# Taokas-10 revisited: Taokas or Atayal?

**Abstract**

This paper reexamines the genetic affiliation of Taokas-10 using the comparative method. Taokas-10 is a wordlist of a Formosan language recorded near Miaoli, Taiwan in the beginning of the twentieth century by the Japanese linguist Naoyoshi Ogawa. It was assumed to be a dialect of Taokas due to its geographic position, and although later researchers noted its resemblance to another Formosan language—Atayal,—no evidence has been presented to justify this claim. This paper presents a thorough examination of its phonology, lexicon, and parts of its morphosyntax. The evidence is clear that the language in the Taokas-10 dataset was in fact a dialect of Atayal, with some lexical borrowings from neighboring Formosan languages (Saisiyat, Taokas, other Atayal dialects). Although it is most closely related to Matu' uwal (Mayrinax) Atayal, it still shows peculiarities in both phonology and lexicon. It was likely a remnant of a larger Atayal population living in the lowlands of Miaoli that was later assimilated by Hakka Chinese, who now dominate the region.

## 1 Introduction

This paper examines the data of an ostensibly Taokas dialect in Tsuchida (1982), with the goal of demonstrating that it is in fact an Atayal dialect closely related to Matu'uwal Atayal.[1] The dataset in question was given the number 10 in Tsuchida' s comparative vocabulary, and is thus called Taokas-10 in this paper. Taokas-10 has unique features in its phonology, lexicon, and even morphosyntax, that point towards a closer relationship with Matu' uwal Atayal specifically.

[1]Matu' uwal, an Atayal dialect spoken in Tai' an township, Miaoli county, has been alternatively known in linguistic literature by its exonym "Mayrinax."

Both Taokas and Atayal belong to the Austronesian language family. Taokas is no longer spoken, as the ethnic Taokas have linguistically assimilated with the surrounding Han Chinese. It used to be prevalent around the lowlands of modern Hsinchu and Miaoli counties. Atayal is an extant language, spoken in the Central Mountain Range in the northern half of Taiwan, including the mountain areas directly inland of territories historically occupied by Taokas. Other languages spoken in the area include Saisiyat (Austronesian), as well as Taiwanese Southern Min and Hakka (Sinitic).

The Taokas-10 data in this paper was sourced from Tsuchida (1982), a comparative vocabulary of the now extinct Western Plains languages of Taiwan. The data was collected by Japanese researchers at the end of the 19th and the beginning of the 20th century. It includes the languages Taokas, Babuza, Papora, Hoanya, and Pazih, all belonging to the Austronesian language family. Data for each language comes from several sources, collected by Japanese linguists and anthropologists in different villages, and Tsuchida' s work puts all of them together in an easily comparable format. Tsuchida gave each dataset a primary number (e.g. Taokas-10), and also a secondary number to datasets from the same village that were obtained by different researchers (e.g. Babuza-1-3). Some datasets also include a letter to indicate that data was collected by a single researcher from different speakers in the same village (e.g. Taokas-3-2b).

Tsuchida sourced the Taokas-10 vocabulary from one of Naoyoshi Ogawa' s notebooks entitled "Taokas," which includes both Ogawa' s and Ino Kanori' s field notes. The Taokas data was collected from at least 1897 to 1917, based on the dates provided by Tsuchida. Although grouped together with Taokas, Taokas-10 shares almost no similarities with the datasets Taokas-1 through Taokas-9, and demonstrably represents an entirely different language.

The language consultant for Taokas-10 was 林金生 (Lin Jin-sheng in Mandarin), a male of unspecified age. Ogawa elicited the data from him in 1901 in a place called Biaukoh-sia (苗閣社), which according to Abe (1938: 162) is an amalgamation of two village names: Biaulik-sia (苗栗社) and Kachikoh-sia (加志閣社). Both were situated in what is now Miaoli county.

When Tsuchida was including Taokas-10 data in his comparative vocabulary, he knew it was different from his other sources. He was familiar with Atayal, and had done fieldwork on several dialects of Atayal, including Matu'uwal Atayal. He noted the resemblance of Taokas-10 to the latter specifically, but did not elaborate beyond providing several examples of lexical correspondences. He still placed it

together with Taokas 'temporarily,' presumably until the actual nature of this language could be resolved.

The sources of the data used in the paper are presented in section 2. Section 3 discusses the methodology. Sections 4, 5, 6 compare Taokas-10 with other Austronesian languages in the area from lexical, phonological, and morphological perspectives, respectively. Section 7 contains a description of phonological and lexical phenomena that are unique to Taokas-10, and a discussion of what this entails.

## 2 Data sources

Taokas-10 data is sourced from Tsuchida (1982), who in turn used Ogawa's field notes. The Atayal data comes mainly from my own field notes, cross-checked against Egerod (1980) and Tesing Silan (2003) for Squliq Atayal, Li (1981) for S'uli' Atayal, and Li (2004) for Matu'uwal Atayal. Saisiyat data for this paper comes from Ferrell's (1969) comparative vocabulary of Formosan languages, and Li's (1978) more comprehensive Saisiyat dialect vocabulary.

Proto-Austronesian (PAn) reconstructions are taken from Blust and Trussel's Austronesian Comparative Dictionary, available online (Blust and Trussel). Proto-Atayal reconstructions are my own (Goderich 2020), but do not contradict Proto-Atayalic reconstructions by Li (1981).[2]

## 3 Methodology

This paper will first examine the possible relation of Taokas-10 to Austronesian languages spoken in the area, beginning with lexical evidence, then moving on to phonological evidence, and finally looking at parts of its morphosyntax.

For lexical evidence, we look at the amount of cognates between Taokas-10 and a given language. The Taokas-10 wordlist contains a total of 198 unique lexical items. 68 of these words are found in the Swadesh list of 100 basic vocabulary items (Swadesh 1971: 283). The percentage of shared cognates from the basic vocabulary list can be used to determine the nature of a genetic relationship between languages. The language with the highest amount of basic vocabulary cognates with Taokas-10 will be the most closely related to it.

[2]Proto-Atayalic is the ancestor of Proto-Atayal and Proto-Seediq.

Once a likely candidate has been established, we look at the phonological evidence, making sure that the sound correspondences between Taokas-10 and its closest relative are systematic. Systematic sound correspondences greatly decrease the possibility of lexical borrowing or chance resemblance leading to similarity between the languages, and strengthen the genetic relationship hypothesis. Phonological evidence may also help establish a genetic relationship, especially in cases where lexical evidence is inconclusive.

Lastly, we inspect the morphosyntax of Taokas-10 for additional evidence. Since the dataset is a wordlist, we cannot expect to find many morphosyntactic features. There is a limited amount of verbal morphology as well as a few phrases that hint at nominal case marking.

# 4 Lexical evidence for a phylogenetic relationship

In this section, we examine the lexical evidence for a close genetic relationship between Taokas-10 and other Austronesian languages spoken in the area around Biaukoh-sia, where the data was collected. We will look at the possibility of Taokas-10 being an aberrant Taokas dialect, a Saisiyat dialect, or an Atayal dialect. Within Atayal, we will compare it with Matu' uwal Atayal specifically, and also with Squliq and S' uli' , all of which are spoken in Miaoli county. The statistics and a preliminary conclusion based on lexical evidence are given in section 4.5.

## 4.1 A comparison of Taokas-10 and Taokas vocabulary

There are only a few Taokas words in the Taokas-10 dataset, and almost all of them have doublets of Atayal origin. These are presented in table 1, with the set marked 'T-10 (1)'being words of Atayal origin, and the set 'T-10 (2)'being words of Taokas origin for the same lemmata. Matu' uwal Atayal is given for comparison with set 1, and two Taokas dialects from Tsuchida (1982) are compared with set 2.

Table 1: Doublets of Atayal and Taokas origin in Taokas-10

| Matu' uwal | T-10 (1) | T-10 (2) | Taokas-3-2a | Taokas-5-2 | Gloss |
|---|---|---|---|---|---|
| yabaʔ | yava | tapu | tāpu | tapu | 'father' |
| yayaʔ | yaya | taai | taai | taai | 'mother' |

| Matu' uwal | T-10 (1) | T-10 (2) | Taokas-3-2a | Taokas-5-2 | Gloss |
|---|---|---|---|---|---|
| quwaw | kuwau | yakau | yakao | yakau | 'alcohol' |
| bawwak | wauwa | kwakwa | kwakwar | kwakwa | 'pig' |
| – | – | walan | yabalan | baran | 'cat' |
| wayluŋ | wailung | tsūtsu | toktor | toxtui | 'chicken' |
| habaŋan | havangan | linlin | lilin | linlin | 'money' |

For all of the words in table 1 except 'cat,' there are two lexical items in the data: one of Atayal origin, and one of Taokas origin. The Taokas origin of the second set is easy to see when compared with Taokas-3-2a and Taokas-5-2. Note that Taokas data is diverse because much like Ogawa's Taokas-10 speaker, the consultants for Taokas 1-9 were likely heritage speakers, and so their pronunciation varied, sometimes considerably.

All of these words, with perhaps the exception of <tsūtsu>[3] 'chicken,' have clear cognates in Taokas. But as they were presented together with a doublet of Atayal origin (except <walan> 'cat'), they were quite clearly learned alongside the Atayal vocabulary, and did not replace it.

