Toward Universal Dependencies for Shipibo-Konibo

Alonso Vasquez[1,3], **Renzo Ego Aguirre**[1], **Candy Angulo**[1], **John Miller**[1], **Claudia Villanueva**[1], **Željko Agić**[2], **Roberto Zariquiey**[1] and **Arturo Oncevay**[1]
[1] Dpto. de Humanidades and Dpto. de Ingeniería, Pontificia Universidad Católica del Perú
[2] Department of Computer Science, IT University of Copenhagen
[3] Department of Linguistics, University of California, Santa Barbara
arturo.oncevay@pucp.edu.pe

Abstract

We present an initial version of the Universal Dependencies (UD) treebank for Shipibo-Konibo, the first South American, Amazonian, Panoan and Peruvian language with a resource built under UD. We describe the linguistic aspects of how the tagset was defined and the treebank was annotated; in addition we present our specific treatment of linguistic units called *clitics*. Although the treebank is still under development, it allowed us to perform a typological comparison against Spanish, the predominant language in Peru, and dependency syntax parsing experiments in both monolingual and cross-lingual approaches.

1 Introduction and Background

Shipibo-Konibo is a language of the Panoan family spoken by around 35,000 native speakers in the Amazon region of Peru. It is a language with agglutinative processes, with a majority presence of suffixes and some clitics (neither a word nor an affix). Additionally, it presents word orders different from the dominant Spanish language.

To the best of our knowledge, there are no other Universal Dependencies (UD) treebanks for an indigenous language of South America, as surveyed by Mager et al. (2018). The closest resource is a treebank developed for a Quechuan variant; however, it was not designed under the UD guidelines (Rios et al., 2008). Another related case is the application of UD for the annotation of the native North American language Arapaho (Algonquian) (Wagner et al., 2016). Thus, Shipibo-Konibo would be the first South American indigenous language with this kind of computational resource[1].

Natural Language Processing (NLP) efforts for Shipibo-Konibo have developed a POS-tagger, a lemmatizer, a spell-checker, and a machine translation prototype with Spanish as the paired language (Mager et al., 2018). Each functionality has been published alongside its annotated corpus. A UD treebank would enhance the NLP toolkit for the language, as it is the core element for being able to train a dependency parser.

This paper describes the steps and decisions made towards a UD treebank for Shipibo-Konibo. First, §2 presents the annotation process. Then, §3 details the information of the UD treebank itself, such as the POS tags, morphological features and dependency relations, including the specific ones for Shipibo-Konibo. Moreover, it describes relevant decisions regarding clitics and word segmentation, including an analysis of the generated multiword tokens. Finally, we take advantage of the built treebank, and perform a typological comparison against Spanish in §4, as well as dependency parsing tests for monolingual and cross-lingual scenarios in §5.

2 Treebank Annotation

The annotation workflow of the Universal Dependencies (UD) treebank for Shipibo-Konibo is described in §2.1. In particular, specific consideration has been given for word segmentation with respect to clitics, which is detailed in §2.2.

2.1 Annotation Workflow

Annotation followed a sequential flow:

1. To annotate Shipibo-Konibo corpus in ChAnot (Mercado et al., 2018) and BRAT (Stenetorp et al., 2012). The former tool was specifically used for the morpheme segmentation of raw text into prefixes, root morphemes and suffixes in appropriate morphological detail. The provided interface with BRAT allows the graphical

[1] The treebank will be available for the next UD release

annotation of syntactic information over the segmentation. We used part of speech and relation names determined prior to the decision to conform to UD v2.0.

2. To compile segmented corpus into UD v2.0 format: Gather all annotations from ChAnot and BRAT into single file in UD v2.0 format. Compress detail segmentation of prefixes and suffixes to only segment on clitic boundaries. Add clitic features, and convert non-standard to UD v2.0 standard universal POS and dependency relation notation.

