Towards a finite-state morphological analyser for San Mateo Huave

Francis M. Tyers
Department of Linguistics
Indiana University
Bloomington, IN
ftyers@iu.edu

Samuel Herrera Castro
Instituto de Investigaciones Antropológicas
Universidad Autónoma de México
México D.F.
sherrera@unam.mx

Abstract

This paper describes the development of a free/open-source computational morphological description for San Mateo Huave, language spoken in the state of Oaxaca in Mexico. The language is of the agglutinative morphological type, with both prefixing and suffixing morphology. Both the nominal and verbal morphology are moderately complex. Huave is under-resourced and this is the first publication describing a computational tool for the language. We use the Helsinki Finite-State Toolkit (HFST) for implementing the finite-state transducer. An automatic evaluation of the coverage of our implementation shows that the coverage is adequate, between 80% and 93% on range of freely available corpora. Both the analyser and the hand-annotated test set are available under a free/open-source licence.

1 Introduction

Mexico is home to 68 languages, 63 of them Indigenous languages. This paper concerns Huave, particularly the variety (ISO-639: huv) spoken in San Mateo del Mar (Oaxaca), referred to as *ombeayiiüts* [o^{m}b^{j}aji͡ts] 'our (incl.) speech' by its speakers. Unlike most of the Indigenous languages of Mexico, in San Mateo del Mar at least, Huave is still widely spoken by adults and children alike. It is worth noting however that the language is seriously under-resourced and this paper is the first published computational tool for the language.

Huave is an agglutinative language where much meaning is expressed via combinations of stems and prefixes and suffixes. For example the word *tatajtüw* /t-a-tajt-üw/ "they cleaned (it)" is composed as follows *t-* 'PAST', *a-* 'SG3', *tajt* < *taad* 'clean' and *-üw* 'PL. Words may also be contracted with each other, for example *timikambajaats* /ti-mi-kambaj-aats/ "in our (incl.) village" is composed of the preposition *ti-* < *tiül* 'in', *mi-* 'POSS.3', *kambaj* 'village', and *-aats* 'POSS.PL-IN'.

This means that for most natural language processing applications an important first step is tokenisation and morphological analysis. Where large amounts of data are available, unsupervised corpus-based approaches can be used, however for Huave, as with all of the Indigenous languages of Mexico, there is a dearth of data, corpus-based or otherwise.

An alternative approach, and one with a long pedigree, is to use finite-state transducers to model tokenisation and morphology (Beesley and Karttunen, 2003). A finite-state morphological analyser returns, for a given surface form, a set of all possible valid analyses. Each analysis is formed of a lemma and sequence of morphological tags. For example given the surface form *iüm* it would return, ij<num><c4> 'two-CLF', iüm<n> 'door' and <pres><sg2>üm<v><iv> 'you weigh'.[1]

Analysers of this type exist for many languages, including for many languages of the Americas, for example Yupik (Chen and Schwartz, 2018), Arawak (Ingunza et al., 2021), K'iche' (Richardson and Tyers, 2021) and Western Sierra Puebla Nahuatl (Pugh et al., 2021) and given an existing grammatical description and lexicon one can be built comparatively rapidly (Washington et al., 2014). An additional benefit of these analysers is that they can be used to generate training data for supervised data-driven morphological models, cf. (Moeller et al., 2018).

The morphological analyser in this paper is based on the Helsinki Finite-State Toolkit (Lindén et al., 2011) due to its support for weighted finite-state transducers and the twol formalism (Koskenniemi, 1983).[2] We chose the formalism to avoid issues with intermediate forms and rule-ordering in rewrite approaches.

[1] The tags are as follows: <num> 'numeral', c4 'type-4 classifier', <n> 'noun', <pres> 'present tense (atemporal)', <sg2> '2nd person singular', <v> 'verb', <iv> 'intransitive'.