A few Taokas-10 words were listed by Tsuchida as Taokas cognates, but are likely of Matu' uwal origin. These are listed in table 2.

Table 2: Dubious Taokas cognates in Taokas-10

| Matu' uwal | Taokas-10 | Taokas-3-2b | Taokas-5-2 | Gloss |
|---|---|---|---|---|
| mamaqisuʔ | taisu | tanasu | tana | 'nine' |
| caŋiyaʔ | sanina | salina | sarina | 'ear' |
| ʔutiq | yutek | yatak | yata | 'earth' |
| balayan | walayan | – | burayan | 'cooking pan' |

<Taisu> 'nine' has an unexpected /q/ deletion word-medially. This happens in a few other places in the dataset (see section 5.3 for discussion). Initial <ta-> is also more reminiscent of Taokas, and could be due to Taokas influence.

Initial /s/ in <sanina> 'ear' is unexpected, as Proto-Atayal *c is otherwise pre-

[3]Throughout this paper, modern forms are written in italics, phonemes are written between slashes, and Ogawa's transcription of Taokas-10 is written in angled brackets.

served as /c/, so this could indeed be a Taokas loan. A medial liquid turning to /n/ happens elsewhere in the dataset, for example <nainin> 'woman' (cf. Matu' uwal *kanayril*). Note that Taokas <salina>/<sarina> and Matu' uwal *caŋiya?* are both reflexes of PAn *Caliŋa 'ear.'

In <yutek> 'earth,' both vowels are distinct from Taokas, but is similar to Matu' uwal *ʔutiq*, as /i/ is frequently transcribed as <e> in the data, especially next to <q> (see section 5.2). For the initial /y/, see further discussion in section 7.

Finally, <walayan> 'cooking pan' can be linked to Matu' uwal *balayan* through vowel correspondences, namely penultimate /a/. Moreover, the Taokas forms for this lemma are all very different, and it is missing entirely for most dialects: Taokas-5-2 and Taokas-7 <burayan>, Taokas-6-1 <bulawan>, Taokas-6-4 <marean> and <tsinnu>.

There are only 7 lexical items in Taokas-10 that are unambiguously of Taokas origin, out of a grand total of 198, and 6 of these also have doublets of Atayal origin. There are 4 other words that were considered by Tsuchida to come from Taokas, but at least 3 of them share more similarities with Matu'uwal Atayal. This makes Taokas an extremely unlikely candidate for a close genetic relationship.

## 4.2 A comparison of Taokas-10 and Saisiyat vocabulary

Some words in the Taokas-10 data share similarities with Saisiyat. The forms with the highest likelihood of coming from Saisiyat are listed in table 3.

Table 3: Words of Saisiyat origin in Taokas-10

| Taokas-10 | Saisiyat | Gloss |
|---|---|---|
| kinsaan | kaysaʔan | 'today' |
| taumu | tawmoʔ | 'banana' |
| kayenni | kayniʔ 'don' t' | 'no' |
| gwaa | wæʔæʔ | 'deer' |
| aha | ʔæhæʔ | 'one' |
| shopa, shupa | ʃəpat | 'four' |

Of these, <kinsaan> 'today,' <taumu> 'banana,' and <kayenni> 'no' seem to have no related terms in any Taokas or Atayal dialect at all. Numerals in Taokas-10 do

not appear to come from a single source. In fact, there are two entries for 'one' : <aha> and <kutu>, the former looks much like Saisiyat *ʔæhæʔ* and the latter resembles Matu' uwal and Squliq Atayal *qutux*. The two ways of saying 'four' —<shopa> and <shupa>—are likely the same word said differently, and are more similar to Saisiyat *ʃəpat* than Matu' uwal *sapaat* or Squliq *payat*.

There are several more words that were listed by Tsuchida (1982: 12) as Saisiyat loans, but they may be Matu' uwal cognates. These are listed in table 4.

Table 4: Dubious Saisiyat cognates in Taokas-10

| Taokas-10 | Saisiyat | Matu' uwal | Gloss |
|---|---|---|---|
| vukush | bokəʃ | bukus 'body hair' | 'hair' |
| isutin | iʔtoʃan | ʔisting | 'short' |
| inalu | ʔinaroʔ | qanaruux | 'long' |
| kumita | komitaʔ | mitaal | 'see' |
| huhu | hœhœʔ | xuxuʔ | 'breast' |
| yo | ʔiyok | ʔiyuk | 'orange' |

The form <vukush> is given as 'hair,' but it does not correspond well with either Saisiyat or Atayal. The Saisiyat form *bokəʃ* has an initial voiced bilabial plosive [b] and the final vowel is a schwa. Matu' uwal *bukus* refers to body hair (sound correspondences are given in section 5). It could be either a loan from Saisiyat, a semantic shift in Taokas-10, or perhaps simply a substitution of a forgotten word with a different one, a tactic that Ogawa' s language informant uses several times throughout the dataset.

Tsuchida identified Taokas-10 <isutin> 'short' as a Saisiyat loan, but Saisiyat *iʔtoʃan* is not nearly as good of a match as Matu'uwal *ʔistiŋ*. The vowel in the final syllable is /a/ in Saisiyat but /i/ in Matu' uwal and Taokas-10; and the order of the plosive /t/ and fricative /ʃ/ in Saisiyat is the opposite to that of Matu' uwal and Taokas-10. Taokas-10 <isutin> has a medial <u> where there is no corresponding vowel in Matu' uwal. The vowel is likely epenthetic, inserted to break up a heterosyllabic cluster.

The item <inalu> 'long' is a bit of a conundrum. On the one hand, it is very similar to Saisiyat *ʔinaroʔ*, but on the other hand there's also a possible Matu'uwal cognate: *qanaruux*. The vowel in the first syllable in Matu' uwal is weakened, but

was likely historically /i/, compare S' uli' *ʔinruyux*, Plngawan *ʔinruɯux*. The initial /q/ deletion in Taokas-10 is unexpected, but /q/ was also deleted prevocalically in Taokas-10 <wunaye> 'sand' (cf. Matu' uwal *bunaqiy*)[4] and <taisu> 'nine' (cf. Matu' uwal *mamaqisuʔ*), though it is more often retained as <k>. It is possible that this word in Taokas-10 came from Saisiyat or that it is a Matu' uwal retention.

The Taokas-10 verb <kumita> 'to see' is more similar to Saisiyat *komitaʔ* than to Matu' uwal *mitaal*. Both the Saisiyat and the Matu' uwal forms are reflexes of PAn *kita 'to see.' The Squliq Atayal reflex is *mitaʔ*, with the Actor Voice prefix *m-* appearing to replace root-initial /k/ (cf. the imperative form *kitaʔ*). In actuality, it is the result of infixation followed by the deletion of the first syllable. This process is called "pseudo nasal substitution" by Blust (2004: 76–80). This nasal replacement strategy in Actor Voice forms may be a later development in Squliq and Matu' uwal. If this is the case, the Taokas-10 form, which has an *-um-* infix, would be a retention.

The remaining two words, <huhu> 'breast' and <yo> 'orange, tangerine,' are similar between Saisiyat and Matu' uwal, but at this point it is not clear whether they represent Saisiyat loans into Matu' uwal.

Of the 198 words in the Taokas-10 dataset, 6 come from Saisiyat, and another 6 might be of Saisiyat or Matu' uwal origin. This is a very low number, and Saisiyat must therefore be excluded from the list of potential genetic relationships for Taokas-10.

## 4.3 A comparison of Taokas-10 and Atayal vocabulary

### 4.3.1 General Atayal vocabulary

A large portion of Taokas-10 words are readily identifiable as Atayal, but could in theory come from a number of dialects. Some of these words are exactly the same or very similar in most Atayal dialects. For example, the word *tunux* 'head' (Taokas-10 <tunu>) has the exact same form in all Atayal varieties.

Other lexemes have different forms in various Atayal dialects, but these differences are obscured in Taokas-10 reflexes, presumably due to the speaker's lan-

[4] Matu' uwal *bunaqiy* and Taokas-10 <wunaye> are reflexes of PAn *bunaj 'sand' with a male register infix, compare also Taokas-3-2b <bunat>. See section 5.3 for sound correspondences between Matu' uwal and Taokas-10, and refer to Li (1983) for more information on the derivation of male register forms.

guage attrition. One example is Taokas-10 <laumu> 'blood,' which could be more closely related to either Matu' uwal *ramuux* or Squliq *ramuʔ*, because neither final glottal stops nor final /x/ were preserved in any way by Ogawa' s language consultant.[5] Some Taokas-10 reflexes are more similar in form to Matu' uwal than other dialects, but of themselves do not constitute evidence of a closer relationship. Table 5 presents some of these cognates.[6]

Table 5: Examples of Atayal cognates in Taokas-10

| Taokas-10 | Matu' uwal | Squliq | Gloss |
|---|---|---|---|
| tunu | tunux | tunux | 'head' |
| kava | qabaʔ | qəbaʔ | 'hand' |
| laumu | ramuux | ramuʔ | 'blood' |
| yava | yabaʔ | yabaʔ | 'father' |
| yaya | yayaʔ | yayaʔ | 'mother' |
| kusa | qusiyaʔ | qəsyaʔ | 'water' |
| kavule | qabuliʔ | qəbuliʔ | 'ash' |
| goala | quwalax | qwalax | 'rain' |
| goage | wagiʔ | wagiʔ | 'sun' |
| watunu | batunux | bətunux | 'stone' |
| kahunek | kahuniq | qəhuniq | 'tree' |
| wanek | kabahniq | qəbəhəniq | 'bird' |
| kōle | qulih | qulih | 'fish' |

There are two words that Taokas-10 shares with Matu' uwal and S' uli' , but not Squliq. These are shown in table 6. Squliq forms are not cognate with the other dialects: *raŋiʔ* 'friend' and *baziŋ* 'egg.'

