A Morphological Analyzer for Shipibo-Konibo

Ronald Cardenas
Charles University in Prague
Faculty of Mathematics and Physics
Inst. of Formal and Applied Linguistics
ronald.cardenas@matfyz.cz

Daniel Zeman
Charles University in Prague
Faculty of Mathematics and Physics
Inst. of Formal and Applied Linguistics
zeman@ufal.mff.cuni.cz

Abstract

We present a fairly complete morphological analyzer for Shipibo-Konibo, a low-resourced native language spoken in the Amazonian region of Peru. We resort to the robustness of finite-state systems in order to model the complex morphosyntax of the language. Evaluation over raw corpora shows promising coverage of grammatical phenomena, limited only by the scarce lexicon. We make this tool freely available so as to aid the production of annotated corpora and impulse further research in native languages of Peru.

1 Introduction

Linguistic and language technology research on Peruvian native languages have experienced a revival in the last few years. The academic effort was accompanied by an ambitious long term initiative driven by the Peruvian government. This initiative has the objective of systematically documenting as many native languages as possible for preservation purposes (Acosta et al., 2013). So far, writing systems and standardization have been proposed for 19 language families and 47 languages.

In this paper, we focus on Shipibo-Konibo (henceforth, SK), also known in the literature as Shipibo or Shipibo-Conibo. SK is a member of the Panoan language family. This family is a well-established linguistic group of the South American Lowlands, alonside Arawak, Tupian, Cariban, and others. Currently, circa 28 Panoan languages are spoken in Western Amazonia in the regions between Peru, Bolivia, and Brazil. Nowadays, Shipibo is spoken by nearly 30,000 people mainly located in Peruvian lands.

The morphosyntax of SK in extensively analyzed by Valenzuela (2003). However, several phenomena such as discourse coherence marking and ditransitive constructions still require deeper understanding, as pointed out by Biondi (2012).

We present the first finite-state morphological analyzer for SK, capable of performing POS tagging as well as morpheme segmentation and categorization. In order to impulse the development of downstream applications and corpora annotation, the tool is freely available[1] under the GPL license.

2 Related Work

The development of freely available basic language tools has proven to be of utmost importance for the development of downstream applications for native languages with low resources. Finite-state morphology systems constitute one type of such basic tools. Besides downstream applications, they are essential for the construction of annotated corpora, and consequently, for development of other tools. Such is the case of Quechua, a native language spoken in South America, for which the robust system developed by (Rios, 2010) paved the way to the proposal of a standard written system for the language (Acosta et al., 2013) and impulsed work in parsing, machine translation (Rios, 2016), and speech recognition (Zevallos and Camacho, 2018).

Initial research regarding SK has centered in the development of manual annotation tools (Mercado-Gonzales et al., 2018), lexical database creation (Valencia et al., 2018), Spanish-SK parallel corpora creation and initial machine translation experiments (Galarreta et al., 2017). Related to our line of research, work by Pereira-Noriega et al. (2017) addresses lemmatization but not morphological categorization. Alva and Oncevay-Marcos (2017) presents initial experiments on spell-checking using proximity of morphemes and syllable patterns extracted from anno-

[1] http://hdl.handle.net/11234/1-2857

tated corpora.

In this work, we take into account the morphotactics of all word categories and possible morpheme variations attested by Valenzuela (2003). We explored and included as many exceptions as found in the limited annotated corpora to which we got access. Hence, the tool presented is robust enough to leverage current efforts in the creation of basic language technologies for SK.

3 Shipibo-Konibo Morphosyntax

In terms of a syntactic profile, SK is a (mainly) post-positional and agglutinating language with highly synthetic verbal morphology, and a basic but quite flexible agent-object-verb (AOV) word order in transitive constructions and subject-verb (SV) order in intransitive ones, as summarized by (Fleck, 2013).

SK usually exhibits a biunique relationship between form and function, and in most cases morpheme boundaries are easily identifiable. It is common to have unmarked nominal and adjectival roots, and few instances of stem changes and suppletion are documented by (Valenzuela, 2003). In addition, the verb may carry one or more deictic-directive, adverb type suffixes, in what can be described as a polysynthetic tendency.

In addition, SK presents a rare instance of syntactic ergativity in an otherwise morphologically ergative but syntactically accusative language.

