is well possible that further research may help resolve the indeterminate status of the forms presented.

Table 8. Symbols used for the reconstruction of the tone class

JR-COGNATES	UniCog	Symbol explanation
A, B, C		tone class in pRk
O		Comparative data on tones not sufficient to reconstruct the tone class with confidence
X		Comparative data available but conflicting correspondence across dialects
(no symbol)	F	Cognate not going back to pRk, so that the reconstruction of the tone class is not relevant
	M	Cognate goes back to pRk but unable to reconstruct its tonal class due to its morphemic status (bound forms etc.)
	-	Reconstruction of tone class on hold

5. Examples of the implementation of UniCog

To test the validity of the framework we propose, we have implemented the UniCog framework on an experimental basis in different lexical data sets. In this section, we report on the results of this implementation.

5.1 NINJAL's Database of Endangered Languages of Japan

The Database of Endangered Languages of Japan was released in its first version in 2016 by NINJAL's collaborative research project Endangered Languages and Dialects in Japan and is composed of one lexical database[8] and one database of oral texts.[9] A new version of the lexical database was released in September 2022 which adopted a data structure model derived from the data model elaborated by the first author for the compilation of Ryukyuan-Japanese dictionaries (see Celik and Oura (2022)[10] for details about the data structure). In the last modification of the database released in March 2023, UniCog's cognate ID system was implemented.

With the implementation of the cognate IDs into the Lexical Database of Endangered Languages of Japan, it became possible to show for each entry the list of all its cognates contained in the database (Figure 4). In the cognate table displayed, cognates are organized according to the geographical location of the doculect, from North-East to South-West, thus allowing for a quick comparison of all the forms. To this date (April 2023), the lexical database contains 16,647 entries from 23 doculects of Japonic. So far, 7,982 entries, or 48 percent of the database have been annotated for cognate ID, representing 2,616 different cognate sets.

[8] https://kikigengo.ninjal.ac.jp/data/tango/search

[9] https://kikigengo.ninjal.ac.jp/data/danwa/search

[10] http://doi.org/10.15084/00003679

ʻゆー *ʔjuː* [名]

沖永良部島上平川

0:00 / 0:01

(1) さかな【魚】。

出典

横山晶子（2017）「沖永良部調査ノート」

同源語

方言（仮名）	方言（国際音声記号）	意味	品詞	ア型	地点
よ	jo	さかな【魚】	名		八丈島三根
よ	jo	さかな【魚】	名		八丈島末吉
イオ	jo	さかな【魚】			鹿児島県甑島里
ユ	ju	さかな【魚】	名		喜界荒木
ʻイユ	ʔiju	さかな【魚】	名		喜界阿伝
ʻイユ	ʔiju	さかな【魚】	名		喜界小野津
ユ	ju	さかな【魚】	名		喜界上嘉鉄
ʻイユ	ʔiju	さかな【魚】	名		喜界湾
ʻゆー	ʔjuː	さかな【魚】	名		沖永良部島国頭
ゆʻー	ʔjuː	さかな【魚】	名		与論東区
ʻゆー	ʔjuː	さかな【魚】	名		沖縄伊平屋村田名
っずう	ddzu	さかな【魚】	名		宮古池間西原
イ゜ズう	zzu	さかな【魚】	名		宮古砂川
いじゅ゜	idʒu	さかな【魚】	名	LO	宮古水納
いゆ	iju	さかな【魚】	名		与那国祖納

Figure 4. Page for the entry *ʔjuː* 'fish' of Okinoerabu Kami-hirakawa with display of cognate list (Screenshot taken 16th April 2023, Japanese page)

By adding the cognate table as another lexical source into the database, and using the potential of regular expressions, it will become possible to execute sophisticated searches directly through pRk. The other merit of implementing UniCog into the Lexical Database of Endangered Languages of Japan is that the transparency of the data is warranted. Indeed, this database not only displays the source of the data, but it also contains the meta-information of the sound recordings accompanying each entry (most entries in the database are provided with sound materials). This way, it is possible to check the accuracy of every piece of data integrated through the Unicog framework, including for example word-form transcription or tone class.

5.2 Dynamic compilation of an etymological dictionary of the Ryukyuan languages

One of the long-term research objectives of the first author is the compilation of an exhaustive etymological dictionary of the Ryukyuan languages. By applying the UniCog framework into existing lexical data, it becomes easy to compile such a dictionary on the fly, applying various filters to create the needed output. The compilation is done in 2 steps. The first involves compiling the data from the various files of lexical data into a single LaTeX file. This step is taken using a Python program. The second step is the compilation of the LaTeX file into a pdf file following a pre-defined dictionary template. We show an example of the resulting output in Figure 5, in