[2] The twol formalism is equivalent in descriptive power to the more familiar xfst rewrite rule formalism (Karttunen, 1993).

Figure 1: A map of Mexico (background) indicating the state of Oaxaca and the district of Tehuantepec (top left). The town of San Mateo del Mar is located on the Isthmus of Tehuantepec (bottom left). It is labelled on the map with other dots indicating outlying villages where the same variety of Huave is also spoken.

The remainder of the paper is laid out as follows: section 2 gives an overview of the sociolinguistic situation and typological features of San Mateo Huave, section 3 describes prior linguistic work, section 4 describes the methodology of completing the analyser, section 5 provides an evaluation of the analyser, and looks qualitatively at the remaining issues. Sections 6 and 7 describe some future directions and offer some concluding remarks.

2 Huave

Huave is a linguistic grouping that is spoken in four municipalities in the districts of Tehuantepect and Juchitán in the state of Oaxaca in the south of Mexico: San Mateo del Mar, Santa María del Mar, San Francisco del Mar, and San Dionisio del Mar. According to *Instituto Nacional de Lenguas Indígenas de México* (INALI; National Institute of Indigenous Languages) there are two principal languages in the grouping: Eastern Huave and Western Huave (INALI, 2016). The Eastern variant includes the languages spoken in the towns of San Francisco del Mar and San Dionisio del Mar, the Western includes the languages spoken in the towns of San Mateo del Mar (see Figure 1) and Santa María del Mar. By far the municipality with the largest number of speakers is San Mateo del Mar.

As of 2020, the total population of San Mateo del Mar was 15,571 with 15,538 living in households classified as Indigenous and 14,034 speakers over the age of three (INEGI, 2020). The language is widely spoken both at home and in public places. It is used in primary education alongside Spanish, in which the vast majority of speakers are bilingual. In recent years incipient language shift has been seen with a recent survey of secondary school pupils showing only half are speakers of Huave.

Constituent order is flexible (Stairs and Hollenbach, 1981; Herrera Castro, 2010), the object usually follows the verb in transitive clauses, but the subject can appear either before the verb (Subject–Verb–Object) or after the object (Verb–Object–Subject). The difference in order can also have different phonological properties (Pak, 2014). Intransitive verbs prefer the subject to appear after the verb (Verb–Subject). On verbs, tense, aspect and modality is indicated by suffixes or auxiliary words, while person/number is indicated by combinations of prefixes and suffixes. Nouns may be possessed.

Person and number in both the possessive and verbal system has eight combinations as indicated in the following list using: First person singular [+I, -You, -Other] ‘SG1’; First person dual [+I, +You, -Other] ‘DU1’; Second person singular [-I, +You, -Other] ‘SG2’; Third person singular [-I, -You, +Other] ‘SG3’; First person plural (inclusive) [+I, +You, +Other] ‘PL1.INCL’; First person plural (exclusive) [+I, -You, +Other] ‘PL1.EXCL’; Second person plural [-I, +You, -Other] ‘PL2’; Third person plural [-I, -You, +Other] ‘PL3’.

The second and third person singular and plural functions broadly as in other languages, the function of the first person agreement (SG1, DU1, PL1.INCL, PL1.EXCL) is illustrated with the present tense forms of the verb *w* ‘go out’ in (1).

(1) a. *saw*
sa-w
1-go.out
‘I’m going out (me)’

b. *awar*
a-w-ar
3-go.out-DU
‘We’re going out (me and you)’

c. *awaats*
a-w-aats
3-go.out-PL.INCL
‘We’re going out (me, you and them)’

d. *sawan*
sa-w-an
1-go.out-PL.EXCL
‘We’re going out (me and them, not you)’

Verbs are marked for subject agreement, and rarely a limited form of object agreement.