2.2 Clitics and Segmentation

In terms of its morphological profile, Shipibo-Konibo favors synthetic word formations. That is, in Shipibo-Konibo, words are often composed of a root and one or more bound morphemes. Some of these morphemes may be considered *clitics*, linguistic elements that do not fit either the prototype of word or that of affix. Similar elements are labelled *particles* in the Universal Dependencies tradition, but we prefer *clitics*, following the arguments presented in Zwicky (1977, 1985). In the Panoan literature, these intermediate linguistic units have also been called *clitics* (Fleck, 2013; Valenzuela, 2003; Zariquiey, 2015), so we consider it appropriate to follow this terminology in the development of our Shipibo-Konibo treebank.

As *clitics*, these linguistic units exhibit some features that resemble those attested in words. This intermediate nature clashes with the dichotomic division between morphology and syntax, in which linguistic units belong to one of these domains (see Dixon and Aikhenvald (2002); Haspelmath (2011) for discussion).

Taking all this into consideration, we have made the methodological decision of treating clitics as independent syntactic words. Therefore, the relationships between words and clitics is rendered as syntactic and is annotated by means of the appropriate dependency. All clitics in Shipibo-Konibo are phrasal in nature and treating them as independent words captures this in a more precise way (although annotation may be more time-consuming). In section 2.3 we present some examples.

Furthermore, following the principles for tokenizing a surface word into multiple *inflectional groups* (IGs) proposed by Çöltekin (2016, p. 2), we segment clitics as independent words because they and their host may participate in different syntactic relations. For instance, in the Shipibo-Konibo sentence *ea=ra joke* (I came), *ea* is the pronoun (I) in a dependency of `nsubj` from the verb *joke* (came), whereas *=ra* is an evidential clitic in the dependency of `aux:valid`.

Languages with similar morphological profiles have treebanks in Universal Dependencies, such as Finnish (Pyysalo et al., 2015), Turkish (Sulubacak et al., 2016) or Kazakh (Tyers and Washington, 2015). Nevertheless, those treebanks do not tend to systematically label bound morphemes as independent words, as we aim to do in the development of our treebank because of the reasons mentioned above.

2.3 Language Examples

We present two Shipibo-Konibo sentences in anticipation of further discussion.

The sentence *Jatianra en ja maxko bake panshin kírika menike* (So, I give this little boy a yellow book) in Figure 1 presents a ditransitive verb with direct and indirect objects. The clitic *=ra* has an evidential function, hence it projects the dependency relation `aux:valid` to the main verb *menike* (gave). The clitic *=n* expresses nominal case and projects to the token's core word. In Shipibo-Konibo, adjectives tend to precede nominal heads, with determiners preceding both adjectives and nominal heads as shown in the phrase *ja maxko bake*.

The sentence *Joninronki yoyo aká iki: "Jen, enra moa onanke"* (They say the man said, "Ah, I already knew that") in Figure 2 presents a direct speech construction showing two main verbs, each one with a evidentiality clitic. There are two multiword tokens with three syntactic words each, *joni =n =ronki* and *e =n =ra*.

3 Shipibo-Konibo Treebank

Our current Shipibo-Konibo treebank is the result of the syntactic annotation of 407 sentences extracted from parallel Shipibo-Konibo and Spanish educational materials and storybooks – complemented with elicited sentences produced and translated by the Shipibo-Konibo members of our team. This is a small treebank with work still ongoing (Table 1).

3.1 Typological features of Shipibo-Konibo

Shipibo-Konibo presents a basic AOV/SV constituent order (Figures 1 & 2), but it exhibits other

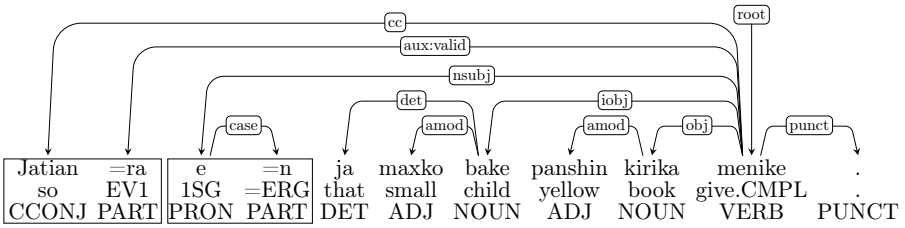

Figure 1: Dependency graph - clitic example (So, I gave that little boy a yellow book.)