***aCare-?**［動］［慌.a.v.2］
慌てる（アワテル）。
［再建］八重山 a:ri-。

[ʔa:ruŋ ^{A} ［自動］（鳩間）
慌てる。急ぐ。騒ぐ。暴れる。『鳩間方言辞典』

***aCar(e)-as-(?)**［動］［慌.a.v.3］
騒がせる（サワガセル）。

a:rasuɴ［動］（波照間北）
騒がす。騒がしくする『波照間方言辞典（仮）』

***aC[o]da**［名］［畚.a.n］琉（X）類（o）
篭輿（オウダ）。畚（モッコ）。運搬用網（ウンパンヨウアミ）。

o:da［名］（与那覇）
（草などを入れる）容器の一種『与那覇調査ノート』
o:da $^{F\#}$ ［名］（皆愛）
（草などを入れる）容器の一種『皆愛調査ノート』
avda［名］（砂川）
もっこ『砂川調査ノート』
auda［名］（友利）
容器の一種『友利調査ノート』
auda［名］（新城）
もっこ『新城調査ノート』
o:da［名］（鏡原山中）
（草などを入れる）容器の一種『山中調査ノート』
auda ^{m1} ［名］（仲筋）
畚『南琉球宮古語多良間方言辞典』
o:da ^{m1} ［名］（仲筋）
畚『南琉球宮古語多良間方言辞典』
a:da ^{m1} ［名］（水納）
もっこ『みんなふつ』
o:da ^{m1} ［名］（水納）
畚『みんなふつ』
auda［名］（川平）
もっこ『川平調査ノート』
[ʔau] da ^{BC} ［名］（鳩間）
もっこ（畚）。藁縄を網状に編んで、その四隅に吊り紐をつけ、芋や大根、南瓜などの農産物　または肥料や土などを運搬するのに用いる農具。「あうだ（篭輿）」、「Auoda. アヲダ（籠）. 担架に似た一種の簡易寝台で～」『邦訳日葡辞書』の転訛したものか。　アイ　ク［ ai ku］(おふご {EOS}「枕、和名阿布古＜あふこ＞、杖名也」『和名抄』の転訛) の前方と後方に吊るして農産物を運ぶ農具。前後一対の畚に入れた荷を、プ　スカタ　ミ［pu sukata mi］(一荷) という。一方だけでは、カ　タ　ティー［ka ta ti］(片手) といい、前方の荷が重いことを、　マイ　ニー［ mai ni ］(前荷) という。『鳩間方言辞典』
[ʔauda ^{BC} ［名］（古見）
もっこ（畚）。藁縄を網状に編んで、四隅に吊り紐をつけ、芋やその他の農産物、肥料を運搬する用具。『続古見方言の基礎語彙（3）』(158)
auta［名］（新城下地）
もっこ『新城下地調査ノート』
onda［名］（波照間北）
もっこ『波照間方言辞典（仮）』
uda ^{C} ［名詞］（与那国）
モッコ。網籠。馬に乗せる『どぅなんむぬい辞典』

***aC[o]da?**［名］［蛙.a.n］
蛙（カエル）。

auda［名］（川平）
蛙『川平調査ノート』
ottə ^{BC} （竹富）
蛙『竹富島方言アクセント（2）』
[ʔau] ta ^{BC} ［名］（鳩間）
(動) カエル (蛙)。濃緑の体色をもつ小型の蛙もいるが名称は不明。　アウタ　ヌ　ナ　ク　カー　ア　ミ　ヌ　フーン［ auta nu na ku ka　a mi nu　u ŋ］(蛙が鳴いたら雨が降る) 蛙を食べる習慣はなかったが、煎じ薬として食することはあった『鳩間方言

Figure 5. Sample of an etymological dictionary of (Southern) Ryukyuan dynamically compiled

which we have compiled data from Southern Ryukyuan sources annotated with cognate IDs (mainly fieldwork notes of the authors but also some published articles and the dictionaries published by NINJAL). Incidentally, the tone class for **aC[o]da* 'rope basket' is still marked as X in our cognate table, but one look at the Southern Ryukyuan comparative data shown in Figure 5 suffices to realize that this word can safely be reconstructed as belonging to tone class C in proto-Sakishima.

Note that, in a way, the annotation of cognate IDs is very crude. That is, no additional information besides the cognacy of each lexical entry is encoded. That is why the following cautionary note is in order concerning the comparative data compiled with the UniCog framework. It is often the case that in very large lexical sources several words are recorded for a given concept, but only one of them would be considered as the basic term in the said doculect. For example, the basic term for 'head' in Miyako Tarama is *kanamaɭ*, but Tokuyama and Celik (2020) also list *tsɿburu*, which is clearly a loanword and has an additional semantic nuance of 'intelligence' (the reflex inherited from pRk **tuburu* is *tsɿbuɭ*, showing regular sound correspondences and referring to 'calabash'). However, since Tarama *tsɿburu* does indeed belong to the cognate set of pRk **tuburu*, it is annotated accordingly and ends up listed in the comparative data (Figure 6). This implies that, even assuming perfect accuracy in the cognacy judgements, the user still has to make a valid assessment of the comparative data.

