Representation of Yine (Arawak) Morphology by Finite State Transducer Formalism

Adriano M. Ingunza[2*] and **John E. Miller**[1*] and **Arturo Oncevay**[3] and **Roberto Zariquiey**[2]
[1] Artificial Intelligence/Engineering and [2] Linguistics/Humanities
Pontificia Universidad Católica del Perú, San Miguel, Lima, Peru
[3] School of Informatics, University of Edinburgh, Scotland

Abstract

We represent the complexity of Yine (Arawak) morphology with a finite state transducer (FST) based morphological analyzer. Yine is a low-resource indigenous polysynthetic Peruvian language spoken by approximately 3,000 people and is classified as 'definitely endangered' by UNESCO. We review Yine morphology focusing on morphophonology, possessive constructions and verbal predicates. Then we develop FSTs to model these components proposing techniques to solve challenging problems such as complex patterns of incorporating open and closed category arguments. This is a work in progress and we still have more to do in the development and verification of our analyzer. Our analyzer will serve both as a tool to better document the Yine language and as a component of natural language processing (NLP) applications such as spell checking and correction.

1 Introduction

Yine is a low resource indigenous polysynthetic Peruvian language of the Arawak family spoken by approximately 3,000 people living near the Ucayali and Madre de Dios rivers, tributary rivers of the Amazon. Yine is considered "definitely endangered" according to the UNESCO Atlas of the World's Languages in danger (Moseley, 2010).

As noted by Zariquiey et al. (2019), although Yine has a typologically oriented descriptive grammar, documentation and further study of several grammatical aspects are still urgently needed since the Yine language is at risk of entering into an obsolescent and consequently disappearing status. Therefore, such work is vital to not only adequately document the Yine language, but also to support its continued vitality through computer assisted tools such as spell-checkers and machine translators.

Formal and computational representation of morphology is considered a "solved problem" based on Beesley and Karttunen's work and seminal Finite State Morphology text (Beesley and Karttunen, 2003; Karttunen and Beesley, 2005). This does not mean that representing a language is either easy or fast, especially for the case of polysynthetic languages such as Yine.

Our goal is to construct a high coverage finite state transducer (FST) morphological analyzer both to document and preserve the Yine language, and to use it in NLP applications, such as spell checking and correction, that might promote language vitality. Our contributions at this point are: 1. a partial functioning morphological analyzer for nominal and verbal constructions including possessive constructions and verbal predicates, and 2. various project decisions and FST patterns employed so far in construction of the analyzer. Given the incomplete implementation, it is too early to report meaningful project results.

Representation of Yine morphology by a FST is a work in progress. This paper describes relevant morphological features of Yine, representation of these features by FST, particularly challenging representation problems, a preliminary evaluation, and our current and planned future states.

2 Related Work

Beesley and Karttunen (2003)'s Finite State Morphology text is a highly valuable resource for representing morphology by an FST. There are also numerous morphological analyses with FST representations available. Most relevant to this task are analyses performed for other indigenous Peruvian languages: Shipibo-Konibo (Cardenas and Zeman, 2018), Quechua (Rios, 2010) and pan-Ashaninka (Ortega et al., 2020; Castro Mamani, 2020). In particular, the last work includes applications of the FST to spell-checking and segmentation.

While we do not apply our work to spell checking in this paper, that is one of our planned goals.

[*] Authors contributed equally

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 102–112
June 11, 2021. ©2021 Association for Computational Linguistics

Previously we had attempted to develop a HunSpell[1] based spell corrector, but found it too limiting given the polysynthetic nature of the Yine language. This is consistent with Pirinen and Lindén (2010, 2014), who found that FST correctors were essential to achieve performance on par with English for morphologically complex, and typically low resource, languages.

Software, tutorials, and examples for constructing FST morphology are available from the Finite State Morphology book website.[2] We use the Foma library[3] by Hulden (2009), compatible with FST Morphology, and available, along with some fine tutorials. Both applications offer a Python API, but neither is under active development. There is limited community support for Foma.

3 Linguistic Profile and Resources

Yine (ISO 639-3: *pib*) may be considered a morphosyntactically complex language due to its highly polysynthetic profile (mainly related to verbal structures). As noted by Aikhenvald (2020), Arawak languages are synthetic, predominantly head marking and suffixing, with a complex verbal morphology. Yine presents three open word classes: nouns, verbs, and adjectives (mostly by derivation); and four closed word classes: pronouns, adverbs, demonstratives and numerals. In this section, we will only discuss the pronominal system, and some features associated with the verbal and nominal morphology, since they are relevant to the current state of representation of Yine morphology by the FST formalism.

3.1 Morphological profile

As in almost all polysynthetic languages, Yine may express in just one word meanings that would require a whole sentence in other languages. This is illustrated by a complex predicative construction in (1), and a full possessive construction, in (2). Our morphological analysis is based on Hanson (2010)'s grammatical description; glosses have been adapted to the UniMorph schema (Kirov et al., 2018).