We proceed to comment about the most salient morpho-syntactic features relevant to the morphotactics argumentation in section 4.2. The examples presented in this section were taken from Valenzuela (2003).

3.1 Expression of Argument

Verb arguments are expressed through free lexical case-marked nominals, with no co-referential pronominal marking on the verb or auxiliary. That is, verbs and auxiliaries are not marked to agree with 1st, 2nd, or 3rd person of the subject or agent. Instead, verbs are marked to indicate that the action was carried out by the same participant of the previous clause or by another one. We explain this phenomena in section 3.4.

Omission of required subject and object is normally understood as zero third person singular form. There are no systematic morpho-syntactic means of distinguishing direct from indirect objects, or primary versus secondary objects.

3.2 Case Marking

Grammatical cases are always marked as suffixes, except for a couple of exceptions. SK exhibits a fairly rigid ergative-absolutive case-marking system. The ergative case is always marked, whereas the absolutive case is only marked on non-emphatic pronouns. All other grammatical cases are marked, except the vocative case. The vocative case is constructed by shifting the stress of a noun to the last syllable.

3.3 Participant Agreement

Certain adverbs, phrases, and clauses are semantically oriented towards one core participant or controller and receive a marking in accordance with the syntactic function this participant plays, namely *subject (S)* of a intransitive verb, *agent (A)* of a transitive verb, or *object (O)* of a transitive construction. This feature can be analyzed as a type of split-ergativity which might be exclusive to Panoan languages. The following example illustrates this phenomena for the adjunct *bochiki: high up* in S, O, and A orientation (ONOM refers to onomatopeic words).

(1) S orientation
Bochiki-ra e-a oxa-i
up:S-Ev 1-Abs sleep-Inc
"I sleep high up (e.g., in a higher area inside the house)."

(2) O orientation
E-n-ra yami kentí *bochiki* a-ke
1-Erg-Ev metal pot:Abs up:O do.T-Cmpl
"I placed the metal pot high up." (only the pot is high up)

(3) A orientation
E-n-ra yami kentí *bochiki-xon*
1-Erg-Ev metal pot:Abs up-A
tan tan a-ke.
ONOM ONOM do.T-Cmpl
"I hit the metal pot (being) high up." (I am high up with the pot)

3.4 Clause-Chaining and Switch-Reference System

Chained clauses present only one clause with fully finite verb inflection while the rest of them carry same- or switch-reference marking. Reference-marked clauses are strictly verb-final, carry no obvious nominalizing morphology and may precede, follow, or be embedded in their matrix clause.

Same-reference markers encode transitivity status of the matrix verb, co-referentiality or non co-referentiality of participant, and relative temporal or logical order of the two events. This is because most same-subject markers are identical to the participant agreement morphemes and hence correlate with the *subject (S)* or *agent (A)* function played by their controller in the matrix clause. The following example shows three chained clauses. Notice that the matrix verb is *chew*, and the subordinated clause's verbs carry the marker *xon* to indicate that the action was performed by the same agent prior to the action described in the main clause (PSSA: previous event, same subject, *A* orientation).

[[Jawen tapon bi-xon] kobin-a-xon]
Pos3 root:Abs get-PSSA boil-do.T-PSSA
naka-kati-kan-ai.
chew-Pst4-Pl-Inc
"After getting its (i.e., a plant's) root and boiling it, they chewed it."

Same- or switch- reference marking may also be used to encode different types of discourse (dis)continuity.

3.5 Pronouns and Split-Ergativity

The personal pronoun system in SK is composed of 6 basic forms corresponding to the combinations of three person (1,2,3) and two number (singular and plural) distinctions. SK does not differentiate gender or inclusive vs exclusive first person plural. There are no honorific pronouns either.

The ergative-absolutive alignment is used in all types of constructions, except for reflexive pronoun constructions. Reflexive pronouns are marked with the suffix *-n* when referring to both A and S arguments, but remain unmarked when referring to an O argument. Hence, reflexive pronouns constructions clearly present a nominative-accusative alignment.

3.5.1 Clitics

All clitics in SK are enclitics, i.e. they always function as suffixes, but most of them encode clause level features in which case they are attached to the last element of the phrase or clause they are modifying. SK clitics are categorized into case markers, *less-fixed clitics* and *second position clitics*, as proposed by Valenzuela (2003).