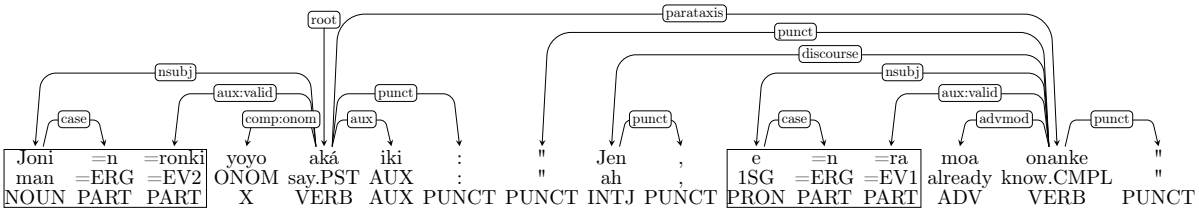

Figure 2: Dependency graph - complex clitic example (They say the man said, "Ah, I already knew that.")

Item	Count
Sentences	407
Orthographic tokens	2706
Syntactic words	3148

Table 1: Corpus Description.

pragmatically conditioned orders. NP-modifiers often precede their head (Figure 1) and verbs do not show either subject or object cross-reference.

As this is first treebank for any South-American indigenous language, there could well be novel grammatical features of Shipibo-Konibo not included in any other treebanks.

3.2 Universal Part of Speech (POS) Tags

Universal Dependencies (UD) introduces a tagset of 17 POS tags, mainly based in the Google universal part-of-speech tags (Petrov et al., 2012). All of them have been employed in the development of the Shipibo-Konibo treebank. The POS tags and frequencies in the treebank are shown in table 2.

The POS tag `X` is used for labelling onomatopoeia, which is a relevant POS in various Panoan languages, including Shipibo-Konibo (Valenzuela, 2003; Zariquiey, 2015, 2011). UD does not have an onomatopoeia POS tag. Hence, we opted to use `X` to label it. In other treebanks, onomatopoeias were ascribed to different POS tags. For example, Badmaeva (2016) in her "Universal Dependencies for Buryat" states that "the case of onomatopoeia is also an interjection" (2016, p. 40). However, onomatopoeias in Shipibo-Konibo are members of a special closed part of speech. They are used in combination with semantically generic verbs or auxiliaries as a productive strategy in order to form new words. Therefore, we considered it appropriate to label them as a different and independent POS.

As discussed in §2.2, Shipibo-Konibo *clitics* are a special type of linguistic unit that ought to be treated as an independent POS. Since Universal Dependencies does not present a clitic POS tag, but it does present a particle POS tag, `PART`, we opted to treat the Shipibo-Konibo clitics as particles, since *clitics* are often called *particles* (§2.2). These linguistic units are divided into three different categories: nominal clitic (expressing case and only used with nominal phrases), second position clitics (mainly expressing evidentiality and following the first constituent of a sentence), and less-fixed clitics (expressing adverbial value and used with any kind of POS). In this sense, it is important to remark that we are not considering them as adpositions `ADP`, since they belong to a closed set of items that occur before (preposition) or after (postposition) a complement composed of a noun phrase, noun, pronoun, or clause that functions as a noun phrase. Thus, they form a single structure with the complement to express its grammatical and semantic relation to another unit within a clause.

The high `PART` frequency noted in table 2 could impact performance in tasks as part-of-speech tagging or even syntax dependency parsing if it would require prior POS tag information. This was dis-

cussed and analyzed by Endresen et al. (2016) in a Russian corpus. We believe it will be important to measure whether the impact would be positive or negative in morphosyntactic tasks for Shipibo-Konibo as well, and thus, we would like to extend the discussion to a multilingual approach as further work.