2.1 Orthography

There have been a number of orthographic proposals for the language, the most notable being the form used by the translation of the New Testament and the Huave–Spanish dictionary of Stairs and Stairs (1981). This orthography uses Spanish-style orthographic conventions to represent /g/, /gw/, /kw/ and /k/, for example *qui* for /ki/ and *cua* for /kwa/ (cf. *miquiej* /mikieh/ 'his blood' and *cuane* /kwane/ 'what').

More recent texts produced by speakers have replaced these with *k* for /k/ and *kw* for /kw/, resulting in *mikiej* 'his blood' and *kwane* 'what', but texts with *k* and *ku* are also found. In this work we have used the recent orthography as proposed and used by speakers themselves.

3 Prior work

One of the earliest published descriptive linguistic works on Huave is a grammar sketch by Belmar (1901). There are no full length grammars, but an technical report was done by Stairs and Stairs (1983) which includes a large number of elicited example sentences and a grammar sketch (Stairs and Hollenbach, 1981) as part of a bilingual Huave–Spanish dictionary (Stairs and Stairs, 1981).

Much existing work has been done on the system of verbal morphology, including a study by Stairs and Hollenbach (1969) which looked at the verbal system from an item-and-arrangement view and an alternative word-and-paradigm model by Matthews (1972). A certain subclass of verbal affixes in Huave are *mobile* meaning that they can appear as prefixes or suffixes. This phenomenon has been studied by Noyer (1993) in the framework of Optimality Theory.

A study of the morphosyntax of the verbal system was published by Herrera Castro (2010) looking particularly at morphosyntactic alignment, person-number marking and valency changing operations (such as causative and reflexive/reciprocal).

Two other relevant works are Noyer (2013), who gives a comprehensive analysis of word-level phonology and Herrera Castro (2016) who makes a detailed study of the semantics of argument noun phrases.

There are two main published dictionaries, the first is a Huave–Spanish dictionary (Stairs and Stairs, 1981) and the second is an etymological dictionary of the Huave languages (Noyer, 2012). The latter dictionary includes reconstructed forms along with attested forms in the contemporary languages and translations in Spanish and English.

4 Methodology

There are three main components to a finite-state transducer: the lexicon (containing the lexemes and their paradigms, or continuation classes); the morphotactic rules which are implemented as continuation classes; and the morphophonological rules. This section deals with each component in turn.

4.1 Lexicon

The lexicon consists of 1,455 lexemes which were added in frequency order (calculated using the New Testament) and with reference to the two available dictionaries (Stairs and Stairs, 1981; Noyer, 2012) for part-of-speech classification. The lexicon has been created in the `lexc` formalism, which is standard in HFST.

Each entry consists of four parts. The first two are a lemma and a morphotactic stem. The lemma comes first, before the colon and is always the dictionary form of the word. The stem, which appears after the colon can contain formal symbols to simplify the writing of phonological rules. For example the transitive verb *taad* 'clean' and the noun *chiig* 'little brother' would have the following entries:

```
taad:ta{:}d V-TV ; ! "limpiar"
chiig:chi{:}g N-mi-A ; ! "hermano menor"
```

Where the IPA length symbol, {:} stands in for either the long vowel itself, *a* as in *tataad* 'He cleaned it' or the aspiration after shortening *j* /h/ as in *tatajtiiw* 'They cleaned it'. The underlying form of the stem is defined as ta{:}d and the length symbol can be realised as either a copy of the vowel, in this case *a* /a/ or as the aspiration *j* /h/.

After the lemma and stem comes a continuation class, this is a classification of the entry by part of speech and morphology. In this case the verb *taad* 'to clean' has the class V-TV, indicating transitive verb and the noun *chiig* 'younger brother' has the class N-mi-A for a noun taking the *mi* set of possessive prefixes and the *A* set of suffixes. Finally, after the exclamation mark, which indicates a comment, comes a gloss in Spanish.