Case markers are attached to noun phrases preceding mood and evidentiality markers in its last constituent word.

Second position clitics are attached to the main clause in the sentence, and they encode evidentiality (+Ev:ra; +Hsy:ronki, ki; e.g. *it is said that ...*), reported speech (e.g. *he says/said that ...*), interrogative focus (+Int:ki,rin; +Em:bi), and dubitative voice.

Less-fixed clitics mark the specific element they are attached to, instead of the whole clause. These are endo-clitics, i.e. they can take any position other than the last morpheme slot in a construction. In this category we can find adverbial, adjectival, and dubitative suffixes.

4 Morphological Analyzer

The analyzer was implemented using the Foma (Hulden, 2009) toolkit, following the extensive morphological description provided by Valenzuela (2003). Besides segmenting and tagging all morphemes in a word form, the analyzer also categorizes the root and the final token in order to account for any sequence of derivational processes. The analysis is of the form

[POS] root[POS.root] morpheme[+Tag] ...

and it is illustrated with an example in Table 1.

The complete list of abbreviations and symbols used for morphological tagging can be found in the Appendix A of (Valenzuela, 2003). Language specific POS tagset was mapped to the Universal Dependencies (Nivre et al., 2016) v2 POS tagset.[2]

In the remaining of this section we provide a thorough explanation of the production rules for the main POS categories and the comment on the limitations of the analyzer.

4.1 The Lexicon

The lexicon was obtained from manually annotated corpus and a digitalized thesaurus kindly provided by the Artificial Intelligence Research Lab of the Pontifical Catholic University of Peru (GIPIAA-PUCP). The annotated corpus was built from folk tales documents and it consists of 12,250 tokens and 2,915 types. The thesaurus provides dictionary entries for 6,750 types.

The extensive work of (Valenzuela, 2003) provides a systematic encoding of morpho-syntactic information for SK. Similar guidelines were followed to design the encoding for Quechua (Rios, 2016), another agglutinative, ergative-absolutive

[2] http://universaldependencies.org/u/pos/

Token	Translation	Analysis
Isábora	the birds	[NOUN] isá[NRoot] bo[+Pl] ra[+Ev]
noyai	are flying	[VERB] noy[VRoot.I] ai[+Inc]

Table 1: Example of analysis produced.

UPOS	Thesaurus	Annotated corpus
NOUN	2557	719
VERB	2284	578
ADJ	601	107
ADV	223	112
PROPN	-	112
PRON	24	36
NUM	-	6
SCONJ	-	2
CCONJ	-	3
AUX	-	2
DET	-	28
ADP	46	19
INTJ	-	15
PART	-	9

Table 2: Number of roots per POS, for each lexicon source.

native language widely spoken in Peru and South America.

The annotated corpus, however, was not annotated following this encoding, and further manual annotation was required. With the help of a digitalized dictionary and an affix thesaurus we manually resolved the mappings and correspondences using the—now widely accepted—morphosyntactic encoding.

The following example illustrates the annotation. The first row shows the raw segmentation of the tokens; the second row, the original annotation (*Clit* stands for clitic, *VS stands for verbal suffix*); the third row, the new annotation following the morphosyntactic tagset proposed by Valenzuela (2003).

Shoko-res	oxa-[a]i	/ pi-ai.
a.little.bit-Clit	sleep-VS	eat-VS
a.little.bit-just	sleep-Inc	eat-Inc

'I'm gonna sleep / eat just a little bit.'

Table 2 presents the number of roots per UD POS category for each lexicon source, for a total of 8,658 roots.

4.2 Morphotactics

Although SK presents a predominantly suffixed morphology, there exists a closed list of prefixes, almost all being body part derivatives shortened from the original noun (e.g. 'head' *mapo* $\rightarrow$ *ma*). These prefixes can be added to nouns, verbs, and adjectives to provide a locative signal.

4.2.1 Nouns

Nominal roots can occur in a bare form without any additional morphology or carry the following morphemes.