POS	Count	%
Open class words		
NOUN	574	18.2
VERB	575	18.3
ADJ	119	3.8
ADV	103	3.3
PROPN	52	1.7
INTJ	7	0.2
Closed class words		
PART	440	14.0
PRON	177	5.6
AUX	162	5.1
DET	123	3.9
CCONJ	93	3.0
ADP	36	1.1
NUM	22	0.7
X (ONOM)	4	0.1
SCONJ	1	<0.1
Other		
PUNCT	654	20.8
SYM	2	0.1

Table 2: Universal POS.

3.3 Universal Morphological Features

The universal morphological features of UD are based on Zeman (2008)'s "Reusable tagset conversion using tagset drivers" with the concept of an expandable feature structure that could support any tagset. Tagset labels aim to "distinguish additional lexical and grammatical properties of words, not covered by the POS tags" (Nivre et al., 2017). A list of the morphological features and values used in the Shipibo-Konibo treebank annotation are given in Table 3; most are already defined in Universal Dependencies. The few morphological features of Shipibo-Konibo that require labels not currently in Universal Dependencies are underlined in Table 3.

The new morphological features are further defined below.

Aspect=And, Ven Shipibo-Konibo uses a set of deictic morphemes which indicate associated motion, *going* (andative) versus *coming* (venitive). Although there is literature arguing that associated motion should be treated as an independent grammatical category, the interaction between associated motion and aspect is well known (Guillaume, 2009).

Feature	Values
Animacy	Inam, Anim
Aspect	Perf, Hab, Iter, Imp, And, Ven
Case	Loc, Ela, Abl, Abs, Dat, Dis, Gen, Ill, Abe, Equa, Erg, Com, All, Tem, Ine, Voc, Chez
Evidentiality	Fh, Nfh
Mood	Jus, Frus, Des, Imp, Prev, Ind, Int
Number	Sing, Plur, Dual
Person	1, 2, 3
Polarity	Neg, Pos
Tense	Past1, Past2, Past3, Past4, Past5, Past6, Fut1, Fut2
VerbForm	Part, Inf
Voice	Mid, Rcp, Act, Cau, App
Clitic	Nomcl, Spcl, Lfcl

Table 3: Features in Shipibo-Konibo

Case=Chez Valenzuela defines a chezative case, which can be translated as "to/at the place where X is/lives" (2003, p. 232). Shipibo-Konibo encodes this case with the clitic *-ibá* ∼ *-ibat*.

Mood=Int Questions in Shipibo-Konibo are encoded by bound morphemes which are labeled by the dependency relation `aux:valid` (see §3.4.1).

Tense=Past1, Past2, Past3, Past4, Past5, Past6 Shipibo-Konibo presents six productive past categories. These tense categories are expressed by verbal bound morphemes. These features are presented in Table 4.

Tense=Fut1, Fut2 Shipibo-Konibo also has two different classes of future tense, expressed by bound morphemes. These features are also presented in Table 4.

Clitic=Nomcl, Spcl, Lfcl In §3.2 we introduced clitics with the `PART` POS tag, while also defining the three clitic categories as nominal clitic (`Nomcl`), second position clitic (`Spcl`), and less-fixed clitic (`Lfcl`).

Features currently annotated in the Shipibo-Konibo treebank are shown in Table 5. These

Universal features	Bound morpheme	Meaning
Past1	-wan	earlier the same day
Past2	-ibat ∼ -ibá	yesterday, a few days ago
Past3	-yantan	some months, a few years ago
Past4	-rabe	ca. 9 months to 3 years ago
Past5	-kati(t)	distant past, many years ago
Past6	-ni	remote past
Fut1	-nonx(iki)	indefinite future
Fut2	-yá ∼ -yat	tomorrow

Adapted from Valenzuela (2003, p. 284-285)

Table 4: Tense Features

have been automatically inferred based on POS tag, dependency relation, lexical, and, in the case of `AdpType`, language type information. Our next work update should deliver manually annotated features as well.