	P	C	PN	N
1	*xi-*, *sa-*, *na-*, *-Vs*			*-Vn*
2	*i-*, *r-*			*-Vn*
3	*a-*			*-Vw*
DU		*a-*	*-Vr*	
1.INCL		*a-*	*-VVts*	

Table 1: Affixes for subject agreement in Huave. These are marked with a hyphen following for prefixes and a hyphen preceeding for suffixes. The uppercase V indicates a vowel governed lexically or phonologically. In the column headings, *P* stands for person, *C* for clusivity and *N* for number.

4.2 Morphotactics

In addition to the lexemes, the lexicon also contains 70 *continuation lexica* which model the morphotactics of the language. Morphemes can appear as prefixes, suffixes and infixes. In our description of morphotactics we follow the examples in the available descriptions (Stairs and Hollenbach, 1981; Herrera Castro, 2010, 2016).

4.2.1 Verbs

As mentioned in §2, verbs inflect for eight combinations of person and number: First person singular, dual, inclusive plural and exclusive plural; second person singular and plural and third person singular and plural. Agreement is for the subject of the sentence.

There are five inflectional markers of tense-aspect-mood, including punctual in *l-*, past in *t-*, progressive in *tea-* and an unmarked present and an irrealis in *ap-*. In addition there is a form used in subordinate sentences, *n-* or *m-*.

The typical structure of a verb is as follows:

> Tense/Aspect/Mood – Agreement1 – Stem – (Derivation) – Agreement2

In general, person agreement goes in the first agreement slot as a prefix and number agreement comes in the second agreement slot as a suffix (cf. Table 1). However this isnot always the case and for example the first person singular prefix can sometimes appear after the verb stem (Herrera Castro, 2010).

Although a small subset of positional stems (e.g. the verbs *chet-* 'be seated', *lemb-* 'be stopped', *py-* 'be lying down', etc.) have the following structure:

> Stem – Tense/Aspect/Mood – Agreement1 – (Derivation) – Agreement2

There are restrictions on which morphemes in each slot can cooccur, where these restrictions are on either side of the stem, for example the suffix *-as* for first person singular can only appear with the past tense *t-*, these are implemented using flag diacritics (Beesley and Karttunen, 2003).

4.2.2 Nouns

Nouns in Huave inflect for possession using a set of possessive prefixes and suffixes. They may also take a plural suffix and an impersonal suffix.

There are four subcategories of prefixes and four subcategories of suffixes, which are assigned independently to nouns. The suffixes are assigned based on the stem vowel, while the prefixes are lexically determined. In (2), the words *pix* 'clothes' and *kwal* 'son, daughter' take different forms of the 3rd person singular possessive prefix.

(2) *apix* *mikwal* *María*
a-pix mi-kwal María
POSS.SG3-clothes POSS.SG3-son María
'María's son's clothes'

The vast majority of nouns do not take a plural suffix, instead indicating plurality either with a modifying numeral, quantifier or demonstrative or with agreement on the verb.

4.2.3 Adjectives

There are two categories of words that can be considered adjectival in Huave. The first category is small and contains property words such as *nine* 'small', *rran* 'white', and *jayats* 'new'. These are not marked for attribution, but may be marked for non-present when they appear as predicates (Herrera Castro, 2010). Compare the word *rran* 'white' in (3) and (4). In (3) it is used predicatively but does not receive any marking in the present tense, whereas in (4) it appears predicatively, but with irrealis morphology, the prefix *ap-*. Note that the word *nakants* glossed as 'red' is classified as a stative verb and not an adjective and so receives the *na-* prefix any tense.

(3) *Mikamix* *a* *naxey rran* *wüx* *angal.*
mi-kamix a naxey rran wüx a-ngal.
POSS.SG3-shirt the man white when SG3-buy
'The man's shirt was white when he bought it.'

(4) *Aaga müx kyaj nakants, oxep*
aaga müx kyaj na-kants, oxep
DET kayak DEM STAT-red, tomorrow
aprran ombas nej
ap-rran ombas nej
IRR-white POSS.SG3-body SG3
'This kayak is red, tomorrow it will be white.'