(1) niklokgimatanaktatkalu
ø-nikloka-gima-ta
ARGNO3SM-swallow-QUOT-LGSPEC1
-na-kta-tka-lu
-LGSPEC2-INDF-PFV-ARGAC3SM
'(The huge snake) swallowed him up somehow, reportedly.'

(2) ragmunateymana
r-gagmuna-te-yma-**na**
PSS3P-tree-PSSD-COM/INS-**PSS3P**
'With their trees'

Note that Yine's morphological complexity involves vowel deletion as seen in (1) and morphemes that may be accounted for as circumfixes, as is the case of possession marking in (2) where possessor indexation is achieved with two elements: prefix *r-* and the suffix *-na*. Its implications for FST expression are very interesting and will be discussed in §4 and §5. In the remaining subsections we present some of the mentioned features. Specifically, we present morphophonological rules, possessive constructions, verbal morphology aspects and argument indexing systems in relation with verbal predicates.

3.2 Morphophonological overview

Yine presents a rich set of morphophonological processes such as vowel deletion and rhotacism of liquid consonants. These processes are presented below.

Deletion between stem and suffix occurs when a specific group of suffixes trigger the deletion of the final vowel in the attached stem as shown in (3), where the frequentative suffix *-je* triggers the deletion of the stem's final vowel. However, this can only occur if vowel deletion does not generate a cluster of three consonants which is an overall restriction in the language as can be seen in (4), where the stem remains complete in its overt realization and avoids the sequence /mkj/.

(3) nnukjetlu
n-nuka-je-ta
ARGNO1S-eat-HAB-LGSPEC1
-lu
-ARGAC3SM
'I eat it (usually)'

(4) numkajetlu
n-gimka-je-ta
ARGNO1S-sleep-HAB-LGSPEC1
-lu
-ARGAC3SM
'I make you sleep (usually)'

Prefixing of possessive morphemes triggers other morphophonological processes that will be

[1] http://hunspell.github.io/
[2] http://www.fsmbook.com
[3] https://fomafst.github.io/

explained in §3.3. In (5) we see /l/ rhotacism, which occurs when an /l/ initial suffix mutates /l/ to /r/ when attached to a stem ending in *i, e, u* or *n*. Example (6) shows how the suffix behaves when attached to a different ending stem. Note that it also occurs an internal-boundary vowel deletion process triggered by the third person suffix.

(5) pnikanru
p-nika-ni-lu
ARGNO2S-eat-DED-ARGAC3SM
'You will eat it (masc)'

(6) pniklu
p-nika-lu
ARGNO2S-eat-ARGAC3SM
'You eat it (masc)'

It is important to notice that the set of morphophonological rules developed by Hanson (2010) is neither exhaustive nor conclusive. The author mentions that a complete description of the morphological patterns of the language is still needed and leaves many issues open for further study. Thus, our application of them is based not only on the explicit description of Hanson (2010) but also in the examples presented by the author which entails some systematizable rules for our work. For example, examples (7) and (8) and how how the same 1PL object morpheme *wu* triggers vowel deletion in (7) and does not in (8) where it would create an identical consonant cluster *ww*. So, although vowel deletion seems to be lexically specified as mentioned by the author, phonological constraints seem to be highly relevant.

(7) yimaka giyolikletwuna
Ø-yimaka giyolika-le-ta
ARGNO3P-teach.hunt-COMP-LGSPEC1
-wu-na
-ARGAC1P-ARGNO3P
'They taught us (how) to hunt'.

(8) kaspukawawuna
Ø-kaspuka-**wa-wu**
ARGNO3P-let.go-**IMPFV-ARGAC1P**
-na
-ARGNO3P
'They are letting us go'.

There are other morphophonological rules applied in word formation which need to be studied in depth. Rules applied to prefixation processes, are presented in the next section.

3.3 Possessive constructions

Possessive constructions in Yine are formed by a possessor prefix (and if needed a linked possessor suffix), a possessed nominal root and, when needed, a 'possession status' suffix. Both morphological elements (i.e. the possessor prefix and the possession status suffixes) are determined by the semantics of the root they attach in terms of alienability. According to Hanson (2010) and Aikhenvald (2020), nominals are lexically specified for alienable versus inalienable possession.

Alienability is a category that makes a morphosyntactic distinction between possession that can be terminated (alienables) and possession that cannot (inalienable) (Payne, 2007). Of course, this is a language specific categorization. For example, in Yine, concepts such as *house* or *language*, are inalienable but a concept like *husband* is alienable. Nevertheless, concepts like *mother* or *hand* tend to be classified as inalienable in those languages that reflect this distinction in their grammar. Additionally, in Yine inalienable nouns present an internal sub-classification distinguishing between kinship terms (like *mother* or *son*) and non-kinship terms (like *hand* or *house*).

Depending on the noun root class and its initial consonant, Yine possessive constructions will use one of the three pronominal sets for possessor indexing.

Class 1 prefixes attach indistinctly to alienable or inalienable roots but only to those beginning with /*g*/. This consonant is always replaced by the pronoun. Additionally, if the first consonant is followed by a /*u*/, it mutates to a /*i*/ (this is always true with the exception of the 2PL prefix).