[5]Note that Matu' uwal *ramuux*, Squliq *ramuʔ*, and Taokas-10 <laumu> are all reflexes of PAn *damuq 'blood.' This etymon was replaced in Taokas: Taokas-3-2b <yataxax>, Taokas-7 <taxa> (cf. also Babuza-1-1 <takka>). It is found in the Taokas-10 dataset, which, as discussed here, is not Taokas but a misidentified Atayal dialect.

[6]Matu' uwal and Squliq *wagiʔ* and Taokas-10 <goage> are reflexes of PAn *waRi, compare Taokas-3-2b <yadidax>, Taokas-7 <zizak>, also Pazih *rizax*.

Table 6: Matu' uwal and S' uli' cognates in Taokas-10

| Taokas-10 | Matu' uwal | S' uli' | Gloss |
|---|---|---|---|
| lauin | rawin | rawin | 'friend' |
| watu | batuʔ | batuʔ | 'egg' |

The number of Taokas-10 words that have cognates in multiple Atayal dialects is 79, out of a total of 198 words. This is a much larger proportion than Taokas or Saisiyat cognates, and this number does not even include words that can be linked to a single, specific Atayal dialect. Uniquely Matu' uwal cognates in Taokas-10 are equally numerous, and are discussed below in section 4.3.2.

### 4.3.2 Unique Matu' uwal cognates

Apart from common Atayal lexical items, a number of words in Taokas-10 can be identified as uniquely Matu' uwal, to the exclusion of other Atayal dialects. Some examples are listed in table 7.

Table 7: Examples of unique Matu' uwal words in Taokas-10

| Taokas-10 | Matu' uwal | Squliq | Gloss |
|---|---|---|---|
| huma | həmaʔ | həmaliʔ | 'tongue' |
| kukui | kukuy | kakay | 'leg/foot' |
| lanek | raniq | tuqi | 'road' |
| hamhom | hamhum | yuluŋ | 'cloud' |
| yutek | ʔutiq | rəhyal | 'earth' |
| imu | ʔimug | ŋasal | 'house' |
| yo | ʔiyuk | yutak | 'orange' |
| tikai | tikay | cikuy | 'few' |
| suvangan | sinbaŋan | soriq | 'spear' |
| shātu | siyatuʔ | lukus | 'clothes' |
| ukas | ʔukas | ʔuŋat | 'not have' |
| asehèn | ʔasi hiiŋ | – | 'sweet' |

Some of these are retentions of PAn etyma, for example Taokas-10 <huma>

'tongue' can be linked to Matu' uwal *həma?*, which is a retention of PAn *Sema.[7] Other dialects have appended suffixes to this root: Squliq *həmali?*, S' uli' *hə-ma?uy*. Such suffixation was one of the possible ways to derive male speech register words, but Squliq and S' uli' have lost the gender register distinction (Li 1982).

Other words are unique Matu'uwal innovations, for example Matu'uwal *hamhum* 'cloud, mist,' which is a likely cognate of Taokas-10 <hamhom(le)>. The parenthesized <-le> is not explained, but most likely signifies two possible variants of this word—<hamhom> and <hamhomle>—which is reminiscent of the male-female speech register distinction for which Matu' uwal is famous (Li 1983: 7).

There are also some distinctions in semantics that are peculiar to Matu' uwal: it uses the word *tikay* to mean both 'small' and 'a few,' but other Atayal dialects have separate words for these two concepts. Taokas-10, like Matu' uwal, uses <tikai> for both.

Some words, like Matu' uwal *ʔiyuk* 'orange' and *siyatu?* 'clothes,' may have been borrowed from Saisiyat or Pazih, because these forms are shared between those languages and Matu' uwal, but not other Atayal dialects. These are also included here, because they are present in Matu' uwal, but see also the discussion in section 4.2.

The last two words in table 7, <ukas> 'not have' and <asehèn> 'sweet,' are special because they are not simple content words, but show parts of the morphosyntax of the language. They are discussed in more detail in section 6.

The number of unique Matu' uwal cognates in Taokas-10 is 69. Added to the number of words that can be cognate with any Atayal dialect, we get a grand total of 148 out of 198 lexical items in the dataset. This is the highest number of cognates with Taokas-10 that we can obtain for any language in the region.

### 4.3.3 Possible Matu' uwal cognates

A small number of words have some similarities with Matu' uwal, but with irregular sound correspondences, or in one case, a large shift in meaning (sound correspondences between Taokas-10 and Matu'uwal are discussed in section 5.3). These are presented in table 8.

[7] Compare Taokas-3-2a <tilax>, Taokas-5-1 <telax>, also Babuza-1-1 <tatsira>, Siraya <dadila>, all meaning 'tongue.'

Table 8: Taokas-10 words possibly related to Matu' uwal

| Taokas-10 | Matu' uwal | Gloss |
|---|---|---|
| watsihun | balihun | 'door' |
| kach' uwin | qacuʔ | 'boat' |
| hakali | taktakaliʔ | 'rabbit' |
| mamaa | samamaʔah | 'sour' |
| haka-utu | ʔutux ( 'spirit' ) | 'lightning' |
| mantan | məhantan ( 'night' ) | 'noon' |
| nanu kahani | nanuwan ku hani | 'what' |

The Taokas-10 word <watsihun> 'door' has a highly irregular correspondence with Matu' uwal *balihun*: <ts> to *l*. This is the only occurrence of such a correspondence in the Taokas-10 dataset, and a very unlikely sound change. However, it is reminiscent of the following correspondences in Atayal dialects: Matu' uwal *lalbiŋ*, Skikun Atayal *ləbiŋ*, Klesan Atayal *cəbiŋ*, Plngawan Atayal *cacabiŋ*, all meaning 'sweet.' This aberrant change may be related to the Atayal male-female register system.

The form <kach' uwin> 'boat' is similar to Matu' uwal *qacuʔ* in its first two syllables, with <k> in Taokas-10 regularly corresponding to Matu' uwal /q/, and <ch' > to Matu' uwal /c/. The final syllable in <kach' uwin> may be a derivational suffix of the male register, and in fact *-iŋ* is one of the suffixes used to derive male register forms in Matu' uwal (Li 1983: 4). Both <watsihun> and <kach' uwin> are discussed further in section 7.2.

In <hakali> 'rabbit' there is an irregular correspondence of Taokas-10 <h> to Matu' uwal /t/. Matu' uwal *taktakaliʔ* also has a reduplicated first syllable, but that may be a later innovation.

Taokas-10 <mamaa> 'sour' is missing the first syllable when compared with Matu' uwal *samamaʔah*, but is otherwise regular.

Taokas-10 <haka-utu> 'lightning' does not share any similarity with Matu' uwal *tinaptap na baŋaʔ*, Squliq *məkəlawiʔ*, or S' uli' *təwilak* and *məkayum*, all meaning 'lightning.' The Taokas-10 form appears to be a compound, and is likely related to Matu' uwal *hakaw na ʔutux* 'rainbow,' with a shift in meaning and a dropped genitive case marker *na*. Since the speaker had an imperfect command of his

heritage language, he made both phonological and semantic mistakes in other parts of the dataset.

The entry <mantan> for 'noon' is likely a mistake, since the exact same form is given for the meaning 'night' elsewhere in the dataset. The latter corresponds to Matu' uwal *məhantan* 'night.'

The phrase <nanu kahani> 'what'likely has a more specific meaning 'what is this?' with <nanu> being the question word (cf. Matu' uwal *nanuwan*, Squliq and S' uli' *nanu?* 'what' ), and <kahani> being the deictic. The possible morphosyntactic implications are discussed in section 6.4.

These 7 entries are most likely Matu' uwal cognates, but due to irregularities in their sound correspondences, they were separated from the rest of the Matu' uwal cognate set.

#### 4.3.4 Non-Matu' uwal Atayal influence

A number of words in the Taokas-10 dataset are more similar to Atayal dialects other than Matu' uwal. There are only a few of these words: some likely come from Squliq or S' uli' , while others may not be Atayal cognates at all. These are listed in table 9.

Table 9: Possible non-Matu' uwal Atayal cognates in Taokas-10

| Taokas-10 | Matu' uwal | Squliq | Gloss |
|---|---|---|---|
| pong | magalpug | məpuw | 'ten' |
| kamin | pawmin | kawin | 'eyebrow' |
| wahoe, vahoi | bayhuw | behuy | 'wind' |
| yungai | ʔuŋay | yuŋay | 'monkey' |
| kuluvan | balayan | kəluban | 'cooking pot' |
| wawatun | saqgag | batul | 'earring' |
| maavi | maqilaap | məʔabiʔ | 'sleep' |
| nanu | nanuwan | nanuʔ | 'what' |

The words <pong> 'ten' looks dissimilar enough from both Matu' uwal *magalpug* and Squliq *məpuw* that it may be unrelated. There are also two other entries for 'ten' in Taokas-10: <lampui> and <wampo>. Numerals in Taokas-10 in general do not closely resemble those of Atayal, and are something of a mystery.