Feature	Value	Count
Clitic	Nomcl	263
Clitic	Spcl	176
Clitic	Lfcl	1
PronType	Int	86
AdpType	Post	36

Table 5: Inferred Features.

3.4 Dependency Relations

UD defines a set of 37 dependency relations, mainly based on "Universal Stanford Dependencies: A cross-linguistic typology" by Marneffe et al. (2014). Thirty-one of these 37 relations were employed in our Shipibo-Konibo treebank. One of the main characteristics of UD is that relations link content words rather than abstract nodes, i.e., *lexicalism* (Nivre et al., 2017). Dependency relations and frequencies in the treebank are reported in Table 6. It is worth mentioning that the frequency of `acl` and `ccomp` relation labels is low due to the choice of annotated sentences rather than a specific property of the language.

Shipibo-Konibo specific relations

While UD aims to provide "a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages" (Nivre et al., 2017), it also allows language-specific subtype relation labels when necessary. For the Shipibo-Konibo treebank, we considered the inclusion of two new subtype relation labels: `aux:valid` and `compound:onom`.

Relation	Count	%
punct	654	20.8
root	407	12.9
nsubj	314	10.0
case	299	9.5
obj	189	6.0
aux:valid	176	5.6
aux	172	5.5
amod	133	4.2
det	130	4.1
advcl	112	3.6
advmod	103	3.3
cc	93	3.0
obl	87	2.8
cop	67	2.1
nmod	67	2.1
compound	46	1.5
conj	39	1.2
iobj	21	0.7
nummod	15	0.5
discourse	6	0.2
appos	6	0.2
flat	4	0.1
vocative	3	0.1
acl	1	<0.1
ccomp	1	<0.1

Table 6: Dependency Relations

3.4.1 Relation subtype - aux:valid

An auxiliary is an element that may express different grammatical categories such as time, aspect, mood, voice and evidentiality. In Shipibo-Konibo, evidentiality and mood are expressed through a subset of clitics. These clitics are ascribed to the relation `aux`, but in order to distinguish them from verbal auxiliaries, they receive the subtype relation label `val`. This subcategory refers to the notion of *validator*, as defined by Cerrón-Palomino (2008, p. 166) for Quechua. For example, the sentence *Enra yapa yoá akai* (I cook fish) uses the first-hand evidentiality clitic *=ra* (Valenzuela, 2003, p. 534) to express that the speaker witnessed the event. See Figures 1 & 2 for more examples.

Note the high frequency of use for `aux:valid` shown in Table 6. At 176 instances, 5.6% of all syntactic words, almost half of Shipibo-Konibo sentences would include an expression of evidentiality (given seldom more than one `aux:valid` is used per sentence). This high frequency expression of evidentiality is an intriguing linguistic phe-

nomenon and worth further study.

3.4.2 Relation subtype - compound:onom

Similar to other languages of the Panoan language family, in Shipibo-Konibo, onomatopoeias are considered as a closed word class (Valenzuela, 2003). In this language there are constructions that include two *semantically generic* verbs: *ati* (do) or *iti* (be) (Valenzuela, 2003, p. 83). These elements may be combined with onomatopoeias in order to create a type of compound verb.

We decided to use the subtype relation label `compound:onom` for those specific types of compound verbs. For example, the verb *yoyo iti* (to speak) corresponds to a compound formed by the verb *iti* (be) and the onomatopoeia *yoyo* (speech noise). In spite of the fact that they are two differentiated entities, both elements constitute a unit at the semantic level, and therefore are compounds in Universal Dependencies. See Figure 2 as another example.

There is a significant use of compounds, 46 instances and 1.5% of syntactic words (Table 6), but only a few are due to onomatopoeia. While deemed important in the language, onomatopoeias have low frequency representation in the current instance of the treebank.