Class 2 prefixes attach also to alienable and inalienable roots with exception of non-kinship inalienable roots. Regarding morphophonology, this class does not attach to stems beginning with /*g*/ and does not replace the initial consonant of the stem. Classes 1 and 2 are almost identical, only differing in the 3rd person masculine/plural prefix: class 1 uses /*r*/ and class 2 uses a ø form.

Class 3 prefixes are attached only with those inalienable stems that do not begin with /*g*/. In the examples below we present the application of each pronominal class. The class 1 prefix pronoun for 1st person singular and its morphophonological effects on an alienable root is shown in (9), Class 2 prefix pronoun for 2nd person singular attached to an inalienable root is shown in (10), and Class

3 prefix pronoun for 3rd person plural is shown in (11). Finally, Class 3 forms for 3rd person plural are shown in (2) and (12).

(9) nutsrukate
n-gitsruka-te
PSS1S-ancestor-PSSD
'My ancestor'

(10) gmeknatjirne
g-meknatjir-ne
PSS2S-brother in law-PL
'Your brothers in law'

(11) gikamrurna
gi-kamruru-na
PSS3P-work-PSS3P
'Their work'

A last consequence of lexical specification of nominal stems is the usage of the so called 'possessed status suffixes'. These are affixed to alienable stems when possessor is expressed, as shown in (9) with *-te*, and to inalienable stems when possessor is not expressed as in (13) where *-chi* is used.

3.4 Verbal and verbal predicate morphology

Hanson (2010) treats morphological elements corresponding exclusively to the verbal stem separately from verbal predicate elements. She makes this separation to better leverage the commonality between verbal, nominal and adjectival predicates also attested to in Yine. Verbal stem morphology is exclusive to verbal stems, whereas predicative morphology may be applied to any predicate type.

Verbal stem morphology includes noun incorporants, oblique markers, evidentials, adverbial incorporants, aspect and subordination information, stem closure morphology, applicative suffixes and voice and mood morphemes.Verbal stem complexity is shown in (12). Notice that the example is not a simple stem but a predicate. Bolded morphemes correspond to what Hanson (2010) considers stem morphology.

(12) rustakatsyeggimatanrona
r-**gistaka-tsa-yegi-gima**
ARGNO3P-**cut-cord.of-PROX-QUOT**
-ta-na-
-LGSPEC1-LGSPEC2
-lo-na
-ARGAC3SF-ARGNO3P
'They cut the rope near her, reportedly'

Argument indexing and 'external aspect' specification do not correspond to the verbal stem but to the predicative morphology. Argument indexing is achieved by using prefixation for subjects and suffixation for objects. As for possessor indexing, 3PL forms are indexed by two morphological elements: prefix *r* and suffix *-na*. The pronominal forms are almost the same as the ones used for possessive constructions. The main distinction is that only classes 1 are 2 are used. Pronominal indexes are also classified in two classes and follow a regular pattern.

3.5 Available linguistic resources

Linguistic resources used for this paper such as analysis and corpora, come from three principal sources: Hanson (2010) which is a comprehensive typological oriented grammar, a Yine-Spanish/Spanish-Yine dictionary by Wise (1986) , and a theoretical guide developed by Zapata et al. (2017). Additionally, we used a Yine corpus by Bustamante et al. (2020) for evalution purposes (see §6).

4 Finite State Morphology

In the FST morphology formalism (Figure 1A), parallel language representations (tapes) are mapped one to the other, where by convention the upper tape corresponds to the morphological analysis and and lower tape corresponds to the word form. Each level accepts (generates) valid strings in their respective tape, and either level can be transduced to corresponding (possibly multiple) strings on the other level. FSTs can be stacked so that a lower or upper tape feeds into the corresponding tape of another FST. In summary: 1. words can be transduced to morphological analyses, 2. morphological analyses can be transduced to words, 3. only valid representations are accepted (generated) on either side, 4. a valid input representation may result in multiple output representations, and 5. transducers can be stacked to multiple levels.

Scripting for FST (see Figure 1B) includes an optional *Lexc* language for lexicons and an expansive *FST* language. While *Lexc* is a good fit for ordered concatenative morphology and is accessible for entering inventories of open category roots, it is not a natural fit for the highly agglutinative polysynthetic Yine language with its relatively free order of suffixes. Instead open category root inventories are edited in spreadsheets and exported via