The word <kamin> 'eyebrow' resembles Squliq a little more, but the /m/ to /w/ correspondence makes it dubious. The exact same form is also listed for the meaning 'nail,' which adds uncertainty. Taokas-10 <kamin> meaning 'fingernail, toenail' is uncontroversial: compare Matu' uwal *kakamil* 'nail' and Squliq *kəmamil* 'to scratch with fingernails.' The same form as 'eyebrow' may be an erroneous entry.

One of the only words with an obvious Squliq/S' uli' sound correspondence is the double entry <wahoe>/<vahoi> 'wind' —likely the same word pronounced differently several times, unsurprising given that the speaker tends to conflate /b/ and /w/. The Matu' uwal cognate is *bayhuw* with a long /u/ in the final syllable (marked with a glide in the orthography). The Proto-Atayal form had a final *ɹ here, for which the regular reflex in Matu' uwal is ∅ with compensatory lengthening on the preceding vowel. In fact, this is seen in Taokas-10 data: compare Taokas-10 <maliku> and Matu' uwal *mamalikuw* with Squliq *məlikuy* 'man,' or Taokas-10 <taka> and Matu' uwal *taka* 'frog' with Squliq *takay*. Therefore, it is likely that this word was borrowed from Squliq or S' uli' Atayal, which both have the sound change PA *ɹ > *y* in all environments.

Similarly, Taokas-10 <yungai> 'monkey' could be a Squliq or S' uli' borrowing: compare Matu' uwal *ʔuŋay* and Squliq *yuŋay*. However, Taokas-10 may have had its own unique reflexes of word-initial PA *ɹ, the first segment in this word; see section 7 for discussion.

The word <kuluvan> 'cooking pot' is most likely a cognate of Squliq *kəluban* (same meaning), but it is given alongside <walayan>, which is a cognate of Matu' uwal *balayan* 'cooking pot.' In this situation, it appears likely that a Squliq word was borrowed, but did not replace the Matu' uwal word. The wordlist notes that <walayan> is a pot made of iron, while <kuluvan> is a copper pot, so there was a semantic difference between the two.

The last two words, <maavi> 'sleep' and <nanu> 'what,' are not found in Matu' uwal, but exist in most other Atayal dialects. These could be loanwords from Squliq/S' uli' into Taokas-10, or the Matu' uwal forms could be later innovations.

The total number of non-Matu' uwal Atayal cognates in Taokas-10 is 8 (<wahoe>/<vahoi> are counted as a single item). Out of a total of 198, this number is more suggestive of lexical borrowing, or perhaps resemblance due to drift.

## 4.4 Words of other origin

Finally, there are some words in the data that come from other sources, or are of unknown origin. The latter are listed in table 10. These words appear to have no cognates in Atayal, Taokas, or Saisiyat, but it is also possible that none have been found so far, and the meaning may have shifted.

Table 10: Taokas-10 words of unknown origin

| Taokas-10 | Gloss |
|---|---|
| kinale | 'forehead' |
| tsisule | 'body' |
| yarim | 'alcohol' |
| kuima | 'armlet' |
| lumlum | 'star' |
| kuli | 'snake' |
| ivui tsauni | 'come' |
| ivuima keleta | 'speak, say' |
| vuivui | 'fast' |
| sili | 'leopard' |

The word <yarim> 'alcohol' is one of three lexical items for the same lemma, the other two being <yakau>, a Taokas cognate, and <kuwau>, an Atayal cognate. See also the discussion in section 4.1.

There are also two words that appear to have come from Taiwanese Southern Min (TSM): <pana> 'aboriginal people,' possibly from 番仔 *huan-á*, a derogatory term for aboriginal people; and <thi> 'iron,' from 鐵 *thih*. The latter is also used in Matu' uwal: *təhiʔ*, but it is unclear when it was borrowed or why it resembles the TSM pronunciation more than the Hakka *tied* in a region that has historically been dominated by Hakka speakers.

There are 10 words whose origin is so far completely unknown, and 2 words that can ultimately be traced to Taiwanese Southern Min. Surprisingly, there are no Hakka loanwords in the Taokas-10 dataset. One explanation is that the speaker had enough metalinguistic knowledge to separate Hakka vocabulary from Taokas-10, but not to distinguish Atayal, Saisiyat, and Taokas lexical items as coming from different sources.

## 4.5 Statistics

### 4.5.1 Statistics for the dataset as a whole

The overwhelming majority of Taokas-10 words are cognates with Atayal only, because Atayal shares little of its vocabulary with other Austronesian languages (Ferrell 1969: 63–69). Of these, almost half can be unambiguously linked to Matu' uwal, to the exclusion of any other Atayal dialect: some because of unique sound correspondences, and others because they are lexemes unique to Matu' uwal, not shared with other Atayal dialects.

A small number of words are not of Matu' uwal origin. Some of them resemble forms found in Saisiyat, others look like Taokas words, and a few are likely Squliq loanwords. The statistics on Taokas-10 cognacy with different languages are presented in table 11.

Table 11: Taokas-10 lexicon breakdown by origin

| Cognate with: | Number | % of total |
|---|---|---|
| Only Matu' uwal | 69 | **35** |
| Any Atayal dialect | 79 | **40** |
| Possible Matu' uwal | 7 | 3.5 |
| Atayal (excluding Matu' uwal) | 8 | 4 |
| Taokas | 7 | 3.5 |
| Saisiyat | 6 | 3 |
| Taiwanese Southern Min | 2 | 1 |
| Dubious Saisiyat | 6 | 3 |
| Dubious Taokas | 4 | 2 |
| Unknown | 10 | 5 |
| *Total*: | 198 | |

Only unique lexical items were counted in the data. Repeated words, including multiple spellings of the same word, were counted as a single occurrence. The calculations included only single words, and did not include the several phrases that occurred in the dataset. The total number of unique words examined was 198.

Of the total number, 35 percent (69 words) can be identified as uniquely Matu' uwal. About half of these are uniquely shared between Taokas-10 and Matu'

uwal, with no cognates in Squliq and S' uli' . The other half is matched with Matu' uwal based on sound correspondences, which include pretonic vowels, the affricate /c/, the uvular plosive /q/, and the reflexes of the Proto-Atayal retroflex approximant *ɹ.

Another 40 percent (79 words) could potentially be linked to a number of Atayal dialects, because the forms are the same or very similar in various villages. Some of the words in this category are exactly the same across all Atayal varieties (e.g. *ŋarux* 'bear,' compare Taokas-10 <ngalo>), while others appear in several, but not all dialects. Crucially, all 79 forms are found in Matu' uwal, and the sound correspondences are as systematic as in the unique Matu' uwal cognates.

The "possible Matu' uwal" entry includes words that are most likely Matu' uwal, but have irregularities in either sound or meaning. These aberrances prevent them from being included in the Matu' uwal cognate set with certainty.

There is a small number of words (8 words) that have correspondences in Squliq Atayal, but not in Matu' uwal. These are likely to be later loanwords from Squliq, although the time of borrowing cannot be determined. Some of the words in this category may not in fact come from Squliq: for example, the word <kamin> meaning 'eyebrow' (cf. Matu' uwal *pawmin*, Squliq *kawin*) might be simply a mistake, since it has the exact same form as <kamin> 'fingernail' (cf. Matu' uwal *kakamil* 'fingernail,' Squliq *kəmamil* 'to scratch with fingernails' ). The word <yungai> 'monkey' looks more similar to Squliq *yuŋay* than to Matu' uwal *ʔuŋay*, but may in fact reflect a unique sound change in Taokas-10; see section 7.1 for further discussion.

The rows labelled "dubious Saisiyat" and "dubious Taokas" include all lexemes from sections 4.1, 4.2 that have a high likelihood of being of Matu' uwal origin, but may also be connected to Saisiyat or Taokas. Due to the phonological mergers and inconsistencies in the data, it is impossible to link them to either language with certainty. Nevertheless, the likelihood that at least some of these lexemes are more closely related to Matu' uwal is high. The actual number of unique Matu' uwal cognates in the data is probably a bit higher than the table might suggest, and 35% is a conservative number.

The "unknown" words are the 10 lexical items whose origin cannot be established. These were presented in section 4.4.

### 4.5.2 Swadesh list statistics

Looking at the dataset as a whole is insightful, but in order to verify the phylogenetic relationship of a language, historical linguistics uses shared innovations. Shared innovations can appear in any area of language, including the lexicon. Within the lexicon, basic vocabulary items are considered more stable and less prone to change and borrowing (though not completely immune to either). Shared innovations in the basic vocabulary thus serve as the primary indicator of a phylogenetic relationship.

The most commonly used basic vocabulary list is the Swadesh list (Swadesh 1971: 283). Of the 100 items on the Swadesh list, 68 can be found in the Taokas-10 dataset (the word 'one' has both an Atayal and a Saisiyat form, so the total number of lexical items is 69). The numbers are presented in table 12.