***tuburu**［名］［頭.t.n］琉（C）類（x）
頭（アタマ）。夕顔（ユウガオ）。瓢箪（ヒョウタン）

tsɿburu［名］（与那覇）
頭脳『与那覇調査ノート』
tsɿguɳ［名］（鏡原山中）
植物の一種。夕顔『山中調査ノート』
tsɿbuɭ m1［名］（仲筋）
ひょうたん『南琉球宮古語多良間方言辞典』
tsɿburu m1［名］（仲筋）
頭。知識。知恵『南琉球宮古語多良間方言辞典』
tsʁburu m1［名］（水納）
頭脳『みんなふつ』

Figure 6. Sample of an etymological dictionary of (Southern) Ryukyuan (**tuburu* entry)

6. Standardized orthography of the Ryukyuan languages for comparative purposes

As discussed in section 5, implementing UniCog on existing lexical data renders possible the compilation of comparative data, the first step towards reconstructing proto-Ryukyuan. However, even if we succeeded in integrating all the existing lexical data on the Ryukyuan languages, we would still be impeded in our comparative purposes by the diversity of orthographies and transcription practices found in the sources. Indeed, depending on the source, word-forms are transcribed in a very narrow phonetic notation (for example Kibe ed. (2012)), while others adopt an idiosyncratic orthography based on a specific phonological analysis of the doculect under inves-

tigation. As a result, word-forms taken from different sources, even when documenting the same doculect, cannot readily be compared, as evidenced in (6).

(6) Transcription of the word meaning 'to walk' as reported in the main lexical sources of Miyako Tarama
a) aʃkˢï (Hirayama 1983)
b) aʃkï (Takahashi 1993)
d) alkɨ (Shimoji 2017)
f) a⌊kɭ (Tokuyama and Celik 2020)

In order to be able to make valid historical comparisons between word-forms, for example by producing alignment data, it is therefore necessary to first standardize the transcription of word-forms. In this section, we propose a standardized phonological orthography for all the Ryukyuan languages, which is both specifically designed for comparison between word-forms and based on the International Phonetic Alphabet (IPA). Below, we explain the broad principles of this orthography.

6.1 Vowels

6.1.1 Central vowels

High central vowels, or sounds acoustically close to it, are found in dialects of both Northern and Southern Ryukyuan. These sounds have been variously transcribed as [ï] (Miyanaga (1980), Jarosz (2015a), Hirayama (1983, 1986, 1988), Uemura (1959), Miyagi (2003), Tomihama (2013), Honda (2021) etc.), as [ɨ] (Ryudai Hogen Kenkyu Kurabu (1969), Nakamatsu (1987), Kato (2022), Nakama et al. (2022) etc.), or, in the case of Miyako and some dialects of Yaeyama, as [ɿ] (Karimata (2005, 2008), Kibe ed. (2012), Celik (2022b) etc.). Although the same symbol [ï] [ɨ] has been used for both Northern and Southern Ryukyuan, these sounds are not equivalent. From a diachronic perspective, Northern Ryukyuan high central vowel goes back to pRk *e while the sound transcribed as [ï][ɨ][ɿ] in Southern Ryukyuan goes back to pRk *i or *u (Thorpe 1983). From a synchronic perspective also, the Southern Ryukyuan sound transcribed as [ï][ɨ][ɿ] should not be analyzed as a central vowel (see for instance Aoi (2012)). It also exhibits phonetic characteristics like spirantization not found in Northern Ryukyuan dialects. For these reasons, we believe it is not optimal in a comparative framework to use the same transcription for Northern and Southern Ryukyuan. The choice of symbol is admittedly arbitrary, but we propose the following transcription practice (7)(8).

(7) The central high vowel derived from *e in Northern Ryukyuan is written as [ï], and the corresponding mid vowel as [ë].
(8) The so-called 'central' high vowel derived from *i or *u in Southern Ryukyuan is written as [ɨ], and the corresponding mid vowel as [ɘ].

6.1.2 Low vowels

The low vowel is written as [a], but the marked low vowel derived from the long vowel is written as [ɑ] to distinguish it from [a]. This rule is needed to handle the two different 'a' sounds found in Southern Ryukyuan Kuroshima and Taketomi.

6.1.3 Long vowels
Length of vowels is written as [ː].

6.2 Consonants
6.2.1 Palatal consonants
No dialect of Ryukyuan has been found to make a phonological distinction between post alveolar [ʃ], [ʒ] and alveolo-palatal [ɕ], [ʑ]. Although for some doculects [ɕ], [ʑ] is a better phonetic match of the actual pronunciation of palatal fricatives, we decided for two reasons to uniformly use [ʃ] and [ʒ]. First, these symbols are more common, and second, they are visually more easily recognized.

Palatalized [n] is written as [ɲ]. However, if there is no distinction between [ni] and [ɲi], it is written as [ni].

6.2.2 Fricatives and affricates
If there is no opposition between voiced alveolar fricatives [z], [ʒ] and voiced alveolar affricates [dz], [dʒ], we use the latter voiced alveolar affricate symbol whatever the exact phonetic value (affricate vs. fricative) of the consonants in question. The reason for that is that in the dialects that do distinguish between affricates and fricatives, fricatives are derived from approximants, while affricates correspond to voiced alveolar affricates/fricatives of other dialects.