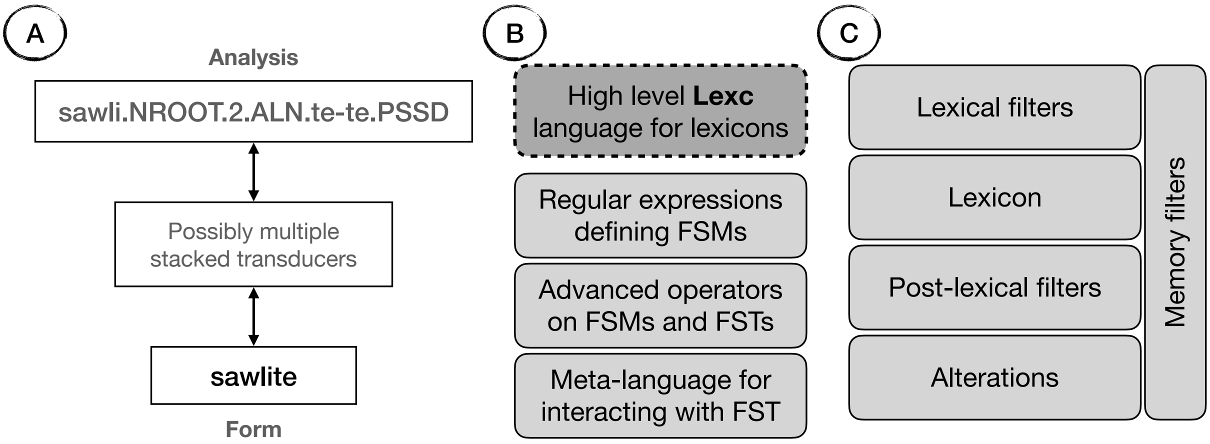
A
Analysis
sawli.NROOT.2.ALN.te-te.PSSD
Possibly multiple stacked transducers
sawlite
Form
B
High level **Lexc** language for lexicons
Regular expressions defining FSMs
Advanced operators on FSMs and FSTs
Meta-language for interacting with FST
C
Lexical filters
Lexicon
Post-lexical filters
Alterations
Memory filters

Figure 1: Language views: A) Upper analysis and lower form, B) By level/domain, C) By function.

Python scripts to FST source files. All morphological analysis is coded in *FST*, consistent with efforts by (Cardenas and Zeman, 2018; Ortega et al., 2020; Castro Mamani, 2020) for other Amazonian languages.

The *FST* language can be viewed as divided into regular expressions (defining finite state machines (FSMs)) typically used for string searching or pattern matching, advanced operators on FSMs or FSTs, and a meta-language for interacting with FSTs. Regular expressions largely suffice for the analysis tape; cross-product, rewrite rule, composition, and containment advanced operators are essential for operating on FSMs and FSTs; define, apply, file related, and virtual stack machine related meta-commands let us construct and interact with FSTs and the operating system.

FST components may also be grouped functionally as lexical, post-lexical and memory filters; lexicon; and alterations (Figure 1C). Filters which restrict lexical generation precede the lexicon; they serve to restrict the allowable combinations of constituent morphemes that might be generated by the lexicon. The lexicon, originates all constituent morphemes from both open and closed morpheme classes generating all possible valid (mostly) lexical sequences.

Sometimes it is difficult to prospectively generate only valid analyses, and so filters may be used to prevent over-generation. Similarly, some problems of over-generation (e.g., duplication) are more readily solved after generation with post-lexical filters. Phonological and morphophonological processing often imposes constraints on surface form realization of the morphological analyses, e.g., final vowel elision or rhotacism. Such constraints are implemented as alterations of the lexical analysis. Long range or discontinuous morphological relations are not readily handled by FSTs, but with use of limited memory based filters, with diacritic flags, even these problems can be resolved.

We chose to divide and conquer the analyzer project based on (Hanson, 2010)'s Yine Grammar structure. We define common terms, closed class morphemes, and open class roots, followed by higher level constructs expressible as single words: adjective, noun, noun phrase, verb, nominalization, predicate, and clause.

In the next section on morphological analysis we will see cogent examples combining language analysis from the previous section and finite state morphology described here.

5 Morphological analysis

We report several morphological analyses and snippets of corresponding FST code. FST is a multi-use term applying to simple definitions, regular expressions, filters, alterations, lexicon and the entire analyzer. All the terms beginning with /•/ are symbolic terms defined in file `common-u.foma`; their corresponding implementation specific and Unimorph terms are substituted on evaluation. Listing 1 shows a snippet of label definitions.

```
define •NRoot ".NROOT";
define •VRoot ".VROOT";
...
define •Quot ".QUOT";  # quote
define •Infer ".INFER";  # inference
```

Listing 1: Label definitions

```
define NRoot [
  [ {kamruru} [•NRoot •PossPfx3
    •Inalienable]:0 ]
| [ {gagmuna} [•NRoot •PossPfx1
    •Alienable •PossSfxte]:0 ]
];
```

Listing 2: Noun root snippet

5.1 Open word categories

Open word vocabulary is processed using Python scripts to construct root constituents with coded lexical information. The snippet in listing 2 defines noun roots, *kamruru* and *gagmuna* with form, alienability, and possessor prefix class. Inalienable nouns are further marked with `•Kin` when a kinship term. Alienable nouns are marked for their possessed suffix type. Possessor prefix class is largely determinable from alienability, kinship, and whether the initial sound segment is /g/, but it was more convenient, to index it directly. Note the use of `define` to define the FST of all noun roots and assign it to `NRoot`. The form `{kamruru}` is expanded to a string of characters and available on both the upper and lower tapes of this transducer. The regular expression `[•NRoot •PossPfx3 •Inalienable]:0` groups together the sequence of analysis terms on the upper tape as `.NROOT.3.NALN` and and maps them to ø on the lower tape via the `:` cross-product operation.