Table 12: Taokas-10 Swadesh list percentages

| Cognate with: | Number | % of total |
|---|---|---|
| Only Matu' uwal | 22 | 32 |
| Any Atayal dialect | 38 | 55 |
| Taokas | 1 | 1.5 |
| Saisiyat | 5 | 7 |
| Unsure/unknown | 3 | 4.5 |
| *Total*: | 69 | |

The number of Atayal cognates in the basic vocabulary is even higher than in the dataset as a whole: 32% uniquely Matu' uwal cognates, and another 55% cognate with any Atayal dialect, including Matu' uwal. This patterning is consistent with the hypothesis that Taokas-10 was an Atayal dialect that was influenced by surrounding languages.

One word, the numeral 'one,' has the doublets <kutu> and <aha>, the former being of Atayal origin and the latter coming from Saisiyat: compare Matu' uwal, Squliq *qutux*, Saisiyat *ʔæhæʔ*, all meaning 'one.' Of the 38 Atayal cognates, all but one can be found in Matu' uwal; the exception being <maavi> 'to sleep' : compare Squliq, S' uli' *məʔabiʔ*, Matu' uwal *maqilaap/maqaylup*. Matu' uwal is the only Atayal dialect that does not have a cognate of this form, and it may be a later innovation in this dialect (and thus a retention in Taokas-10).

The total number of basic vocabulary cognates between Taokas-10 and Matu' uwal is thus 60 out of 69, or 87%, which is by far the closest relationship Taokas-10 has with any other language.

# 5 Phonological evidence for a phylogenetic relationship

## 5.1 Atayal phonology

The phonological systems of different Atayal dialects do not differ greatly. The consonant inventory of Matu' uwal Atayal is shown in table 13 as an example of a more conservative phonology. The pronunciations are identical to their IPA values unless indicated otherwise.

Table 13: The consonant inventory of Matu' uwal Atayal

| | | | | | |
|---|---|---|---|---|---|
| voiceless plosives | p | t | k | q | ʔ |
| affricates | | c [t͡s] | | | |
| voiceless fricatives | | s | x | | h [ħ] |
| voiced fricatives | b [β] | | g [ɣ] | | |
| nasals | m | n | ŋ | | |
| laterals | | l | | | |
| rhotics | | r [ɾ] | | | |
| glides | w | y [j] | | | |

The vowel system of Matu' uwal is very simple, with only three cardinal vowel phonemes: /a/, /i/, and /u/.

The main differences between the phonology of Matu' uwal and those of Squliq and S' uli' are the phonological mergers that occurred in the latter two. Both Squliq and S' uli' lack the phoneme /c/, and S' uli' additionally does not have /q/ in its consonant inventory. Some dialects of Squliq have the fricative [ʑ] as a quasi-phoneme. It is the result of palatal glide fortition, and is marginally contrastive in some varieties of Squliq (H.J. Huang 2015).

## 5.2 Tentative phonology of Taokas-10

Before examining the phonological evidence that shows how Taokas-10 is related to Matu' uwal specifically, we need to address the way the data was originally transcribed and how it should be analyzed phonemically. Ogawa' s transcription is assumed to be faithful to his language consultant' s speech, so any inconsistencies are the result of the speaker' s native language influencing the phonology and syntax of his heritage language. Table 14 shows the consonant inventory of Taokas-10 based on Ogawa' s data. In square brackets are the assumed phonetic values of the transcriptions, where they would have differed from the IPA.

Table 14: The consonant inventory of Taokas-10

| | | | | |
|---|---|---|---|---|
| voiceless plosives | p | t | k | |
| affricates | | ts [t͡s], ch' [t͡ʃʰ] | | |
| voiceless fricatives | | s, sh [ʃ] | | h |
| voiced fricatives | v | | g [ɣ~g] | |
| nasals | m | n | ng [ŋ] | |
| laterals | | l | | |
| glides | w | y [j] | | |

The consonant <g> is placed together with <v> in the 'voiced fricatives' row to make the inventory more balanced and match the distribution in other Atayal dialects. The actual pronunciation used by Ogawa' s language consultant is, of course, unknown and unknowable.

Ogawa used a total of 10 vowel symbols in his Taokas-10 transcriptions (<a>, <i>, <u>, <e>, <o>, <ā>, <ō>,<ū>, <è>, <ü>), but it was likely a phonetic rather than phonemic representation of the data.

For example, the mid vowels <e> and <o> in the Taokas-10 data mostly appear where in Matu' uwal there is an adjacent post-dorsal consonant, as demonstrated in table 15.

Table 15: High vowel lowering in Taokas-10

| Taokas-10 | Matu' uwal | Gloss |
|---|---|---|
| lauwe | rawwiq | 'eye' |
| ulake | ʔulaqiʔ | 'child' |

| Taokas-10 | Matu' uwal | Gloss |
|---|---|---|
| hamhom | hamhum | 'cloud' |
| kamhe | qamhit | 'flea' |

The post-dorsal consonants /h/ and /q/ have a lowering effect on adjacent high vowels in Atayal, both when preceding and when following the consonant (Li 1980: 354). Ogawa' s language consultant did not have /q/ as a distinct phoneme (and may have pronounced /h/ as a glottal fricative instead of a pharyngeal one), but he still preserved the lowered vowels in these historical environments.

Some mid vowels occur outside this environment, for example Taokas-10 <yake> and Matu' uwal *yaki?* 'grandmother,' or Taokas-10 <ngalo> and Matu' uwal *ŋarux* 'bear.' Conversely, there are cases where the presence of a /q/ or /h/ phoneme does not trigger lowering in Taokas-10: for example Taokas-10 <kusa> and Matu' uwal *qusiya?* 'water,' or Taokas-10 <tsukuli> and Matu' uwal *cuquliq* 'person' (for correspondences of Matu' uwal /q/ in Taokas, see section 5.3). For the most part, the presence of mid vowels is consistent with the vowel lowering hypothesis.

The five vowels with diacritics appear a total of only seven times in the dataset. All seven lexical items are presented in table 16.

Table 16: Vowels with diacritics in Taokas-10

| Taokas-10 | Matu' uwal | Gloss |
|---|---|---|
| pāyu | payux | 'many' |
| kaā | kaal | 'sky' |
| shātu | siyatuʔ | 'clothes' |
| kōle | qulih | 'fish' |
| tsūtsu | (wayluŋ) | 'chicken' |
| asehèn | ʔasi hiiŋ | 'like honey' |
| tsüla | cəlaq | 'paddy' |

The symbols <ā>, <ō>, and <ū> presumably indicate long vowels. Atayal does not have vowel length as a feature of its phonology. Since this length diacritic occurs on only some vowels, and only in five words in the whole dataset, we will assume that it is not indicative of a phonemic contrast.

The grapheme <è> occurs only once in the data, in the word <asehèn> 'sweet,' which is most likely the Matu' uwal phrase *ʔasi hiiŋ* 'like honey.' It is uncertain what vowel quality Ogawa wanted to represent with the grave accent (possibly, openness). The special vowel quality may be connected to the fact that the word *hiiŋ* 'honey, sugar' has a hiatus with two adjacent high front vowels, preceded by a pharyngeal fricative. The combination of the lowering effect of /h/ with a hiatus or reflex thereof may have given the special vowel quality that Ogawa indicated in his transcription.

Like <è>, the symbol <ü> also occurs only one time in the whole dataset, found in the word <tsüla> 'field, paddy.' The Matu' uwal cognate for this word is *cəlaq*, with a weak central vowel corresponding to <ü>. There are few correspondences of Matu' uwal [ə] in the Taokas-10 data: one example is Taokas-10 <huma> and Matu' uwal *həmaʔ* 'tongue,' where the weak vowel in Matu' uwal corresponds to <u> instead. The <ü> in Taokas-10 <tsüla> may have been an attempt to represent a more lax or centralized sound than a regular high back vowel.

## 5.3 Comparison between Taokas-10 and Matu' uwal phonology

This section provides a brief description and comparison of Taokas-10 and Matu' uwal phonology. Based on lexical evidence from section 4, Taokas-10 is likely to be an Atayal dialect particularly closely related to Matu' uwal. We would therefore expect its phonology to be similar to that of Matu' uwal or other Atayal varieties.

From that point of view, the dataset has a number of phonological mergers. Some of these mergers are more or less systematic, but others show a considerable amount of variation when compared to Atayal data, as well as reductions that would be expected from a native speaker of a Sinitic language.

Taokas-10 lacks several phonemes that are present in Matu' uwal, and has inconsistent reflexes for others. Some of the correspondences depend on the phonological environment: Taokas-10 tends to delete word-final plosives and fricatives, though they are occasionally preserved. The full list of correspondences can be seen in table 17.

Table 17: Sound correspondences between Matu' uwal and Taokas-10

| Matu' uwal | Taokas-10 | Matu' uwal | Taokas-10 |
|---|---|---|---|
| p | p | m | m |
| t | t, ∅ /_# | n | n |
| k | k, ∅ /_# | ŋ | ng, n, m |
| q | k, ∅ | l | l, ∅, n /_# |
| ʔ | ∅ | r | l |
| b | v, w | w | w, v, u /_]σ |
| g | w, gw, k, ∅ | y | y, i /_]σ |
| c | ts, ch' | a | a |
| s | s, sh | i | i, e |
| x | h, ∅ /_# | u | u, o |

The plosives /p/, /t/, and /k/, affricate /c/, fricatives /s/ and /h/, nasals /m/ and /n/, liquid /l/, glide /y/, are represented fairly consistently in Taokas-10 before vowels, but only the nasals and glides are regularly preserved in coda position.