- Body part prefix (+Pref), to indicate location in the body.
- Plural marker (+Pl:bo), meaning more than one. Dual number distinction is not made in nouns, but in verbs.
- N-marker and other case markers. The suffix *-n* can mark the ergative (+Erg), genitive (+Gen), and interessive (+Intrss, to denote interest), and instrumental (+Inst) cases. Other marked cases in SK include absolutive (+Abs:a), dative (+Dat:ki), locative (+Loc:me,ke), allative (+All:n,nko), ablative (+Abl:a), and chezative (+Chez:iba). The allative case always follows a locative case marker, both of them presenting several allomorphs.
- Participant agreement marker (+S:x), to indicate the subject of a transitive verb.
- Distributive marker (+Distr:tibi), produces quantifier phrases, e.g. day+Distr $>$ *every day*.
- Adjectival markers , such as diminutive (+Dim:shoko), deprecatory (+Deprec:isi), legitimate (+Good:kon, +Bad:koma), proprietive (+Prop:ya) and privative (+Priv:oma,nto).
- Adverbial markers.
- Postpositional markers.

- Second position clitics, exclusively the focus emphasizer (+Foc:kan).

It is worth mentioning that only the first plural morpheme has precedence over the others suffixes, and clitics are required to be last. Plural, cases, and adverbial markers can occur multiple times. There is no gender marking in SK. Instead, the words for woman (*ainbo*) and man (*benbo*) are used as noun modifiers. Consider the example

Títa-shoko-bicho-ra oxa-ai
mom:Abs-Dim-Adv-Ev sleep-Inc
'Mommy sleeps alone.'

The diminutive *shoko* is denoting affection instead of size. Notice that the adverbial suffix *bicho* would have to be constructed as a separate adjunct in English and it is attached to the noun, not the verb.

Derived Nominals Verbal roots can be nominalized by adding the suffix *-ti* or past participle suffixes *a, ai*. Zero nominalization is only possible over a closed set of verbs, e.g. *shinan- 'to think, to remember / mind, thinking'*.

On the other hand, adverbial expressions and adjectives may function as nominals and take the corresponding morphology directly without requiring any overt derivation.

4.2.2 Adjectives and Quantifiers

Adjectival roots can optionally bear the following morphemes.

- Negative (+Neg:ma), to encode the opposite feature of an adjective.
- Diminutive (+Dim:shoko), deprecatory (+Deprec:isi), intensifier (+Intens:yora).
- Adverbial markers.
- Interrogative clitics (+Int:ki,rin; +Em:bi).

Derived Adjectives Nominal roots can be adjectivized when adding proprietive (+Prop:ya) or privative (+Priv:oma,nto) markers, e.g. *bene-ya* [husband+Prop] $\rightarrow$ *married (woman)*.

In regards to verbs, participial tense-marked verbs can function as adjectives. Transitive verbs and a closed set of intransitive verbs can take an agentive suffix (+Agtz:mis,yosma,kas) to express *one who always does that action*.

As with nominalization, adverbs take zero morphology to function as adjectives.

4.2.3 Verbs

Verbal morphology presents by far the most complex morphotactics in SK, allowing up to 3 prefixes and 18 suffixes following a relatively strict order of precedence, as follows.

- Prefixes related to body parts, providing locative information about the action.
- Plural marker (+Pl:kan).
- Up to 2 valency-changing suffixes, depending whether we are increasing or decreasing transitivity, whether the root is transitive or intransitive, or whether the root is bisyllabic or not.
- Interrogative intensifier (+Intens:shaman), to bring focus on the action in a question.
- Desiderative marker (+Des:kas), to indicate that the clause is desiderative (e.g. *I want to* V).
- Negative marker (+Neg:yama).
- Deictive-directive markers are identical or similar to motion verbs and encode a movement-action sequence, e.g. *V-ina* $\rightarrow$ 'go up the river and V'.
- Adverbial suffixes, depending whether the verb is marked as plural or not. Here in this slot we find the suffix *bekon* that indicates dual action.
- Habitual marker (+Hab:pao), to encode that the action is done as a habit.
- Tense markers.
- Adjectival (+Dim:shoko; +Deprec:isi; +Intens:yora) and adverbial suffixes.
- Preventive marker (+Prev:na), to express warning, a situation to be prevented.
- Final markers, including participial and reference markers depending whether the verb is finite or non-finite in the clause. Reference markers encode agreement with the agent or subject of the clause (S vs A agreement), whether it is even the same agent and the point in time the action was carried out.
- All second position clitics.