5.2 Noun root examples

Yine noun roots from the example just above are shown in (13) and (14). Inalienable nouns are preferentially possessed and are marked with the suffix *-chi* when unpossessed. Alienable nouns can readily occur without a possessor (unmarked) and are marked with their possessed suffix when possessed.[4]

(13) kamrurchi
kamruru-chi
work-UNPSSD

‘(the unpossessed) work’

`kamruru.NROOT.3.NALN-chi.UNPSSD`

(14) gagmunate
gagmuna-te
tree-PSSD

‘(a possessed) tree’

`gagmuna.NROOT.1.ALN.te-te.PSSD`

Listing 3 shows how noun possession is defined by FST. Inalienable unpossessed state is marked with *-chi* by selecting inalienable nouns, `$[•Inalienable]`, from noun roots, `NRoot`, writing the noun root and `-chi •Unposs` on the upper tape, and noun root and `^V chi` on the lower tape. `$[•Inalienable]` is a lexical filter which when composed, `.o.`, with noun roots from the lexicon selects only inalienable noun roots. The intermediate flag `^V` subsequently triggers a final vowel elision, defined by `VElision`.[5] Alienable possessed state is defined similarly except that for possessed suffix *-te* there is no final vowel elision.

```
define NounInalienUnposs $[•Inalienable]
    .o. [NRoot %-:"^V" {chi} •Unposs:0]
    .o. VElision;

define NounTe $[•Alienable •PossSfxte]
    .o. [NRoot %-:0 {te} •Poss:0];
```

Listing 3: Noun possession regexes

5.3 Nominal example

The noun shown in (15) is copied from (2) above. Word construction shows several phenomena taken into account by the FST: 1. possessor class 1 (stem with initial /*g*/), 2. comitative noun case, 3. elision alteration of initial /*g*/, 4. discontinuous dependency for possessor 3rd person plural.

(15) ragmunateymana
r-gagmuna-te-yma-na
PSS3P-tree-PSSD-COM/INS-PSS3P

‘With their trees’

`r.PSS3P-gagmuna.NROOT.1.ALN.te`
`-te.PSSD-yma.COM/INS-na.PSS3P`

The snippet in listing 4 shows 3rd person singular and plural prefixes from possessor prefix class 1. For the singular case, `t •3SgFPssr -` is written to the upper tape, and `t ^g` to the lower tape. The intermediate flag `^g` subsequently triggers an alteration due to the initial /*g*/. The plural case adds complexity with a diacritic flag being set to positive by `@P.PSSR.3PL@` for both upper and lower tapes, in addition to writing `r •3PlPssr -` to the upper tape and `r ^g` to the lower tape. The diacritic flag with feature `PSSR` remembers its setting and permits completion of the word with the `PSS3P`

[4] The annotations shown in (13, 14) use standard four-line glossing format customary in contemporary grammatical description. Output from the FST morphological analyzer is added as a fifth line of the gloss.

[5] Intermediate flags are an essential technique for triggering alterations. See alteration rule examples in (Hulden, 2011).

suffix.[6]

```
define PronNPfxScl [
  ...
  | {t} [•3SgFPssr %-] : "^g"
  | "@P.PSSR.3PL@" {r} [•3PlPssr %-]:"^g"
];
```

Listing 4: Possessor paradigm 1 (initial 'g')

The snippet in listing 5 presents three mutually exclusive noun case alternatives of which comitative is matched in analysis; and so the comitative `-yma •Com` is written to the upper tape and `yma` to the lower tape. None of the cases trigger vowel elision.

```
define NounCase [
  %-:0 {yma} •Com:0
  | %-:0 {yegi} •Circ:0
  | %-:0 {ya} •Loc:0
];
```

Listing 5: Comitative noun case

The snippet in listing 6 decides whether or not to show the `PSS3P` suffix based on the `PSSR` diacritic flag setting. If the flag setting meets the `3PL` requirement, then `-na •3PlPssr` is written to the upper tape and `^Vu na` is written to the lower (intermediate) tape. The intermediate flag `^Vu` subsequently triggers an alteration of final vowel elision except for /*u*/. If the `PSSR` diacritic flag is not set then nothing is written to either tape; in this way the FST can accept the discontinuous 3rd person plural possessor.

```
define Pron3PlNSfx [
  %-:"^Vu" "@R.PSSR.3PL@" {na} •3PlPssr:0
  | "@D.PSSR@"
];
```

Listing 6: Possessor 3rd person plural suffix

The snippet in listing 7 generates the noun from optional possessor class 1 prefix, noun root, optional noun plural, optional noun case and diacritic flag determined 3rd person plural suffix. The alteration FSTs are composed with the lexical output to handle changes due to initial /*g*/, final vowel elision, or final vowel elision for vowels other than /*u*/.

```
define Nouns [•Noun:0 [
  (PronNPfxScl) NounPfx1 (NounPlural)
  (NounCase) Pron3PlNSfx
  ...
  .o. gAlteration
  .o. VElision
  .o. VuElision;
```

Listing 7: Noun generation

The word *ragmunateymana* shows application of both the initial /*g*/ and final vowel elision except for /*u*/ alterations. The snippet in listing 8 shows how an initial *gi* is rewritten as /*u*/ or /*g*/ is rewritten as /*ϕ*/ after the `^g` intermediate flag in the lower tape; subsequently the flag itself is erased from the lower tape.