6.2.3 Glottalized and aspirated consonants
In dialects that have glottalized (or unaspirated) consonants and thus make a phonological distinction between glottalized and aspirated consonants, glottalized consonants are written as [C$^{\text{ʔ}}$a] and aspirated as [C$^{\text{h}}$a]. In the case of dialects where this phonological distinction is only active at the beginning of words, corresponding consonants within words are written [Ca]. In the case of dialects exhibiting phonetic but not phonemic glottalized consonants, these are written as geminate consonants (e.g. Irabu [tta] ‘came’).

6.2.4 Moraic nasal
If moraic nasals are not phonemically distinctive at the place of articulation, they are represented by the uvular nasal [ɴ] regardless of the actual place of articulation.

6.2.5 Notation of glottal stop
In dialects without distinctive glottal stop, phonetic [ʔ] is omitted. In dialects that have phonological [ʔ] (presence or absence of [ʔ] is distinctive), the phonological distinction involving the glottal stop is written as follows: wutu ‘husband’ / ʔutu ‘sound’, jiɴ ‘relationship’ / ʔiɴ ‘dog’, ɴni ‘breast’ / ʔɴni ‘rice plant’.

6.2.6 Labial fricative
The bilabial and labiodental fricatives [ɸ] [f] are widely found in Ryukyuan languages. They have many sources and variants and it is often hard to decide which symbol more accurately reflects the main phonetic value for a specific doculect (no variety of Ryukyuan is known to possess a phonological contrast between [ɸ] and [f]). Accordingly, the choice of symbol has been made based on which of the two symbols has traditionally been used in the literature on a specific doculect.

6.2.7 Distinction between [ɸu] and [hu]

In dialects where the phonological distinction between [ɸu] and [hu] is absent, we use [ɸu] which is a closer match to the labialized phonetic realization of this syllable[11]. For dialects which distinguish between [ɸu] (or [fu]) and [hu], we write [ɸu] (or [fu]) and [hu] (marked) separately.

6.2.8 Others

Labiodental approximant [ʋ] is not used instead of [w]. Notation of devoicing is omitted except for some dialects of Yaeyama, in which devoicing seems to be a phonological feature (e.g. Hateruma, Shiraho, Aragusuku, Kabira).

6.3 Notation of tone class

Concerning tone, we need to distinguish between the tone class to which the word belongs and the pitch realization of the citation form. Tone classes are specific to each doculect but they can be standardized by referring to their correspondence with the tone classes of pRk. That is, we simply use the letters of the reconstructed tone classes in pRk, A, B, and C (see section 4.5 for the prosodic model of pRk). In the case where tone classes have merged, we show the correspondence by using the letters for both the merged categories: AB, BC, AC[12]. In the case of dialects like Miyako Hirara where all tonal classes have merged, we leave the information of tone class empty. As for the pitch realization of the citation form, we use " [" and "] " to indicate the locus of pitch rise or fall. Examples are given in Table 9.

Table 9. Tone class correspondence notation

Doculect	Word-form	Tone corr. notation	Meaning
Ie	tʰiti]dʒa	AC	'goat'
Shuri	ɸi:dʒa:	BC	'goat'
Tarama	pinda	C	'goat'
Ishigaki	[pi]bidʒa	A	'goat'
Yonaguni	hibida	C	'goat'

In the case of verbs, only two tonal classes are reconstructed in pRk (Matsumori 2012)[13]. One of them corresponds to tone class A, but the other corresponds to tone class B or C depending on the dialect and the conjugation form (for a discussion of B and C tone alternations found in the conjugation of verbs in some of the Southern Ryukyuan dialects, see Uwano (2011), Celik (2020), Nakazawa (2023)). We therefore use the symbols A and ¬A (i.e. 'not A') to show the tone correspondences in verbs. Again, no tone correspondence is assigned in the case of doculects without tonal distinctions (Table 10).

[11] Diachronically, this syllable may come from pRk *pu, *po, *ku, *ko.

[12] In most cases, this notation can be equated with the notation of the synchronic tone classes. However, in some specific cases, like for example Hateruma in which one tone class split into two classes through tonogenesis (Aso and Ogawa 2016), this notation is not equivalent to the notation of the synchronic tone class.

[13] Matsumori (2012) reconstructs only two tone classes for adjectival roots also, but we need to posit three classes for adjectival roots in Miyako (Igarashi et al. 2018, Celik 2020), necessitating the use of all three labels A, B, C.

Table 10. Tone class correspondence notation (verb)

Doculect	Word-form	Tone corr. notation	Meaning
Ie	matʃu]ɴ	A	‘wind’
Tarama	makɨ	A	‘wind’
Ishigaki	[ma]kuɴ	A	‘wind’
Ie	matʃu[ɴ	¬A	‘sow’
Tarama	makɨ	¬A	‘sow’
Ishigaki	[makuɴ	¬A	‘sow’

6.4 Examples

In Table 11, we give some examples of the standardized orthography.