Another category are stative verbs which require a stative prefix *na-* to be used both attributively and predicatively (5).

(5) *Teon tangal nawaak tixem.*
Teon t-a-ngal na-waak tixem.
Antonio PAST-SG3-buy STAT-be.dry prawn
'Antonio bought dry prawns.'

These are handled with two different stem categories in the lexicon.

4.2.4 Other categories

There are two prepositions in the language *tiül* 'in', which can be contracted to the following noun (6) and *wüx* 'on'.

(6) a. *tiül mikambajaats*
tiül mi-kambaj-aats
in POSS.SG3-village-PL.INCL
'in our village'

b. *timikambajaats*
ti-mi-kambaj-aats
in-POSS.SG3-village-PL.INCL
'in our village'

There are a set of personal pronouns, following the same person–number schema as the possessives and verbal morphology. These appear in both neutral and emphatic forms, *ikoots* 'we', *ikootsa* 'we-EMPH'. In addition there are several determiners, the most widely used being *aaga* 'the' and the set of demonstratives *kam*, *küy* 'this', *kyaj* 'that' *kyün* 'yon'.

4.3 Morphophonology

We used morphographemic rules in the *twol* formalism to model the phonological alternations. This formalism was first proposed by Koskenniemi (1983) and consists of finite-state constraints over possible input–output string pairs. In HFST (Lindén et al., 2011), constraints are applied in parallel via the composition operator and the output of each of the constraints is intersected. There are currently a total of 8 constraints, covering palatalisation, vowel reduction, assimilation, devoicing and aspiration. Figure 2 gives an example of three of the constraints

```
"Long vowel without suffix"
%{:%}:Vx <=> Vx _ ;
     except
        Vow _ Cns+ %>: [ .#. | %>: %{A%}: ] ;
     where Vx in ( a e i o ü ) ;


"Aspiration after shortened vowels"
%{:%}:j <=> Vow _ Cns+ %>: [ .#. | %>: %{A%}:
] ;


"Desonorisation of occlusive sonorant"
Cx:Cy <=> Vow %{:%}: _ %>: [ .#. | %>: %{A%}:
] ;
     where Cx in ( b d g )
           Cy in ( p t k )
     matched ;
```

Figure 2: Three phonological constraints to implement the forms *andüüüb* /a-ndüüb/ 'She follows' and *tandüjpiw* /t-a-ndüjp-iw/ 'They followed' from the stem *ndüüüb* /ndü:b/ 'follow'. When adding a suffix of one syllable, the long vowel in the stem is shortened /üü/ → /üj/ and aspirated and as a result of the vowel no longer being long, the final stop appears as /p/ (/b/ appears after long vowels).

covering the process that happens when a single-syllable suffix is added to a stem with a long vowel.

5 Results

The output in Figure 3 is the result of running a sentence from Herrera Castro (2010) through the morphological analyser. The remainder of this section describes the evaluation procedure.

5.1 Corpora

There are few published texts that are available online in text format. The largest available text is *Jayats nanderac wüx miteatiüüts Jesucristo* (Wycliffe Bible Translators, 2009), the translation of the New Testament. This consists of over 230k tokens and was produced over a number of years. It was used in the development of the analyser for checking forms and for determining which stems to add to the lexicon.

To evaluate the text on unseen data, we obtained some other texts from a range of domains, but none consisting of more than 20,000 tokens. The texts we used were two texts from the domain of education: (1) a book for teaching adult literacy, *Sateow at sarang nawiig wüx ximbeay* (INEA, 2014).[3] and (2) a Licenciate thesis on education (Buenavista, 2012). The second two texts were smaller and more literary, a collection of stories in Huave collected by the Summer Institute of Linguistics (Stairs and