In *ragmunateymana* the initial /*g*/ of the noun root is elided and the /*r*/ of the pronoun prefix added. The case for final vowel other than /*u*/ elision is more complex, in that the vowel is not elided if it would result in a three consonant cluster. Such is the case here and so the final /*a*/ of *-yma* need not elide before *-na*. Since the three consonant cluster includes nasal consonants, the final /*a*/ could be elided resulting in the alternative valid word form *ragmunateymna* (Hanson, 2010).

```
define gAlteration [[g i -> u || "^g" _]
  .o. [g -> 0 || "^g" _ ]
  .o. ["^g" -> 0]
];
```

Listing 8: 'g' alteration

5.4 Verb predicate mega example

The verb predicate shown in (16) is not testified to by the Yine corpus, but rather is a *tour de force* act of word creation based on the grammar by Hanson, comparable to verb predicate phrase creation in non-polysynthetic languages. The analysis shown is based on the FST analysis and shows several important word generation features: 1. subject prefix class 1 (stem with initial /*g*/), 2. associate prefix *gim-*, 3. alteration due to initial /*g*/, 4. discontinuous dependency of form for 3rd person plural, 5. multiple incorporants for verb stem, 6. open category noun incorporant, 7. marker for closure of incorporants, 8. multiple incorporants for verb predicate, 9. vowel elision.

(16) rumustakasijnegimananjetyanupluna
r-gim-gustaka-siji-ne
ARGNO3P-LGSPEC3-cut-corn-PSSD
-gima-nanu-je-ta
-QUOT-EXTNS-HAB-LGSPEC1
-ya-nu-pa-lu-na
-APPL-DED-ALL-ARGAC3SM-ARGNO3P
'It is said that they, and someone else
(usually) cut their (masc) corn during
a specific time lapse'

```
 r.ARGNO3P-gim.LGSPEC3-gustaka.VROOT.AMBI
-siji.NROOT.2.ALN.ne-ne.PSSD-gima.QUOT
-nanu.EXTNS-je.HAB-ta.LGSPEC1-ya.APPL
-nu.DED-pa.ALL-lu.ARGAC3SM-na.ARGNO3P
```

[6] Diacritic flags are a powerful yet difficult to understand addition to FST. See (Hulden, 2011) for a brief introduction and (Beesley and Karttunen, 2003, pp 339-373) for an in depth explanation with examples.

The subject pronoun prefix class 1 (with inital /*g*/) is similar to that of possessor prefix class 1 with nouns. Discontinuous behavior for `•Subj3Pl` is also similar to that for `•3PlPssr`, noun possessor 3rd person plural, with the obvious difference that the 3rd person plural subject suffix marker -*na* is now very distant from the prefix!

Adding the associative prefix *gim-* to the verb root triggers 'g' alteration for roots with initial /*g*/ similar to subject class 1. The FST, see listing 9, writes `gim •Assoc -` to the upper tape and `gim ^g` to the lower tape. The intermediate flag `^g` subsequently triggers 'g' alteration if the stem has initial /*g*/ as is the case here for the verb *gustaka*.

```
define VerbAssoc [{gim} [•Assoc %-]:"^g"];
```

Listing 9: 'g' alteration with *gim-*

A huge difference in relation nouns is that verbs and verb predicates can have several incorporated morphemes including open noun class morphemes. Individual closed form incorporants are similar in structure to NounCase (listing 5) and VerbAssoc (listing 9) above. With verb stems, multiple incorporants can appear, but each incorporant type only once, and according to Hanson (2010), the order of incorporants is flexible. The snippet in listing 10 shows forming the union of individual incorporants, and the snippet in listing 11 shows how this union is repeated over 1 to 9 iterations. While not obvious from the union (because everything is via definitions), the lexical form and analysis for each incorporant are written to the upper tape and the lexical form and a unique filter flag are written to the lower tape. The filter flags will be used to enforce the no more than one of each incorporant type rule. [7]

```
define VerbIncorporantsNoCoda [
    %-:0 NounAlienPoss 0:"^I.A"
    ...
    | VerbAspect2 0:"^I.H"
    | VerbAspect3 0:"^I.I" ];
```

Listing 10: Verb stem incorporant union

When verb stem incorporants are used, they must be followed by marking of incorporant list closure, or by a causative which also effects closure, `[VerbClosure | VerbCausative]`. While repetition for 1 to 9 iterations of the union of incorporants assures no more than 9 incorporants, it does not prevent repetition of some of the incorporants. This is where the filter flags, e.g., `"^I.H"`, are used. Composing `~[detectIncorporantDuplicates]` with the lower tape from verb incorporants excludes all cases where the same filter flag is repeated, thus eliminating repeated incorporants from the FST.

```
define VerbIncorporants
        [VerbIncorporantsNoCoda^{1,9}
        [VerbClosure | VerbCausative]]
    .o. ~[detectIncorporantDuplicates]
    .o. eraseIncorporantFlags;
```