Table 11. Examples of the standardized orthography (‘-’: no cognate found in the source)[14,15]

Doculect	**kapa* A ‘well’	**uma* B ‘horse’	**woto* A ‘husband’	**ijo* A ‘fish’	**pune* C ‘boat’	**tume* A ‘nail’
Asama	[kʰo:	[ʔma:	[wutu: A	[ʔju:	ɸu:[nï C	[tsʔïmï: A
Ie	ha]: AC	ʔma[: B	wutu] AC	ʔju]: AC	kʰuni] AC	simi] AC
Nakijin	[ha: A	ʔma[: B	wu[tu: A	[ʔju: A	[pʰu]ni C	tʃʔi[mi: A
Shuri	ka: A	ʔɴma BC	wutu A	ʔiju A	ɸuni B	tsimi A
Ikema	ka: B	-	butu A	ddzu B	funi C	tsɨmi B
Tarama	ka: A	mma ‘horse (zodiac)’	butu A	zzu A	funi C	tsɨmi A
Ishigaki	[ka]: A	[ɴma BC	[bu]du A	[ʔi]dzu A	[ɸuni BC	[tsɨ]mi A
Taketomi	kɑ: A	ɴɴma	butu	idʒu A	ɸuni BC	sumi A
Hateruma	ke: A	ɴmaɴ BC	butu A	ju: A	fu̥ni BC	ʃi̥mi A
Yonaguni	kʰa A	ɴma B	butu A	iju A	ɴni C	ɴmi A

7. Further research possibilities and applications

As described in this paper, the UniCog framework is primarily designed for the compilation of comparative data with the specific aim of achieving a detailed reconstruction of proto-Ryukyuan. As a framework linking lexical materials through cognacy, it can easily be extended in the future to cover the whole of the Japonic family. At the same time, the use of this framework is not limited to the reconstruction of the proto-language. Since cognates are linked across different sets of lexical data whatever their status in pRk, we can use for example the information on the morphological structure of the cognate sets to investigate inside one doculect the relationship between the tone class of compound nouns and that of its constituents.

There are also practical applications of this framework. Provided that they are annotated with the cognate IDs, it becomes possible to subtract the entries of two different lexical sources and retrieve the list of entries not shared by them. In the case of relatively close dialects, we can thus use the biggest available lexical source and subtract from it the lexical source on the target dialect to create lexical questionnaires.

[14] Data taken from the following sources. Asama: Uwano (2017a, 2017b), Ie: Oshio (2009), Nakijin: Nakasone (1983), Shuri: National Language Research Institute (1963), Ikema: Nakama et al. (2022), Hirara: Jarosz (2015b), Tarama: Tokuyama and Celik (2020), Ishigaki: Miyagi (2003), Taketomi: Maeara (2011), Hateruma: Celik et al. (2023), Yonaguni: Yonaguni hōgen jiten henshū iinkai (2021).

[15] In the Asama dialect, the tonal distinctions seem to have been lost in the case of two-syllable words that underwent mono-syllabification through phonological change (‘well’, ‘horse’, ‘fish’).

References

Aoi, Hayato (2012) The phonetic interpretation of the "Central Vowel" in the Tarama Variety of Miyako Ryukyuan. *Gengo Kenkyu* 142: 77–94.

Aso, Reiko, Kenan Celik and Kohei Nakazawa (2022) Possibilities of meta-research on Japonic descriptive linguistics: Evaluation of 40 years of lexical research on Southern Ryukyuan and its implications for future methodology. *NINJAL Research Papers* 23: 75–98.

Aso, Reiko and Shinji Ogawa (2016) A three-pattern accent system in Hateruma Ryukyuan. *Gengo Kenkyu* 150: 87–115.

Celik, Kenan (2020) *Minami Ryukyu Miyakogo-shi*. PhD Thesis. Kyoto: Kyoto University.

Celik, Kenan (2022a) Jodai-nihongo no ko-rui o_1 ni taio suru Ryukyu-sogo no mo hitotsu no ontaio ni tsuite. *The Society for Japanese Linguistics 2022 Spring Conference proceedings*, 115–120.

Celik, Kenan (2022b) Miyakogo Uruka hogen no goishu. *Gengo Kijutsu Ronshu* 14: 157–209.

Celik, Kenan and Tatsuo Oura (2022) Digital data of the glossary of the Minna dialect (31st March 2022 edition). Tokyo: The National Institute for Japanese Language and Linguistics, Language Variation Division.

Celik, Kenan, Reiko Aso and Kohei Nakazawa (2021) Meiji-ki no Yaeyamago no goishiryo "Kainan shoto tango-hen". *Gengo Kijutsu Ronshu* 13: 139–177.

Celik, Kenan, Reiko Aso and Kohei Nakazawa (2023) Minami Ryukyu Yaeyamago Hateruma hogen jiten ni kansuru chukan hokoku. *Gengo Kijutsu Ronshu* 15: 193–358.

Den, Yasuharu, Toshinobu Ogiso, Hideki Ogura, Atsushi Yamada, Nobuaki Minematsu, Kiyotaka Uchimoto and Hanae Koiso (2007) The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics. *Japanese Linguistics* 22: 101–123.

Hattori, Shiro (1958) Amami gunto no shohogen ni tsuite: Okinawa, Sakishima shohogen to no hikaku. Reprinted in: Shiro Hattori (1959) *Nihongo no keito*, 275–294. Tokyo: Iwanami Shoten.