3.5 Segmentation and Multiword Tokens

Our decision to split orthographic tokens on clitic boundaries in §2.2 results in an abundance of multiple syntactic word tokens (Table 7) with 402 multiword tokens (MWTs) of 2706 total tokens. The clitic of second position, `Spcl`, invokes the dependency relation `aux:valid` typically with the clausal head and not with the core word of the MWT. The nominal clitic, `Nomcl`, invokes the dependency relation `case` with the core word of the MWT.

The cases where a token contains multiple clitics, the `Spcl` comes later. This has the effect of preserving projectivity. We continue to follow this issue of multiple clitic MWTs and projectivity.

3.6 Multiword Tokens vs Other Languages

Indeed, Shipibo-konibo has proportionally many more Multiword tokens (MWTs) than Spanish or Turkish, a language considered agglutinative, but less than Hebrew. Table 8 shows the differences where ∼15% of Shipibo-Konibo tokens are multiword versus ∼3% for Turkish, much less for Spanish, and ∼32% for Hebrew.

Property	Value	Count
MWTs	All	402
Num words	2	362
Relation	case	260
Relation	aux:valid	138
Relation	other	4
Head	not MWT core	137
Num words	3	40
Relation	aux:valid	35
Relation	other	5
Head	not MWT core	39

Table 7: Multiword Tokens

The big differences in MWT relative frequency is surprising given the UD documentation's explicit encouragement to use MWTs for annotating clitics (Universal Dependencies contributors, 2018). Our decision to segment tokens by phrasal clitic boundaries likely explains part of this large difference versus even other agglutinative languages.

Item	Quantity			
	Shipibo	Spanish	Turkish	Hebrew
Sentences	407	17680	5635	6216
Tokens	2706	547681	56422	115535
Multiword Tokens				
Count	402	1887	1640	37035
% tokens	14.86	0.34	2.91	32.06

See Spanish (Martínez Alonso and Zeman, 2017), Hebrew (Goldberg et al., 2017), and Turkish (Sulubacak et al., 2016) treebanks.

Table 8: Multiword Tokens Comparison

4 Word Order vs Spanish

We examined word order differences between the dominant Spanish and Shipibo-Konibo. Spanish results are from the training set of the Es-Ancora treebank (Martínez Alonso and Zeman, 2017), while Shipibo-Konibo results are from our treebank. Table 9 reports counts and relative frequencies of a constituent *preceding* its head. Constituents are reported either by their dependency relation with their head or POS in the case of single syntactic word constituents. Relative frequency of following the head is just the complement of that of preceding the head.

Direct and oblique objects usually follow the head (typically a verb) in Spanish and precede the head in Shipibo-konibo. Auxiliary verbs usually precede the head in Spanish and follow the

head in Shipibo-Konibo. Spanish uses prepositions and Shipibo-Konibo postpositions, but determiners precede their heads in both languages. Similar differences and similarities follow for the less common constituents as well.

Constituent ≺ Head	Shipibo		Spanish	
	Count	%	Count	%
obj	157	83.1	898	24.3
obl	64	73.6	209	15.3
iobj	19	90.5	62	71.3
nmod	59	88.1	8	0.3
acl	*	*	0	0.0
advcl	75	67.0	18	3.1
ccomp	*	*	1	0.4
advmod	91	88.4	298	52.1
amod	106	79.7	261	18.4
nummod	12	80.0	80	77.7
appos	1	16.7	0	0.0
cop	37	55.2	181	99.5
AUX	2	1.2	737	95.3
ADV	91	91.0	333	55.9
DET	123	100.0	5661	99.1
ADJ	77	75.5	279	16.8
ADP	1	2.8	5373	98.8

* Zero or one occurrence in Shipibo-Konibo corpus.

Table 9: Phrase or word order - Shipibo vs Spanish

Full confirmation of Shipibo-Konibo features versus the WALS database (Dryer and Haspelmath, 2013) awaits further progress. But a review of word order from Table 9 versus WALS largely confirms comparable word order features in WALS. An exception is adjective and noun head order. Our corpus shows ∼75% adjective preceding head (∼80% for adjective preceding *noun* head). So adjective precedes noun head order *dominates* versus the earlier finding by Faust (1973) reported in WALS of *no dominant order*.