Listing 11: Verb stem incorporants

Listing 12 shows a snippet for the FST of all duplicate filter flags. Each line such as `$["^I.A" ?* "^I.A"]` denotes the language containing that filter flag duplicated, and the union over all such flags denotes the union of languages with duplicate flags. Taking the complement of this results in all languages without duplicate flags, and composing this complement with the actual group of incorporants, excludes any cases where there are duplicate flags. This is a powerful operator!

```
define detectIncorporantDuplicates [
    $["^I.A" ?* "^I.A"]
    | $["^I.B" ?* "^I.B"]
    ...
    | $["^I.I" ?* "^I.I"]];
```

Listing 12: Incorporant test for duplicates

Alienable possessed nouns or inalienable nouns (possessed root form) can serve as incorporants. This augments the expressiveness of the verb stem dramatically in that the number of verb stem combinations now gets multiplied by the number of alienable nouns and by the number of inalienable nouns. Gloss (16) incorporates the possessed alienable noun *siji-ne*, 'corn'.

Elision processes are the same or similar for nouns and we don't repeat the FST code here. Note that with so many components in the word and multiple elision processes it is not obvious to the non-native speaker, how to derive the final word form with all applied elisions and other alterations.

5.5 Ambiguity

There may be multiple analyses for individual words of the language and similarly multiple word representations for the same analysis. This ambiguity can happen because: 1. elision of final vowels of morphemes so that forms are no longer distinct, 2. elision is optional so that inherently there are multiple forms, or 3. the same form is used across multiple morphemes. Language use is a constant

[7] Beesley and Karttunen (2003, pp 299-230) explains a lexical filter version of this. In our implementation, filter flags are written to the lower tape and post-lexical filters applied to eliminate duplicate incorporant types.

process of negotiation between ambiguity of expression and efficiency of communication.

6 Evaluation

For unit testing of noun, verb, and verb predicate analyses, we constructed forms for several distinct analyses each of 20 nouns sampled over possessor class and 20 verbs sampled over subject and object classes. Diverse analyses varied possessor/-subject/object person, number, and gender as well as noun or verb incorporants. While resulting derived forms were largely consistent with analyses, we discovered and corrected several cases of lexically specified vowel elision and rhotacism not covered in Hanson (2010)'s grammar.

For coverage on test data we sampled words matching on known root forms with 25 each of noun roots and verb roots sampled at random from a Yine corpus by Bustamante et al. (2020). This resulted in many out of vocabulary words from longer root forms than those used for selection. Yet, there remained numerous other words unrecognized (not covered) by the analyzer even though sharing the expected root. So we performed a detail error analysis from a sub-sample of 63 unrecognized words to diagnose errors and make model improvements.

The error analysis is reported in table 1. Some forms suffered from multiple errors and so error counts exceed the number of words sampled. For nouns major reasons for lack of coverage are: 1. morpheme not in FST vocabulary, 2. non-verbal predicate, 3. verbalizer changed category to verb, 4. noun root entry incorrect. For verbs major reasons for lack of coverage are: 1. morpheme not in FST vocabulary, 2. elision and rhotacism alterations, 3. nominalizer changed category to noun, 4. morpheme has more flexible order.

Corrections and improvements from easy to hard are: 1. Correct out of vocabulary, entry, and orthographic errors of roots on vocabulary spreadsheets. 2. Correct intermediate flags and alterations for elision and rhotacism. 3. Add missing suffixes and more flexible order for morpheme out of vocabulary and order errors. 4. Prioritize development of non-verbal predicate, nominalizer, and verbalizer functions to address non-verbal predicate and change of category errors.

Cardenas and Zeman (2018) obtained 78.9% average coverage over multiple domains on test data for a completed FST morphology of an Amazonian polysynthetic language. Our ≈15% coverage in a preliminary evaluation on multiple domain test data should be interpreted as a measure of the effort still to go on this project. Our goal remains a high coverage FST morphological analyzer.

Error	**Nouns**	**Verbs**
Root out of vocabulary	9	6
Morpheme out of vocabulary	7	10
Morpheme out of order	1	3
Elision incorrect	0	9
Rhotacism incorrect	0	6
Orthographic mismatch	0	2
Change of category	5	7
Non-verbal predicate	7	0
Root entry incorrect	5	1
Error counts	34	44
Sample size	30	33
Total words sampled	574	1292
Percentage recognized	11.3%	16.1%

Table 1: Lack of Coverage Reasons

7 Conclusion

We have shown our initial steps in developing noun, verb, verb predicate and pronoun categories for a morphological model of the Yine language, illustrating analyses performed and FST patterns used to solve challenging problems. Testing for analyzer coverage with real world data revealed several deficiencies, some expected (nominalizers, verbalizers, non-verbal predicate) and some surprises (unexpected elision, rhotacism, and missing morpheme errors). We will continue to improve the analyzer by fixing problems and adding major word categories and functions, now with added emphasis on testing with external data. Goals for the analyzer include both language documentation and use as a component of natural language processing (NLP) applications such as spell checking and low resource machine translation.