Hattori, Shiro (1976) Ryukyu hogen to hondo hogen. In: Iha Fuyu seitan hyakunen kinenkai (ed.) *Okinawagaku no reimei: Iha Fuyu seiten hyakunen kinenshi*, 7–55. Okinawa: Okinawa Bunka Kyokai.

Hattori, Shiro (1978–1979) Nihon sogo ni tsuite 1–22. *Gengo* 7(1)–7(3), 7(6)–8(12).

Higa, Matsuyoshi and Yoriko Takaesu (2021) *Okinawa Tonaki hogen jiten*. Tonaki: Tonaki mura.

Hirayama, Teruo (1983) *Ryukyu Miyako Shoto hogen kiso goi no sogoteki kenkyu*. Tokyo: Kadokawa Shoten.

Hirayama, Teruo (1986) *Amami hogen kiso goi no kenkyu*. Tokyo: Ofusha.

Hirayama, Teruo (1988) *Minami Ryukyu no hogen kiso goi*. Tokyo: Ofusha.

Hirayama, Teruo, Ichiro Oshima, Makio Ono, Makoto Kuno, Mariko Kuno and Takao Sugimura (ed.) (1992–1994) *Gendai Nihongo hogen daijiten*. Tokyo: Meiji Shoin.

Hirayama, Teruo, Ichiro Oshima and Masachie Nakamoto (1967) *Ryukyu Sakishima hogen no sogoteki kenkyu*. Tokyo: Meiji Shoin.

Honda, Hirotaka (2021) *Tokunoshima kotoba jiten ko (Jo-Ge): Ino kawa chushin*. Haebaru: Kyodo Bunka Kenkyukai.

Igarashi, Yosuke (2016) Akusento-gata no taio ni motozuite nichiryu-sogo wo saiken suru tame no goi-risuto "Nichiryu ruibetsu goi". *The Society for Japanese Linguistics 2016 Autumn Conference proceedings*, 233–238.

Igarashi, Yosuke (2019) Nichiryugo ruibetsu goi v.7 (2019/05/17) digital data. https://researchmap.jp/multidatabases/multidatabase_contents/detail/238738/86b4e49faf0f27bff29c1cf84fc6bc37?frame_id=729332 (last accessed 2023/03/29).

Igarashi, Yosuke (2021) Bunkigakuteki shuho ni motozuita nichiryu-shogo no keito-bunrui no kokoromi. In: Tomohide Kinuhata and Yuka Hayashi (eds.) *Fiirudo to bunken kara miru Nichiryu shogo no keito to rekishi*, 17–51. Tokyo: Kaitakusha.

Igarashi, Yosuke, Yukinori Takubo, Yuka Hayashi, Thomas Pellard and Tomoyuki Kubo (2012) The Ikema Dialect of Miyako Ryukyuan has a three-, not two-, pattern accent system. *Onsei Kenkyu* 16(1): 134–148.

Igarashi, Yosuke, Yukinori Takubo, Yuka Hayashi and Tomoyuki Kubo (2018). Tonal neutralization in the Ikema dialect of Miyako Ryukyuan. In: Haruo Kubozono and Mikio Giriko (eds.) *Tonal change and*