Verbal roots must always bear either a tense marker or at least one final marker. All other suffixes are optional. The following example illustrates how the deictive-directive marker can encode a whole subordinated clause.

Sani betan Tume bewa-kan-*inat*-pacho-ai
Sani and Tume sing-Pl-go.up.the.river-Adv-Inc
'Sani and Tume always sing while going up the river.'

Derived Verbs Nominal roots are turned into transitive verbs by adding the causativizer +Caus:n. The auxiliary marker +Aux:ak can be added to nominal, adjectival, and adverbial roots to form transitive verbs.

4.2.4 Pronouns

Personal pronouns can bear the following suffixes.

- Ergative (+Erg:n) and absolutive (+Abs:a) case marker. This last one is only used on singular forms and first person plural.
- Chezative (+Chez:iba), dative (+Dat:ki), and comitative (+Com:be) case markers.
- Post-positional suffixes.
- Interrogative and evidential clitics.

The ergative case construction also renders possessive modifiers, with the exception of the first and third singular form, which have a different form with no marking. Possessive pronouns are formed by adding the nominalizer +Nmlz:a to possessive modifiers.

Emphatic pronouns present the marker +S:x when agreeing with the S argument and no marker when agreeing with the A argument. Special attention was taken for the third person singular pronoun *ja-*, which presents a tripartite distribution: *ja-n-bi-x* for S, *ja-n-bi* for A, *ja-bi* for O.

Interrogative pronouns *who, what, where* can be marked for ergative, absolutive, genitive, chezative, and comitative cases. The participant agreement suffix for these pronouns presents a tripartite distribution: +S:x, +O:o, +A:xon for S, O, A agreement, respectively. The following example illustrates the behavior of pronoun *jawerano*: where.

(4) S orientation
Jawerano-a-x-ki mi-a jo-a
where:Abl-S-Int 2-Abs come-Pp2
'From where did you come?'

(5) O orientation
Jawerano-a-ki mi-n paranta be-a
where:Abl-O-Int 2-Erg banana:Abs bring-Pp2
'From where did you bring banana?'

(6) A orientation
Jawerano-xon-ki epa-n pi-ai
where-A-Int uncle-Erg eat-Pp1
'Where is uncle eating?'

Interrogative pronouns *how, how much, how many* are marked only for participant agreement using an ergative-absolutive distribution (+S:x, +A:xon). In addition, all interrogative pronouns can take interrogative, focus, and emphasis clitics.

Demonstrative roots can function both as pronouns and determiners. In the first case, they bear all proper pronoun morphology. In the second case, they can only bear the Plural nominal marker +Pl:bo.

4.2.5 Adverbs

Adverbs can be suffixed with evidential clitics. However, whenever an adverb is modifying an adjective, it takes participant agreement morphology (+S:x,ax,i; +A:xon) in order to agree with the syntactic function of the noun the adjective is modifying.

Adverbial roots can also function as suffixes and be attached to nouns, verbs, adjectives, and even other adverbial roots.

Derived Adverbs Adverbs can be derived from demonstrative roots by adding locative case markers depending of the proximity of the entity being referred to. Adjectival roots function as adverbs by receiving the +Advz:n morpheme. Nouns and quantifier roots take the locative case marker +Loc:ki in order to form adverbs.

4.3 Postpositions

There are only 20 postpositional roots in SK, all of them can take second position clitics. In the same fashion as adverbial roots, postpositional roots can also function as suffixes. Adverbial roots can function as postpositions by taking the locative marker sequence +Loc: ain-ko.

4.3.1 Conjunctions

All conjunction roots take participant agreement markers (+S:x, +A:xon), except coordinating conjunctions *betan* (and) and *itan* (and, or). These

markers encode inter or intra-clausal participant agreement, often used as discourse discontinuity flags.

Subordinating conjunctions can take the following morphemes.

- Locative, ablative, and similitive (+Siml:ska) case markers.
- Completive aspect markers, also found as participials in verbs at the *final* slot.
- Reference agreement mark +P:ke, to encode discourse continuity.
- Second position clitics.

In the following example, we analyze the behavior of the conjunction root *ja*.