The low vowel /a/ is consistent between Matu' uwal and Taokas-10. High vowels in Matu' uwal often surface as mid vowels in Taokas-10, especially when adjacent to a uvular /q/ or pharyngeal /h/: compare Matu' uwal *qulih* 'fish' with Taokas-10 <kōle>. Even though the post-dorsal segments are no longer present in Taokas-10, the lowering effect they have on neighboring high vowels can still be observed.

The phonemes that we would expect to be most difficult for a Sinitic speaker to produce are /q/ and /r/. Moreover, no Sinitic language has pharyngeal fricatives, or a phonemic distinction between [x] and [h]. Voiced fricatives are largely absent from Sinitic languages, but Hakka is one of the languages that has a voiced labiodental fricative/approximant, and Hokkien has the voiced plosives [b] and [g].

The above observations are consistent with the data. Both /q/ and /r/ are completely absent from Taokas-10, /r/ fully merging with /l/, and /q/ either merging with /k/ or being deleted. The voiced bilabial obstruent /b/ surfaces as both <v> and <w>, with no apparent regularity, and /g/ shows the most variability, being transcribed as <w>, <k>, <ng>, <l>, or deleted word-finally. The voiceless velar fricative /x/ and the voiceless pharyngeal fricative /h/ are completely merged,

both transcribed as <h> before vowels, but deleted word-finally.

Matu' uwal /q/ usually corresponds to Taokas-10 <k>. Unlike other obstruents, it is usually preserved word-finally: Taokas-10 <lanek> and Matu' uwal *raniq* 'road,' or Taokas-10 <hapunek> and Matu' uwal *hapuniq* 'fire' (note the vowel lowering effect from section 5.2). Yet there are lexical items where the corresponding segment in Taokas-10 does not appear. This happens more often in word-final position: Taokas-10 <lauwe> and Matu' uwal *rawwiq* 'eye,' or Taokas-10 <tsukuli> and Matu' uwal *cuquliq* 'person.' There are also three words in the dataset where the /q/ to zero correspondence occurs before a vowel: Taokas-10 <wunaye> and Matu' uwal *bunaqiy* 'sand,' Taokas-10 <taisu> and Matu' uwal *mamaqisu?* 'nine,' Taokas-10 <inalu> and Matu' uwal *qanaruux* 'long.' The Taokas-10 lexical items <taisu> and <inalu> are dubious cognates, and may not be of Atayal origin at all.

Ogawa also recorded what were likely free variations in his consultant' s speech: both /s/ and /c/ in Matu' uwal each have two distinct correspondences in Taokas-10. The fricative /s/ is represented as either <s> or <sh>, and the affricate /c/ as either <ts> or <ch' >. These variants occur in identical environments, and are unlikely to indicate an allophonic variation or a phonemic split. Instead, they are most likely a symptom of the linguistic attrition that is seen elsewhere in the dataset.

Hakka dialects spoken in Taiwan have several features that may shed light on apparent irregularities in the Taokas-10 data. Ogawa' s language consultant lived in an area with a high percentage of Hakka speakers, and may have been a native speaker himself.

Hakka has a voicing contrast in labiodental and post-alveolar fricatives. It also has a place contrast between dentals and post-alveolars in its affricates and fricatives. The full consonant inventory is presented in table 18, sourced from Gu (2005: 119).

Table 18: Taiwanese Hakka consonants

| | | |
|---|---|---|
| p $p^h$ | t $t^h$ | k $k^h$ |
| | ts $ts^h$ | |
| | tʃ $tʃ^h$ | |
| f v | s | h |
| | ʃ ʒ | |

Table 18: Taiwanese Hakka consonants

| | | |
|---|---|---|
| m | n | ŋ |
| | l | |

The contrasts in Hakka may explain multiple correspondence sets in the Taokas-10 dataset. Matu' uwal /c/ can correspond to <ts> and <ch' >, which are phonemic in Hakka. Hakka also lacks voiced plosives, and does not have velar fricatives. The lack of a /g/ phoneme in Hakka may be responsible for the large number of correspondences of Matu' uwal /g/ [ɣ] in Taokas-10. Unlike Hakka, Taiwanese Southern Min has voiced plosives /b/ and /g/, but lacks a dental-alveolar contrast in affricates and fricatives, and is thus a less likely candidate.

The mergers observed in the dataset point to Ogawa' s language consultant (Lin Jin-sheng) being a native speaker of Hakka, based on the phonology and phonotactics of his idiolect. His knowledge of the Atayal dialect dubbed "Taokas-10" was rudimentary, and appears to have been limited to a few hundred words and very simple phrases, making him a heritage speaker.

## 5.4 Diachronic phonology of Atayal and Taokas-10

As explained in section 5.3, Taokas-10 was most likely elicited from a native speaker of a Sinitic language, and as such shows a variety of mergers, some more systematic than others, that are signs of language contact. Despite this hindrance, there is still enough evidence to identify the reflexes of Proto-Atayal (PA) phonemes and retentions that are specific to Matu' uwal.

Below is a list of five sound changes from Proto-Atayal that occurred in both Squliq and S' uli' dialects (except *q > ʔ, which happened only in S' uli' ), but not in Matu' uwal. A comparison of Taokas-10 data shows that none of these changes occurred in that dialect.

- **PA *c,*s > s**. Both Squliq and S' uli' have merged Proto-Atayal *c and *s into /s/. Compare Squliq *səlaq* and S' uli' *səlaʔ* 'mud' to Matu' uwal *cəlaq* and Taokas-10 <tsüla> (the <ü> in Taokas-10 presumably stands for a centralized high back vowel).
- **PA *t > c /_i**. Squliq dialects as well as some (but not all) S' uli' dialects affricatized all instances of /t/ before /i/. Compare Squliq and S' uli' *cimuʔ* 'salt' with Matu' uwal *timuʔ* and Taokas-10 <timu>.

- **PA *ɹ,*y > y**. Squliq and S' uli' merged Proto-Atayal *ɹ with the approximant *y, while in Matu' uwal this proto-phoneme has different reflexes depending on the environment, most of the time being deleted, but also surfacing as either /ʔ/ or /w/ word-initially before the low vowel /a/, depending on the subdialect (Li 1981: 264). In the latter case, Taokas-10 may have unique reflexes of this proto-phoneme, which are discussed in section 7. For non-initial reflexes, compare Squliq *takay* 'frog' with Matu' uwal *taka* and Taokas-10 <taka>, or Squliq *kayal* 'sky' with Matu' uwal *kaal* and Taokas-10 <kaā>.
- **PA *q,*ʔ > ʔ**. Only S' uli' has this merger. It merged *q into the glottal stop in all positions. The Taokas-10 speaker could not produce this sound, but he pronounced it as [k] in 27 out of 35 lexical items where it should occur. Compare S' uli' *ʔutux* 'one' with Matu' uwal *qutux* and Taokas-10 <kutu>.
- **Vowel weakening**. Squliq and S'uli'preserve vowel distinctions only in the last foot. All vowels preceding it are weakened, either into [ə] or into [a], depending on the dialect. Matu' uwal still preserves vowel distinctions in prepenultimate position, and Taokas-10 data shows the same distinctions. For example, Taokas-10 <holake> and Matu'uwal *hulaqiy*, but Squliq *həlaqi* 'snow' ; Taokas-10 <hauinu> and Matu' uwal *hawinuk*, but Squliq *həwinuk* 'waist, lower back.'

The aforementioned changes are presented in compact form in table 19, illustrating the contrast between Matu' uwal and Taokas-10 on the one hand, and Squliq and S' uli' on the other hand. The table shows the reflexes of Proto-Atayal phonemes in the four dialects.

Table 19: Sound changes in Matu' uwal, Taokas-10, Squliq, and S' uli'

| Proto-Atayal | Matu' uwal | Taokas-10 | Squliq | S' uli' |
|---|---|---|---|---|
| *c | c | ts, ch' | s | s |
| *t /_i | t | t | c | t/c |
| *ɹ | ∅/w | ∅/h | y | y |
| *q | q | k/∅ | q | ʔ |
| V /_σσ# | V | V | ə | ə/a |

Squliq and S' uli' share three of the sound changes (a fourth one, *t palataliza-

tion before /i/, is not attested in all S' uli' dialects). Taokas-10 shares none of the sound changes with either Squliq or S' uli' ; in fact, the only sound that underwent a change in that dialect is Proto-Atayal *ɪ. Likewise, Matu' uwal only changed Proto-Atayal *ɪ, although in a different way. This is explored further in section 7.1.

# 6 Morphosyntactic evidence

There is very little morphosyntax that can be seen in the data, but certain words and phrases contain affixes, nominal case markers, auxiliary predicates, and deictic pronouns that are shared with Matu' uwal Atayal.

## 6.1 Affixation

The common Austronesian Actor Voice infix *-um-* surfaces two times in the data: in <h-um-akai> 'to walk,' and <k-um-ita> 'to see.' This is important for Atayal data, because only Matu' uwal and Plngawan preserve the vowel on the infix, with the rest of the dialects just having a single consonant *-m-* infix.