[3]The title translates into English as "I read and write in my language".

```
^Pe/pero<cnjcoo>$
^tasaj/<past><sg3>saj<v><tv>$
^mikwal/<px2sg>kwal<n>/<px3sg>kwal<n>$
^nej/nej<prn><pers><sg3>$
^natangüy/<rel>tang<v><iv><refl>$
^nüx/nüx<n>$
^este/este<part>$
^marang/<subj><sg3>rang<v><tv>$
^che/che<evid>$
^najngow/najngow<n>$
^./.<sent>$
```

Figure 3: Example output from the analyser for the sentence *Pe tasaj mikwal nej natangüy nüx este marang che najngow.* "But he told his older daughter to make soup." (lit. But he said to his grown daughter that she make soup.). Each line is a token, the part before the / is the surface form and following that is a list of potential analyses, each separated by a slash. There is one ambiguous token, the word *mikwal* < *kwal* 'child of someone' which with the *mi-* prefix could be analysed as either second person singular (POSS.2SG) or third person singular (POSS.3SG). The word could be disambiguated in a subsequent step by the following third person pronoun, *nej*.

Stairs, 2006) and a collection of poetry by Raúl Rangel González entitled *Andeak xemeaats* (Rangel González, 2020).[4]

5.2 Naïve coverage

Our first method of evaluation was to calculate the naïve coverage and mean ambiguity on freely available corpora. Naïve coverage refers to the percentage of surface forms in a given corpora that receive at least one morphological analysis. Note that forms counted by this measure may have other analyses which are not delivered by the transducer. The mean ambiguity measure was calculated as the average number of analyses returned per token in the corpus. The results can be found in Table 2. Each corpus was first split into 10 equal-sized parts and then the coverage was calculated and the mean and standard deviation obtained.

5.3 Precision and recall

We tested the precision and recall by using a test corpus created from the glossed example sentences in Herrera Castro (2010). We made a unique wordlist from these sentences and selected 100 words at random. We then passed these through out analyser and edited the output to include all the analyses found in the glosses — that is, if there was a gloss in the text for which the corresponding analysis was not present, we added the analysis. This gave a list of words where each word had all of its possible analyses according to the glossed texts (we call this the gold standard).

We define true positives, tp, as those analyses which were in both the transducer's output and in the gold standard list of analyses. We define false positives, fp, as those analyses that were in the transducer output but not in the gold standard list of analyses. And we define false negatives, fn as those analyses which were in the gold standard list, but not in the transducer output. Incorrect analyses by the transducer are counted as false positives as we consider them to be incorrect hypotheses generated by our transducer about a form in the language and to distinguish them from false negatives. False negatives are typically caused by something missing in the implementation or lexicon, while false positives are typically caused by incorrect implementation or incorrect lexicon or category assignment.

Tokens which received no analyses were counted as false positives. We defined precision, P (1), recall, R (2) and F_1-score (3).

$$P = \frac{\text{tp}}{(\text{tp} + \text{fp})} \tag{1}$$

$$R = \frac{\text{tp}}{(\text{tp} + \text{fn})} \tag{2}$$

$$F_1 = 2\frac{PR}{P + R} \tag{3}$$

Intuitively, precision is the likelihood of an analysis presented by the transducer being an analysis found in the gold standard, while recall is the likelihood of an analysis found in the gold standard being in the transducer. Table 3 presents the results of the evaluation.

Note that this method is only an approximation of the precision and recall of the analyser as the corpus may not contain all valid analyses for a given token.

6 Future work

In its current state, the analyser handles most of the inflectional processes of the language and some derivation. The lexicon is still very small and we would like to expand it, to include at least all the stems in the two dictionaries we have available.

There are two processes which are not well supported currently, the first is infixation of *-ra-* for forming the passive with some transitive

[4] The title translates into English as "The voice of my spirit".