(7) *Ja-tian* jawen bene ka-a ik-á
that-Temp Pos3 husband:Abs go-Pp2 be-Pp2
iki jato onan-ma-i ...
AUX 3p:Abs know-Caus-SSSI ...
'By that moment her husband had gone to teach them (i.e. the Shipibo men) ...'

(8) Jo-xon jis-á-ronki ik-á iki
come-PSSA notice-Pp2-Hsy be-Pp2 AUX
Inka Ainbo wini wini-i.
Inka woman:Abs cry cry-SSSI
'When (he) returned, he saw the Inka Woman crying and crying.'

(9) *Ja-tian* jawen bene-n raté-xon
3-Temp Pos3 husband-Erg scare:Mid-PSSA
yokat-a iki: "Jawe-kopí-ki mi-a wini-ai?"
ask-Pp2 AUX why-Int 2-Abs cry-Inc
"Then her husband got scared and asked (her): 'Why are you crying?'"

While the first instance of *jatian* in (9) coincides with the introduction in subject function of the male Inka and hence with a change of subject, the second instance in (11) does not. In fact, the subjects in (10) and (11) have the same referent, but *jatian* is used to indicate a switch from narrative to direct quote in the chain. Note that in (11) the subject 'her husband' is overtly stated so that the hearer does not misinterpret *jatian* as indicating a change in subject.

4.4 Limitations

The analyzer processes token by token without considering context, restricting it from discarding hypothesis based on fairly rigid constructions, e.g. future tense with auxiliary verbs, modal verbs, nominal compounds, among others.

There exist a group of morphemes that present multiple possible functions in the same position of the construction template. Hence, they can be mapped to more than one morphological tag. Consider the case suffix *-n* in the following example. The square brackets indicate that even though *-n* is attached to *nonti*, it acts as a phrase suffix that modifies the whole phrase (*you canoe*).

E-n [mi-n nonti]-n yomera-i ka-ai
1-Erg 2-Gen canoe-Ins get.fish-SSSS go-Inc
"I am going to fish with your canoe."

In this case, the analyzer outputs all possible tag combinations, such as +Erg:ergativo, +Inst:intrumental, +Gen:genitive, +Intrss:interessive, and +All:allative. Other suffixes with this kind of behavior are completive aspect suffixes and past tense suffixes in verbs. Disambiguation of these morphemes requires knowledge of the syntactic function of the word in the clause. Such sentence level disambiguation is out of the scope of the analyzer.

5 Evaluation

We evaluate the robustness of our analyzer by testing the coverage of word forms. A coverage per type of 94.99% was achieved for the training data (annotated corpus + thesaurus). A closer look into the remaining non-recognized types revealed that in all cases they contain an already covered root or affix but with different diacritization. This is to be expected since the only diacritization rules existent for SHK were proposed recently by Valenzuela (2003) and the text the annotated data was based in was written way before the proposal of the diacritization rules.

Table 3 shows type and token coverage over raw text not used during development. These corpora span several domains such as the bible, educational material, legal domain, and folk tales. This last domain—same as the domain of the annotated corpus—has the highest coverage.

As expected, the lowest coverage is obtained over the legal domain, a specialized domain with complex grammatical constructions and specialized vocabulary. For example, legal documents

	Number of Words		Coverage (%)	
Subset	Tokens	Types	Tokens	Types
Bible - New Testament	210,828	20,504	79.11	49.49
Elementary School Books	31,127	4,395	76.59	45.12
Kindergarten Text Material	15,912	2,581	76.90	55.29
Constitution of Peru	12,319	2,645	70.83	40.57
Folk tales	10,934	2,737	94.38	85.42
Total	281,120	28,133	78.93	47.12

Table 3: Coverage on corpora from different domains of raw corpora.

Error type	Count
Alternative spelling	43
Proper nouns	20
Common nouns	4
Other OOV	25
Foreign word	8

Table 4: Error analysis of the 100 most frequent unanalyzed word types in raw corpora.

must be precise about semantic roles of the participants, information partially encoded through morphology in SK.

In contrast, educational material for kindergarten level presents the second highest coverage, quite possibly because only basic grammatical constructions are used at this level of education.