Acknowledgements

We are grateful to the bilingual teachers from NOPOKI Nimia Acho and Remigio Zapata. Similarly, we acknowledge the research grant of the "Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica" (CONCYTEC, Peru) under the contract 183-2018-FONDECYT-BM-IADT-MU.

References

Alexandra Aikhenvald. 2020. Morhology in arawakan languages. *Oxford Research Encyclopedia of Linguistics*.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Studies in Computational Linguistics Online. CSLI publications, Stanford University, Stanford, CA, USA.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No data to crawl? monolingual corpus creation from PDF files of truly low-resource languages in Peru. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Ronald Cardenas and Daniel Zeman. 2018. A morphological analyzer for Shipibo-konibo. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 131–139, Brussels, Belgium. Association for Computational Linguistics.

Richard Alexander Castro Mamani. 2020. Ashaninka-morph. github at https://github.com/hinantin/AshMorph.

Rebecca Hanson. 2010. *A Grammar of Yine (Piro)*. Ph.D. thesis, La Trobe University, Victoria, Australia.

Mans Hulden. 2009. Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32. Association for Computational Linguistics.

Mans Hulden. 2011. Morphological analysis with fsts. Document in github: https://fomafst.github.io/morphtut.html.

Lauri Karttunen and Kenneth R. Beesley. 2005. Twenty-five years of finite-state morphology. In Antti Arppe, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund, and Anssi Yli-Jyrä, editors, *Inquiries into Words, Constraints and Contexts*. CSLI publications, Stanford University.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Arya D. McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, , and Mans Hulden. 2018. Unimorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Christopher Moseley, editor. 2010. *Atlas of the world's languages in danger*, 3rd edition. UNESCO, Paris, France.

John Ortega, Richard Alexander Castro-Mamani, and Jaime Rafael Montoya Samame. 2020. Overcoming resistance: The normalization of an Amazonian tribal language. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 1–13, Suzhou, China. Association for Computational Linguistics.

Thomas E Payne. 2007. *Describing Morphosyntax: A guide for field linguists*. Cambridge University Press.

Tommi A. Pirinen and Krister Lindén. 2010. Creating and weighting hunspell dictionaries as finite-state automata. *Investigationes Linguisticae*.

Tommi A. Pirinen and Krister Lindén. 2014. State-of-the-art in weighted finite-state spell-checking. In *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing 2014, pages 519–532, Berlin, Heidelberg. Springer-Verlag.

Annette Rios. 2010. *Applying Finite-State Techniques to a Native American Language: Quechua*. Ph.D. thesis, Universitaẗ Zuřich.

Mary Ruth Wise, editor. 1986. *Diccionario Piro*. Summer Institute of Linguistics, Yarinacocha, Perú.

Remigio Zapata, Nimia Acho, and Gerardo Zerdin. 2017. *Guía teórica del idioma yine*. Universidad Católica Sedes Sapientae.

Roberto Zariquiey, Harald Hammarström, Mónica Arakaki, Arturo Oncevay, John Miller, Aracelli García, and Adriano Ingunza. 2019. Obsolescencia lingüística, descripción gramatical y documentación de lenguas en el perú: hacia un estado de la cuestión. *Lexis*, 43(2):271–337.

Appendix: Unimorph and Hanson Grammar Terms Used in Paper

Unimorph	Hanson (2010)	Description
ALL	ELV	Allative / Ellative
APPL	APPL	Applicative
ARGNO1S	1SG	First person singular 'subject'
ARGNO2S	2SG	Second person singular 'subject'
ARGNO3P	3PL	Third person plural 'subject'
ARGNO3SM	3SGM	Third person masculine 'subject'
ARGAC1P	1PL	First person plural 'object'
ARGAC3SM	3SgM	Third person singular masculine 'object'
ARGAC3SF	3SgF	Third person singular feminine 'object'
COMP	SUBD	Comparative (subordination function)
COM/INS	COM	Commitative (and instrumental)
DED	SUBD	Deductive (subordination function)
EXTNS	EXTNS	Extensive aspect
HAB	CONTIN	Habitual / Continuative
INDF	GENZ	Indefinitness in time
IPFV	IMPFV	Imperfective aspect
LGSPEC1	VCL	Verb Stem Closure
LGSPEC2	CMPV	Completive aspect
LGSPEC3	ASSOC	Associative
PFV	PFV	Perfective aspect
PL	PL	Plural
PROX	VICIN	Proximative
PSSD	PSSD	Possessed noun
PSS1S	1SGPSSR	First person singular possessor
PSS2S	2SGPSSR	Second person singular possessor
PSS3P	3PLPSSR	Third person plural posessor
QUOT	QUOT	Quotative (epistemic marker)
UNPSSD	UNPSSD	Unpossessed noun

Table 2: FST Morphology - UniMorph categories with Hanson (2010)'s glossing equivalents.