Corpus	Genre	Tokens	Coverage (%)
Jayats nanderac wüx miteatiiüts Jesucristo	Religion	235,267	92.74 ± 0.16
Sateow at sarang nawiig wüx ximbeay	Education	18,490	82.49 ± 1.82
Licenciate thesis (Buenavista, 2012)	Education	18,343	82.05 ± 0.44
Cuentos Huaves II	Fiction	5,793	91.63 ± 1.00
Andeak xemeaats	Poetry	2,235	82.54 ± 2.75
Total:		280,307	91.26 ± 0.16

Table 2: The naïve coverage of the analyser over the available text corpora. The coverage is calculated over ten equal size splits and standard deviation is given. It is clear that the coverage is greatest and the estimate most reliable on the New Testament, which is also our largest text. However it is also noteworthy that the coverage over the fiction texts is also over 90%.

	Precision	Recall	F_1-score
All	70.1	64.7	67.3
Known	90.4	81.5	85.7

Table 3: Precision, recall and F-score for the test set. The metrics are substantially higher for the set of words that are known as they are ones where both a stem exists in the lexicon and the morphotactics are implemented as these are generally cases of missing alternative stems, for example a noun stem is present in the lexicon for a given surface form, but not a verb stem which generates the same surface form.

verbs, it is not clear from the literature what morpho/phonotactic restrictions there are. The one example that we found (7) is functional, but the solution has yet to be extended to other stems.

(7) a. *Aaga pet atsamb a miüs.*
aaga pet a-tsamb a miüs
DET dog SG3-bite the cat
'The dog bites the cat.'

b. *Tatsaramb a miüs*
t-a-tsa⟨ra⟩mb the cat
PAST-SG3-bite⟨PASS⟩bite the cat
'The cat is bitten.'

Another process is reduplication, examples of this have so far been lexicalised, but other solutions exist. It remains to be seen how productive the process is in texts (from our 100-token test set there was one word exhibiting reduplication).

Object agreement (as found in other languages of Mesoamerica such as Nahuatl or the Mayan languages) is also found in Huave but it is very reduced compared to other languages, only appearing in clauses where there is a singular subject and plural object, for example in (8) from Herrera Castro (2010).

(8) *Tanajawüw a pet*
ta-na-jaw-üw a pet
PAST-S.SG1-see-O.PL3 the dog
'I saw the dogs.'

This last phenomenon has yet to be implemented, but we anticipate it being straightforward.

7 Concluding remarks

We have presented the first morphological analyser for San Mateo Huave, a language isolate spoken in the state of Oaxaca in the south of Mexico. The analyser comprises a finite-state transducer based on the Helsinki Finite-State Tools. It covers a reasonable percentage — 80–93% of tokens in running text over a number of freely available corpora of San Mateo Huave. The analyser is available as free/open-source software under the GNU General Public Licence.[5]

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments.

References

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Studies in Computational Linguistics. CSLI Publications, Stanford, California.

Francisco Belmar. 1901. *Estudio del huave*. Instituto Nacional de Antropología e Historia, Oaxaca.

Hugo Alberto Hidalgo Buenavista. 2012. Makiaacheranüw wüx namix mongich ikoots matajküw majneaj andeaküw ombeay mol tiül noik nendüy najiüt meawan, monjüy amb anayiw neat akiaacheyej niüng monkiajchay nenüt Moisés Sáenz najlüy tikambaj,

[5] https://github.com/apertium/apertium-huv/

kiaj kawak okwiaj latiük, tiül miiüt pajtiam. Master's thesis, Escuela Normal Bilingüe e Intercultural de Oaxaca.

Emily Chen and Lane Schwartz. 2018. A morphological analyzer for St. Lawrence Island / Central Siberian Yupik. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Samuel Herrera Castro. 2010. Alineamiento y frase verbal en Huave de San Mateo Del Mar, Oaxaca. Master's thesis, Escuela Nacional de Antropología e Historia, México, D.F.

Samuel Herrera Castro. 2016. *Sintaxis y semántica de la frase nominal en Huave de San Mateo Del Mar, Oaxaca*. Ph.D. thesis, Colegio de México, México, D.F.