neutralization, 83–128. Berlin: Walter de Gruyter.
Izena-jima hogen jiten henshu iinkai (2004) *Izena-jima hogen jiten*. Izena: Izena village board of education.
Jarosz, Aleksandra (2015a) *Nikolay Nevskiy's Miyakoan dictionary: Reconstruction from the manuscript and its ethnolinguistic analysis. Studies on the manuscript*. Ph.D. thesis. Poznań: Adam Mickiewicz University.
Jarosz, Aleksandra (2015b) *Nikolay Nevskiy's Miyakoan dictionary: Reconstruction from the manuscript and its ethnolinguistic analysis. The reconstructed dictionary*. Ph.D. thesis. Poznań: Adam Mickiewicz University.
Kajiku, Shin'ichi (2020) *Hatoma hogen jiten*. Tokyo: NINJAL Language Variation Division.
Karimata, Shigehisa (2005) Okinawa-ken Miyako-jima Hirara hogen no foneemu. *Nihon Toyo Bunka Ronshu* 11: 67–113.
Karimata, Shigehisa (2008) *Ryukyu Yaeyama hogen no hikaku rekishi hogengaku ni kan suru kisoteki kenkyu. Heisei 17 nendo - 19 nendo kagaku kenkyuhi hojokin (kiban kenkyu (C)) kenkyuseika hokokusho*. Research report.
Kato, Kanji (2022) Kagoshima-ken Tokunoshima Isencho. In: Kenan Celik, Nobuko Kibe, Yosuke Igarashi, Hayato Aoi and Hajime Oshima (eds.) *Nihon no shometsu kiki gengo hogen no bunpo kijutsu*, 335–362. Tokyo: NINJAL Language Variation Division.
Kibe, Nobuko (ed.) (2012) NINJAL Collaborative Research Project Reports 12-02. General research for the study and conservation of endangered dialects in Japan. Research Report on the Miyako Dialects of Southern Ryukyuan. Tokyo: NINJAL Language Variation Division.
Kiku, Chiyo and Toshizo Takahashi (2005) *Yoron hogen jiten*. Tokyo: Musashino Shoin.
Kindaichi, Haruhiko (1974) *Kokugo akusento no shiteki kenkyu: genri to hoho*. Tokyo: Shima Shobo.
Kinuhata, Tomohide (2021) Ryukyu Sogo to Jodai Nihongo kara mita sogo no shiji taikei shiron. In: Tomohide Kinuhata and Yuka Hayashi (eds.) *Fiirudo to bunken kara miru Nichiryu shogo no keito to rekishi*, 190–213. Tokyo: Kaitakusha.
Kyushu Hogen Gakkai (1969) *Kyushu Hogen no kisoteki kenkyu*. Tokyo: Kazama Shobo.
List, Johann-Mattis, Simon Greenhill, Tiago Tresoldi and Robert Forkel (2018) *LingPy. A Python library for quantitative tasks in historical linguistics*. Jena: Max Planck Institute for the Science of Human History.
Maeara, Toru (2011) *Taketomi hogen jiten*. Ishigaki: Nanzansha.
Maeo, Yoan (2002) *Iriomote hogenshu*. Taketomi: private edition.
Martin, Samuel E. (1987) *The Japanese language through time*. New Haven: Yale University Press.
Matsumori, Akiko (2000a) The development of word lists for Ryukyuan accent research, Based on the dialects of Okinoerabu Island. *Onsei Kenkyu* 4(1): 61–71.
Matsumori, Akiko (2000b) An examination of so-called "multi-patterned" accent systems in Ryukyuan dialects, focusing on three-syllable words. *Kokugogaku* 51(1): 93–108.
Matsumori, Akiko (2010) Tarama no san-kei akusento to keiretsubetsu goi. In: Zendo Uwano (ed.) *Nihongo no kenkyu no 12 sho*, 490–503. Tokyo: Meiji Shoin.
Matsumori, Akiko (2012) Toward a categorized vocabulary for Ryukyuan field research. *Onsei Kenkyu* 16(1): 30–40.
Miyagi, Shin'yu (2003) *Ishigaki hogen jiten*. Naha: Okinawa Taimususha.
Miyanaga, Masamori (1980) Miyanaga Masamori zenshu 8. Tokyo: Daiichi Shobo.
Nakama, Hiroyuki, Yukinori Takubo, Shoichi Iwasaki, Yosuke Igarashi and Natsuko Nakagawa (2022) *Minami Ryukyu Miyakogo Ikema hogen jiten*. Tokyo: NINJAL Language Variation Division.
Nakamatsu, Takeo (1987) *Ryukyu hogen jiten*. Haebaru: Naha Shuppansha.
Nakasone, Seizen (1983) *Nakijin hogen jiten: Nakijin hogen no kenkyu goi-hen*. Tokyo: Kadokawa Shoten.
Nakazawa, Kohei (2023) Diachronic study on the tonal alternation in the verbal paradigm of Southern Ryukyuan Yonaguni Dialect. *NINJAL Research Papers* 24: 169–194.
Nakazawa, Kohei and Akiko Yokoyama (2021a) Stop series in Japonic. In: Hiroyuki Suzuki and Mitsuaki Endo (eds.) *Studies in Asian and African Geolinguistics I "Stop series"*, 30–31. Tokyo: Research Institute for Languages and Cultures of Asia and Africa (ILCAA) Tokyo University of Foreign Studies.
Nakazawa, Kohei and Akiko Yokoyama (2021b) Interesting sounds and sound changes in Japonic. In: Hiroyuki Suzuki and Mitsuaki Endo (eds.) *Studies in Asian and African Geolinguistics I "Stop series"*, 93–101. Tokyo: Research Institute for Languages and Cultures of Asia and Africa (ILCAA) Tokyo University of Foreign Studies.