Likewise, the Actor Voice prefix *ma-* (used mostly with stative or reciprocal predicates) is only found in Matu'uwal and Plngawan, whereas in other Atayal dialects it has become *m-*. This prefix is found in a number of Taokas-10 lexical items, listed in table 20.

Table 20: Prefix *ma-* in Taokas-10

| Taoka-10 | Matu' uwal | Gloss |
|---|---|---|
| ma-kilu | ma-kilux | 'hot' |
| ma-olang | ma-ʔurag | 'dirty' |
| ma-tana | ma-tanah | 'red' |
| ma-ova | ma-ʔubaʔ | 'white' |
| ma-sha | ma-siyaq | 'to laugh' |
| ma-t' ao | mantahuuk | 'to sit' |
| ma-avi | – | 'to sleep' |
| ma-koas | ma-quwas | 'to sing' |
| ma-patus | (patus) | 'to fire a rifle (?)' |
| ma-hoke | mə-nahuqil | 'to die' |

| Taoka-10 | Matu' uwal | Gloss |
|---|---|---|
| ma-ch' uvu | (c<um>buʔ) | 'to shoot' |

Most of the words in the table have direct counterparts in Matu' uwal. Three appear to differ: <ma-avi> 'to sleep' is similar to Squliq *məʔabiʔ* (Matu' uwal has *maqilaap*/*maqaylup* instead), <ma-ch' uvu> 'to shoot' uses a prefix instead of an infix like Matu' uwal *c<um>buʔ*, and <ma-patus> is not used in Matu' uwal (*cumbuʔ* is used in this sense), but a related term can be found in Squliq *matus*, derived from *patus* 'gun.' [8]

A derivational prefix *mas(i)-* can be seen in the Taokas-10 word <mashu-ulake> 'to give birth,' with the Matu' uwal cognate *masʔulaqiʔ*, derived from *ʔulaqiʔ* 'child.' This prefix is not found in the Squliq word *melaqiʔ* 'to give birth.'

## 6.2 Nominal case markers

The Taokas-10 data includes very few phrases, and those appear to be mostly stripped of any case markers and linkers, and adhere more to Sinitic word order. There is only one entry where a case marker can be identified: <itsasan> 'tomorrow,' which can be split into the case marker *ʔi* and the stem *casan*.

The citation form for the word 'tomorrow' in Matu' uwal is *casan*, but it is always used with the marker *ʔi*, as are all future temporal adverbs. All past temporal adverbs take the marker *cu* instead, for example *cu hisaʔ* 'yesterday.' Taokas-10 data does not include any temporal adverbs in the past, but we would expect them to be preceded by *cu*.

## 6.3 Auxiliary verbs

The Taokas-10 data includes two existential verbs: <kia> 'to have, to exist,' and the negative existential <ukas> 'not to have, not to exist.' The former is found in both Matu' uwal (*kiya*) and Squliq (*kya*), but its use in Squliq is restricted, as it co-exists with two other existential verbs: *makiʔ* and *nyux*/*cyux*. In Matu' uwal it is the most common existential verb.

[8] Firearms were obviously unknown to the Atayal before their introduction by foreigners, but this word used to refer to a type of hand-powered slingshot, and is not a recent borrowing.

Of extant Atayal dialects, the negative existential *ʔukas* exists only in Matu' uwal. Skikun has *ʔuka*, while other dialects use *ʔuŋat* in this function. Matu' uwal *ʔukas* and Taokas-10 <ukas>, as well as Skikun *ʔuka*, are likely inherited from PAn *uka 'negative existential,' but the final /s/ in this word is unique to Matu' uwal and Taokas-10.[9]

The Taokas-10 data also has what appears to be the auxiliary verb *ʔasi* in the entry <asehèn> 'sweet.' The word for 'sweet' is *lalbiŋ* in Matu' uwal, and *səbiŋ* in Squliq. It is likely that the speaker could not remember the basic word, and used a descriptive construction instead: what in Matu' uwal would be *ʔasi hiiŋ*, literally 'like honey.' The form <heng> can be seen in the wordlist glossed as 'sugar,' but its primary meaning is 'honey.' Using *ʔasi* in this way is a Matu' uwal feature, whereas Squliq would use the verb *giwan* in an identical construction.

## 6.4 Deictics

Only one deictic pronoun can be seen in the data, occurring in the phrase <nanu kahani>, glossed as 'what,' but literally meaning 'what is this.' It is unclear if there is a word boundary in <kahani> (e.g. *ka* + *hani*). The proximal deictic 'this' is *hani* in Matu' uwal, S' uli' , and Plngawan; whereas in Squliq it is *qani*.

If <kahani> is a single unit, it may be a reflex of the original Proto-Atayal form that later became *hani*/*kani*/*qani*, depending on the dialect. If it is not, and is instead preceded by a marker or linker, then it is identical to the Matu' uwal form, although the function of *ka* here is unclear.

If the phrase can be analysed as *nanu ka hani*, we would expect it to be a nominative case marker. The nominative case marker is *ku* in Matu' uwal and S' uli' , and *qu* in Squliq; but Plngawan has *ka* and Skikun has *qa*, so it is entirely possible for Taokas-10 to have had *ka* as a case marker retained from Proto-Atayal. Alternatively, the *ka* could be a reflex of the topic marker *ga. Unfortunately, there is not enough syntactic data in the wordlist to determine this.

[9] An anonymous reviewer points out that this final /s/ may come from a captured case marker that has been reanalyzed as part of the function word. This is possible, although no case markers in contemporary Matu' uwal start with /s/ (Li 1995; L.M. Huang 1995). Regardless of its origin, this innovation is uniquely shared between Matu' uwal and Taokas-10.

# 7 Unique features of Taokas-10

Sections 4, 5, 6 showed how Taokas-10 is closely related to Matu' uwal in its lexicon, phonology, and morphosyntax. However, even discounting borrowings and imperfect language retention, there is some data that distinguishes it from Matu' uwal. Apart from a possible unique reflex of the proximal deictic pronoun, discussed in section 6.4, it had unique reflexes of Proto-Atayal *ɹ, and some unique lexical items in the male-female lexical register.

## 7.1 Unique reflexes of Proto-Atayal *ɹ

Proto-Atayal *ɹ underwent mergers in different dialects of Atayal, merging with *y in Squliq, S' uli' , and Skikun. In Plngawan, it remained a separate phoneme /ɹ/. In Matu' uwal, it was usually deleted, but in word-initial position before the vowel *a, it is reflected either as /ʔ/ or as /w/, depending on the subdialect. Different reflexes in different villages mean that Proto-Atayal *ɹ was still present word-initially as a separate phoneme some time after Matu' uwal split off from the rest of Atayal.

Taokas-10 likely had a different reflex of Proto-Atayal *ɹ in some positions, specifically word-initially before certain vowels. Before the low vowel /a/, it was transcribed by Ogawa as <h>, as shown in table 21. Plngawan and Squliq cognates are given for comparison, where they can be found.

Table 21: Word-initial *ɹ before /a/ in Taokas-10

| Proto-Atayal | Taokas-10 | Matu' uwal | Plngawan | Squliq | Gloss |
|---|---|---|---|---|---|
| *ɹamil | hamin | (w)amil | (sapit) | yamil | 'footwear' |
| *ɹaŋ[rl]ic | haŋli | (w)aŋriʔ | ɹaŋlit | yəŋəliʔ | 'housefly' |

In Matu' uwal words that historically had a word-initial *ɹ followed by a low vowel, there are different reflexes in the otherwise mostly identical subdialects of Tabilas and Sahiyang (Li 1981: 264). It was reflected as /w/ in Tabilas and as /ʔ/ in Sahiyang, but only in this position. Squliq has /y/ and Plngawan has /ɹ/ in all positions for this proto-phoneme. Interestingly, Taokas-10 shows an <h> here instead. In the rest of the data, the grapheme <h> corresponds only to /h/ or /x/ in Matu' uwal.

An interesting development can be observed in the Taokas-10 word <ake> 'bad.' Plngawan *ɹakih* and Squliq *yaqih* seem to indicate an initial Proto-Atayal *ɹ here as well, but Matu' uwal only has the form *ʔaqih* and not the expected doublet **waqih.[10] This may mean that this lexeme sporadically lost the initial rhotic in Matu' uwal and Taokas-10.

There are also forms with reflexes of Proto-Atayal initial *ɹ followed by *u in the data. These words have initial <y> in Taokas-10, and at first glance look like loans from Squliq or S' uli' , where all instances of *ɹ became /y/ (/y/ stands for IPA [j], the palatal approximant). They are shown in table 22.

Table 22: Reflexes of PA word-initial *ɹ before *u in Taokas-10

| Proto-Atayal | Taokas-10 | Matu' uwal | Plngawan | Squliq | Gloss |
|---|---|---|---|---|---|
| *ɹuŋay | yuŋai | ʔuŋay | ɹuŋiy | yuŋay | 'monkey' |
| *ɹutiq | yutek | ʔutiq | (raxal) | (rəhyal) | 'earth' |

Matu' uwal has no alternative reflexes of *ɹ before /u/: it is always a glottal stop word-initially. Other dialects still have their regular reflexes: Plngawan /ɹ/, Squliq and S' uli' /y/. There are only two words in the dataset reflecting Proto-Atayal word-initial *ɹ followed by a high vowel: <yungai> 'monkey' and <yutek> 'earth.' Both Squliq and S' uli' Atayal have the form *yuŋay* 'monkey,' and one might assume that the Taokas-10 form was borrowed from one of these dialects.