Error Analysis: We further analyze the unrecognized words in the raw corpora. We manually categorize the 100 most frequent unrecognized word types, as shown in Table 4. It can be noted that the most common error is due to alternative spelling of the final word form, mostly due to the absence—or presence—of diacritics or due to the presence of an unknown allomorph. Most of the errors of this kind can be traced back to tokens in the Bible domain. The Bible was translated to SK in the 17th century and it has remained almost intact since then. Hence, some constructions are considered nowadays ungrammatical (e.g. a verb must always carry either a participant agreement suffix or a tense suffix) or some suffixes are obsolete (e.g. the n-form +Erg:*sen*; the infinitive form +Inf:*ati*).

Furthermore, the high presence of OOV words other than nouns or proper nouns is an indicative that the root lexicon upon the analyzer is based is still limited and far more entries are needed.

6 Conclusion and Future Work

We presented a robust and fairly complete (in morphotactics, not in lexicon) finite-state morphological analyzer for Shipibo-Konibo, a low-resourced native language from Peru. The analyzer is capable of performing morphological segmentation and categorization, as well as part-of-speech tagging of the root and the whole final token.

Experiments over corpora from different domains show promising coverage given the limited root lexicon available. We performed a thorough analysis of errors over unrecognized words, finding that our analyzer cannot recognize certain obsolete constructions and spellings found in Biblical text, which was written centuries ago. However, for modern day Shipibo-Konibo in non-specialized domains (e.g. legal domain) the tool is quite robust and covers production rules for all word categories.

The work presented in this paper is part of a greater effort to provide the research community with basic language tools that would aid in the construction of treebanks. Future paths considered include the mapping of morphological tags into morphological features defined in Universal Dependencies [3], sentence-level tag disambiguation and parsing, among others.

Acknowledgments

The work was partially supported by the grant 15-10472S of the Czech Science Foundation (GAČR). The authors would like to thank GRPIAA Research Lab at the Pontifical Catholic University of Peru for kindly providing the annotated corpus, dictionaries, and raw corpora used in the experiments of this paper.

[3] http://universaldependencies.org/u/feat/index.html

References

Sullón Acosta, Karina Natalia, Edinson Huamancayo Curi, Mabel Mori Clement, and Vidal Carbajal Solis. 2013. Documento nacional de lenguas originarias del Perú.

Carlo Alva and Arturo Oncevay-Marcos. 2017. Spell-checking based on syllabification and character-level graphs for a peruvian agglutinative language. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 109–116.

Roberto Zariquiey Biondi. 2012. Ditransitive constructions in Kashibo-Kakataibo and the non-distinguishable objects analysis. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 36(4):882–905.

David William Fleck. 2013. *Panoan languages and linguistics. (Anthropological papers of the American Museum of Natural History, no. 99)*. American Museum of Natural History.

Ana-Paula Galarreta, Andrés Melgar, and Arturo Oncevay-Marcos. 2017. Corpus creation and initial SMT experiments between Spanish and Shipibo-Konibo. In *Proceedings of RANLP*.

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: Demonstrations Session*, pages 29–32. Association for Computational Linguistics.

Rodolfo Mercado-Gonzales, José Pereira-Noriega, Marco Antonio Sobrevilla Cabezudo, and Arturo Oncevay-Marcos. 2018. Chanot: An intelligent annotation tool for indigenous and highly agglutinative languages in Peru. In *LREC*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association.

José Pereira-Noriega, Rodolfo Mercado-Gonzales, Andrés Melgar, Marco Sobrevilla-Cabezudo, and Arturo Oncevay-Marcos. 2017. Ship-lemmatagger: Building an NLP toolkit for a peruvian native language. In *International Conference on Text, Speech, and Dialogue*, pages 473–481. Springer.

Annette Rios. 2010. Applying finite-state techniques to a native American language: Quechua. *Institut für Computerlinguistik, Universität Zürich*.

Annette Rios. 2016. A basic language technology toolkit for Quechua.

Diego Maguiño Valencia, Arturo Oncevay-Marcos, and Marco Antonio Sobrevilla Cabezudo. 2018. Wordnet-shp: Towards the building of a lexical database for a peruvian minority language. In *LREC*.

Pilar Valenzuela. 2003. *Transitivity in shipibo-konibo grammar*. Ph.D. thesis, University of Oregon.

Rodolfo Zevallos and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern Quechua. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).