National Language Research Institute (1963) Okinawago jiten. Tokyo: Okurasho Insatsu-kyoku.
Oshio, Mutsuko (2009) *Okinawa Ie-jima hogen jiten*. Ie: Ie village board of education.
Oto, Sokufu (1695) *Shi chi su tsu kana moji zukai kenshuku ryoko shu*. Kyoto: Iseya Kiyobei.
Pellard, Thomas (2008) Proto-Japonic *e and *o in eastern old Japanese. *Cahiers de linguistique Asie Orientale* 37(2): 133–158.
Pellard, Thomas. (2009) *Ogami: Éléments de description d'un parler du Sud des Ryukyu*. PhD thesis. Paris: École des hautes études en sciences sociales.
Pellard, Thomas (2013) Ryukyuan perspectives on the proto-Japonic vowel system. In: Bjarke Frellesvig and Peter Sells (eds.) *Japanese/Korean Linguistics 20*, 81–96. CSLI Publications.
Pellard, Thomas. (2015) The linguistic archeology of the Ryukyu Islands. In: Patrick Heinrich, Shinsho Miyara and Michinori Shimoji (eds.) *Handbook of the Ryukyuan languages: History, structure, and use*, 13–37. Berlin: De Gruyter Mouton.
Pellard, Thomas (forthcoming) Ryukyuan and the reconstruction of proto-Japanese-Ryukyuan. In: Bjarke Frellesvig, Satoshi Kinsui and John Whitman (eds.) *Handbook of Japanese historical linguistics*. Berlin: De Gruyter Mouton.
Ryudai Hogen Kenkyu Kurabu (1969) Kohama hogen. *Ryukyu Hogen* 9–10: 37–91.
Sakamoto, Kiyoe, Kazue Akinaga, Kazuaki Ueno, Eisaku Sato and Yutaka Suzuki (eds.) (1998) Waseda gorui Kindaichi gorui taisho shiryo. *Akusento-shi shiryo sakuin* 13: 225–290
Shimoji, Kayoko (2017). *Tarama futsu jiten*. Tarama: Tarama Village Board of Education.
Shogakutosho (1989) *Nihon hogen daijiten*. Tokyo: Shogakukan.
Sugimura, Takao (2001) The "yotsugana" phonemic system in Kyushu dialect. *Journal of the Phonetic Society of Japan* 5(3): l–18.
Takahashi, Toshizo (1993) Tarama hogen no goi (chukan hokoku). *Tarama-jima chosa hokokusho (1)*, 73–164. Ginowan: Okinawa Kokusai Daigaku Nanto Bunka Kenkyujo.
Tashiro, Antei (1888a) *Tokyo daigaku rigaku toshokan shozo Tashiro Antei shiryo*. Unpublished report. The University of Tokyo Science Library. https://iiif.dl.itc.u-tokyo.ac.jp/repo/s/tashiro/document (last accessed 2023/07/28).
Tashiro, Antei (1888b) Okinawa kenka Miyako-jima oyobi Okinawa-jima taiyaku hogenshu. *Tokyo Jinrui Gakkai Zasshi* 3(29): 323–328.
Tashiro, Antei (1888c) *Miyako-go Naha-go*. Manuscript. National Taiwan University Library. https://dl.lib.ntu.edu.tw/s/Tashiro/item/714623 (last accessed 2023/07/28).
Tashiro, Antei (1888d) *Okinawa-ken torishirabe fuzu*. Manuscript. The University of Tokyo Science Library. https://iiif.dl.itc.u-tokyo.ac.jp/repo/s/tashiro/document/b495504c-3487-1d58-210f-ad5f795286a2 (last accessed 2023/07/28).
Tashiro, Antei (1894) Yaeyama gunto jumin no gengo oyobi shukyo. *Tokyo Jinrui Gakkai Zasshi* 9(96): 229–232.
Thorpe, M. Lawrence (1983) *Ryukyuan language history*. Ph.D. thesis, University of Southern California.
Tokuyama, Shunei and Kenan Celik (2020) *Minami Ryukyu Miyakogo Tarama hogen jiten*. Tokyo: NINJAL Language Variation Division.
Tomihama, Sadayoshi (2013) *Miyako Irabu hogen jiten*. Naha: Okinawa Times.
Uechi, Norio (2021) *Sarahama hogen goi jiten*. Naha: Border Ink.
Uemura, Yukio (1959) Ryukyu shohogen ni okeru "1, 2 onsetsu meishi" no akusento no gaikan. *Kotoba no Kenkyu* 1: 121–140.
Uwano, Zendo (2011) Examples of conjugated forms of verbs in the Yonaguni dialect with particular reference to prosodemes: Part 2. *NINJAL Research Papers* 2: 135–164.
Uwano, Zendo (2017a) Accent data from the Asama dialect in Tokunoshima, Amami: Part 3. *NINJAL Research Papers* 12: 139–161.
Uwano, Zendo (2017b) Accent data from the Asama dialect in Tokunoshima, Amami: Part 4. *NINJAL Research Papers* 13: 290–242.
Yonaguni Hogen Jiten Henshu Iinkai (2021) *Dunan munui jiten daini han*. Ishigaki: Nanzansha.

琉球祖語の再建に向けた比較データ構築用の枠組提案（UniCog）

セリック・ケナン [a]　中澤光平 [b]　麻生玲子 [c]

[a] 国立国語研究所 研究系

[b] 信州大学

[c] 名桜大学

要旨

本稿では，琉球祖語の再建に向けて，琉球諸語の比較データを動的に構築できる枠組みを提案する。UniCog（**Uni**fied **Cog**nacy Framework for proto-Ryukyuan）と呼ぶこの枠組みは，琉球諸語に分布する約 7,400 語の同源語リストを伴っており，その中核には琉球諸語を対象とする既存の全ての語彙データを紐付けるための同源語 ID システムがある。本枠組みを記述した後，具体例を示しながら，この枠組みの実装によって開かれる研究可能性について述べる。最後に，語形の比較を目的とする，琉球諸語の統一的な表記も提案する。この枠組みの導入が日琉諸語の歴史比較言語学の分野においていくらかの貢献を果たすことが期待される。

キーワード：琉球祖語，語彙研究，歴史比較言語学，同源語ID