The borrowing explanation does not work for the lexeme <yutek> 'earth.' No Atayal dialect has a cognate, except for Matu' uwal *ʔutiq* 'earth,' but it is found in Seediq as *rutiq* 'dirty.' Matu' uwal *ʔutiq* has the same sound correspondence with Taokas-10 as Matu' uwal *ʔuŋay* 'monkey,' and Seediq provides evidence for initial *ɹ in Proto-Atayal. Although only two reflexes are found in the dataset, they nevertheless suggest that Taokas-10 reflected *ɹ as /y/ word-initially before /u/.

It should be emphasized that the Taokas-10 dataset distinguishes initial <u-> from initial <yu-> sequences: compare <ulake> 'child' and <utu> 'spirit' with <yutek> 'earth' and <yungai> 'monkey.' We see initial <yu-> in Taokas-10 only in those words that had initial *ɹu- (or *yu-) in Proto-Atayal. This is a regular

[10] A double asterisk (**) is used here to mark an expected but unattested form in a daughter language, based on a reconstruction in a proto-language.

correspondence, and not a case of epenthesis.

It is not a stretch to have a rhotic reflected as both a fricative and a palatal approximant. PAn *R has a wide variety of reflexes, with [r], [g], [h], and *y* [j] being the most common (Conant 1911). Nevertheless, it would be unusual to see such varying reflexes in a single language.

## 7.2 Unique lexical items

Atayal is famous for having a male-female speech register distinction in part of its vocabulary (Li 1982). At least several hundred etyma have two lexical forms: a male and a female one. Female register forms are retentions from the protolanguage, while male register forms are innovations, usually derived from female forms through various strategies, such as suffixation, infixation, segment deletion, or segment substitution (Li 1983). This distinction has been neutralized in most dialects, with the notable exception of Matu' uwal. All other dialects usually preserve just one form for any given etymon, but choose randomly between the male and the female register (Li 1982).

Even though a few elderly speakers of Matu' uwal still preserve the speech register distinction, some forms have been lost, but still surface in other dialects, for example Matu' uwal *mitaal* 'to look, to see' and Squliq *mita?*, ultimately from PAn *kita. The Matu' uwal reflex has a suffix, and so is a derived (male register) form. Table 23 shows the unique reflexes of some words found in Taokas-10, along with their cognates in Matu' uwal and Squliq.

Table 23: Unique lexical items in Taokas-10

| Taokas-10 | Matu' uwal | Squliq | Gloss |
|---|---|---|---|
| bungahi | buŋaʔ | ŋahiʔ | 'sweet potato' |
| kumita | mitaal | mitaʔ | 'look, see' |
| watsihun | balihun | bəlihun | 'door' |
| kach' uwin | qacuʔ | qasuʔ | 'boat' |

The first two words, <bungahi> 'sweet potato' and <kumita> 'to look, to see,' appear more similar to Squliq reflexes, but with all three syllables intact.[11] The latter two forms in all likelihood were originally *bacihun and *qacuwin(g), but

[11]The form *buŋahiʔ* was also reported by Li (1981) to occur in Matabalay, an Atayal dialect

are not attested in any other dialect. The alternation between /c/ and /l/ does not occur in the male register formation in Matu' uwal (Li 1983), but there are signs that it may have been used previously: Matu' uwal *lumiq* vs Squliq *sumiq* 'body louse' (cf. Seediq *cumiq*, with later *c > s in Squliq). Suffixing *-ing* is a possible derivation strategy, for example Matu' uwal *siyatu?* 'clothes (f)' and *situwiŋ* 'clothes (m).'

# 8 Conclusion

This paper has re-examined Taokas-10, a purportedly Taokas dialect recorded from a single speaker in Miaoli county. Even with the very limited amount and low quality of data that is available, it can be demonstrated to have a very high degree of similarity with Matu' uwal Atayal in its lexicon, phonology, and the few areas of its morphosyntax that can be gleaned.

As such, "Taokas-10" is a misnomer, as the data clearly represents an Atayal dialect. A more appropriate name would be Western Plains Atayal, reflecting the geographical area where it was once spoken.

Western Plains Atayal demonstrates unique sound changes that have not been reported in any Atayal dialect to date, especially in its reflexes of Proto-Atayal *ɪ. Furthermore, it retained several items of Atayal vocabulary that are likely remnants of the historical male-female register system.

The "Taokas-10" dataset collected by Ogawa was sourced from just one speaker, who likely did not speak Western Atayal fluently, but was instead a heritage speaker. It includes loanwords from other languages in the area: Saisiyat, Taokas, and other Atayal dialects (Squliq or S'uli') as well. It is uncertain if the loanwords were the speaker's own idiosyncrasies, or part of the Western Plains Atayal dialect. However, the unique sound changes and lexical items in the dataset serve as evidence of a larger Atayal language community in the area.

Western Plains Atayal was most probably a dialect very closely related to, but distinct from, Matu' uwal Atayal, when its unique innovations are taken into account. It was likely spoken in the lowlands of Miaoli county, where Matu' uwal oral histories place their ancestral home. According to these histories, after the arrival of large numbers of Han Chinese, some Atayal were displaced and moved

spoken in Ta-Hsing village, Miaoli county (苗栗縣大興村).

further south and inland, while others stayed and assimilated, eventually becoming completely Sinicized. This, together with the information extracted from the Western Plains Atayal dataset, suggests an Atayal presence on the Western plains of Taiwan dating to at least the 19th century. Such a conclusion goes against previous assumptions of Atayal people living exclusively in mountain areas.

# References


Abe, Akiyoshi. 1938. *Studies on place names in Taiwan [in Japanese]*. Taihoku (Taipei): Bango kenkyūkai.

Blust, Robert. 2004. Austronesian Nasal Substitution: A Survey. *Oceanic Linguistics* 43(1). 73–148.

Blust, Robert and Stephen Trussel. Ongoing. Austronesian comparative dictionary. http://www.trussel2.com/acd/.

Conant, Carlos Everett. 1911. The RGH law in Philippine languages. *Journal of the American Oriental Society* 31(1). 70–85.

Egerod, Søren. 1980. *Atayal-English dictionary* (Scandinavian Institute of Asian Studies Monograph Series 35). London: Curzon Press.

Ferrell, Raleigh. 1969. *Taiwan aboriginal groups: Problems in cultural and linguistic classification* (Institute of Ethnology, Academia Sinica, Monograph). Vol. 17. Taipei: Academia Sinica.

Goderich, Andre. 2020. *Atayal Phonology, Reconstruction, and Subgrouping*. National Tsing Hua University PhD dissertation.

Gu, Guo-Shun. 2005. *An Outline of Taiwanese Hakka [in Chinese]*. Taipei: Wunan.

Huang, Hui-chuan J. 2015. The phonemic status of /z/ in Squliq Atayal revisited. In Yuchau E. Hsiao and Lian-Hee Wee (eds.), *Capturing phonological shades within and across languages*, 243–265. Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

Huang, Lillian M. 1995. *A study of Mayrinax syntax*. Taipei: The Crane Publishing Co., Ltd.

Li, Paul Jen-kuei. 1978. A comparative vocabulary of Saisiyat dialects. *Bulletin of the Institute of History and Philology, Academia Sinica* 49(2). 133–199.

Li, Paul Jen-kuei. 1980. The phonological rules of Atayal dialects. *Bulletin of the Institute of History and Philology, Academia Sinica* 51(2). 349–405.

Li, Paul Jen-kuei. 1981. Reconstruction of Proto-Atayalic phonology. *Bulletin of the Institute of History and Philology, Academia Sinica* 52. 235–301.

Li, Paul Jen-kuei. 1982. Male and female forms of speech in the Atayalic group. *Bulletin of the Institute of History and Philology, Academia Sinica* 53. 265–304.
Li, Paul Jen-kuei. 1983. Types of lexical derivation of men' s speech in Mayrinax. *Bulletin of the Institute of History and Philology, Academia Sinica* 54(3). 1–18.
Li, Paul Jen-kuei. 1995. The case-marking system in Mayrinax, Atayal. *Bulletin of the Institute of History and Philology, Academia Sinica* 66(1). 23–51.
Li, Paul Jen-kuei. 2004. Basic vocabulary for Formosan languages and dialects. *Selected papers on Formosan languages* (Language and Linguistics Monograph Series C3), 1483–1532. Taipei: Institute of Linguistics, Academia Sinica.
Swadesh, Morris. 1971. *The origin and diversification of language*. Edited by Joel F. Sherzer. Chicago: Aldine.
Tesing Silan. 2003. *Atayal-Atayal dictionary*. Nantou: Liao Ying-zhu.
Tsuchida, Shigeru. 1982. *A comparative vocabulary of Austronesian languages of Sinicized ethnic groups in Taiwan part I: West Taiwan* (Memoirs of the Faculty of Letters 7). Tokyo: University of Tokyo.