INALI. 2016. Atlas de las lenguas indígenas de México. Https://atlas.inali.gob.mx/.

INEA. 2014. *Sateow at sarang nawiig wüx ximbeay. Ombeayiüts*.

INEGI. 2020. Censo de población y vivienda 2020. Technical report, Instituto Nacional de Estadística y Geografía. https://www.inegi.org.mx/programas/ccpv/2020/.

Adriano M. Ingunza, John E. Miller, Arturo Oncevay, and Roberto Zariquiey. 2021. Representation of Yine (Arawak) morphologyby finite state transducer formalism. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, page 102–112. Association for Computational Linguistics.

Lauri Karttunen. 1993. Finite-state constraints. In John Goldsmith, editor, *The Last Phonological Rule: Reflections on constraints and derivations*. University of Chicago Press.

Kimmo Koskenniemi. 1983. *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. Ph.D. thesis, Helsingin yliopisto.

Krister Lindén, Erik Axelson, Sam Hardwick, Tommi Pirinen, and Miikka Silfverberg. 2011. HFST—framework for compiling and applying morphologies. *Communications in Computer and Information Science*, 100:67–85.

P. H. Matthews. 1972. Huave verb morphology: Some comments from a non-tagmemic viewpoint. *International Journal of American Linguistics*, 38(2):96–118.

Sarah Moeller, Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2018. A neural morphological analyzer for arapaho verbs learned from a finite state transducer. In *Proceedings of Workshop on Polysynthetic Languages*, pages 12–20.

Rolf Noyer. 1993. Mobile affixes in huave: Optimality and morphological well-formedness. In *Proceedings of the twelfth west coast conference on formal linguistics*, pages 67–82, Stanford. CSLI.

Rolf Noyer. 2012. Diccionario etimológico y comparativo de las lenguas huaves. https://www.ling.upenn.edu/~rnoyer/DECH_August2012.pdf.

Rolf Noyer. 2013. Generative phonology of san mateo huave. *International Journal of American Linguistics*, 79(1):1–60.

Marjorie Pak. 2014. Phonological evidence for the syntax of VOS and SVO in Huave. In *Proceedings of the Workshop on the Sound Systems of Mexico and Central America*.

Robert Pugh, Francis Tyers, and Marivel Huerta Mendez. 2021. Towards an open source finite-state morphological analyzer for zacatlán-ahuacatlán-tepetzintla nahuatl. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 80–85.

Raúl Rangel González. 2020. *Andeak xemeaats*. La voz de mi espíritu.

Ivy Richardson and Francis M. Tyers. 2021. A morphological analyser for K'iche'. *Procesamiento de Lenguaje Natural*, 66:99–109.

Emily Stairs and Barbara Hollenbach. 1969. Huave verb morphology. *International Journal of American Linguistics*, 35(1):38–53.

Emily Stairs and Elena Hollenbach. 1981. *Diccionario huave de San Mateo del mar*, chapter Gramática huave. Instituto Lingüístico de Verano, México, D.F.

Glen Stairs and Emily Stairs. 1981. *Diccionario huave de San Mateo del Mar*. Instituto Lingüístico de Verano, México, D.F.

Glen Stairs and Emily Stairs. 1983. Huave de San Mateo Del Mar, Oaxaca. Technical report, Centro de Investigación para la Integración Social, México, D.F.

Glen Stairs and Emily Stairs. 2006. *Cuentos Huaves II*. Instituto Lingüístico de Verano, México, D.F.

Jonathan Washington, Ilnar Salimzyanov, and Francis Tyers. 2014. Finite-state morphological transducers for three Kypchak languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 3378–3385, Reykjavik, Iceland. European Languages Resources Association (ELRA).

Wycliffe Bible Translators. 2009. *Jayats nanderac wüx miteatiiüts Jesucristo*. Wycliffe Bible Translators.