5 Parsing for Shipibo-Konibo

Dependency syntax parsing is a complex task that usually requires a lot of annotated data, thus we decided to perform experiments in two different scenarios. The first one treats the treebank as an isolated corpus using monolingual methods, whereas the second one presents a cross-lingual experiment to identify which other languages from the UD v2.0 collection can support the parsing task for Shipibo-Konibo.

5.1 Monolingual Parsing

A straightforward test was performed using a greedy transition-based parser (Parsito) (Straka et al., 2015) from UDPipe (Straka and Straková, 2017) and the Yara Parser (Rasooli and Tetreault, 2015), which is also a transition-based method but uses beam search. The obtained results with 10-fold cross-validation are presented in Table 10, where we perform parses with POS gold annotations and raw text.

Input	Parser	UAS	LAS
Gold POS	Parsito	83.66±4.12	77.81±4.33
	Yara	87.32±2.90	81.25±3.45
Raw text	Parsito	37.68±1.23	30.39±1.34
	Yara	42.15±6.20	29.19±3.90

Table 10: Monolingual parsing accuracy for unlabeled (UAS) and labeled (LAS) attachment with gold POS tags and raw text as inputs

With the gold annotations, UAS and LAS scores from Parsito are greater than the language average of 78.59% and 72.81%, respectively, from Straka and Straková (2017); and the Yara Parser provides slightly better results in most cases. The low difference may be caused by the different search approaches (greedy versus global beam search) in the transition-based parsers. Meanwhile, parsing raw text scored much worse, which was expected for the corpus size. However, most of the cross-validation results has presented high variance; and thus, these results must not be treated as definitive ones, and only as a reference, as there could be overfitting and scarcity issues.

5.2 Cross-Lingual Parsing

We conducted an experiment with single-source cross-lingual delexicalized parser transfer from the UD v2.0 source languages into Shipibo-Konibo as the target language, in the vein of Zeman and Resnik (2008).

In the experiment, we used the `mate-tools` graph-based parser by Bohnet (2010) with default settings. The entire Shipibo-Konibo treebank was our test set. We tagged the treebank for POS using MarMoT (Mueller et al., 2013) via 10-fold cross-validation with a mean accuracy of 93.94±1.38 (s.d.). As we performed delexicalized transfer, all training and test data used only the following CoNLL-U features: ID, POS, HEAD, and DE-

kk	66.42	pl	54.02	hr	48.80	got	43.97
ja_ktc	63.26	lv	53.81	ja	48.46	no	42.26
eu	58.77	cs_cac	53.29	en	48.29	nl	42.22
tr	58.73	ro	53.29	sv	47.86	vi	41.57
ta	57.49	el	53.04	sv_lines	47.78	swl	41.49
fa	57.01	grc	52.69	sa	46.71	pt_bosque	40.98
hi	56.89	cs	51.63	id	46.49	grc_proiel	40.72
hu	56.46	sl	51.50	es_ancora	46.15	fi	39.31
et	55.77	cop	50.60	gl_treegal	46.15	it	39.31
bg	55.56	ru	50.51	pt	45.98	la_proiel	38.11
fi_ftb	55.43	sk	50.04	es	45.42	nl_lassysmall	37.98
de	55.35	gl	49.79	en_esl	45.38	cu	37.38
la_ittb	55.26	ru_syntagrus	49.57	ca	45.17	da	36.10
sl_sst	54.88	la	49.02	zh	45.17	ar	33.96
ug	54.58	en_lines	48.97	uk	44.91	ga	30.50
cs_cltt	54.23	fr	48.85	pt_br	44.31	he	23.22

Table 11: Cross-lingual parsing accuracy (UAS) for single-source delexicalized transfer parsers with Shipibo-Konibo as the target language. The source treebanks and their codes are from UD v2.0.

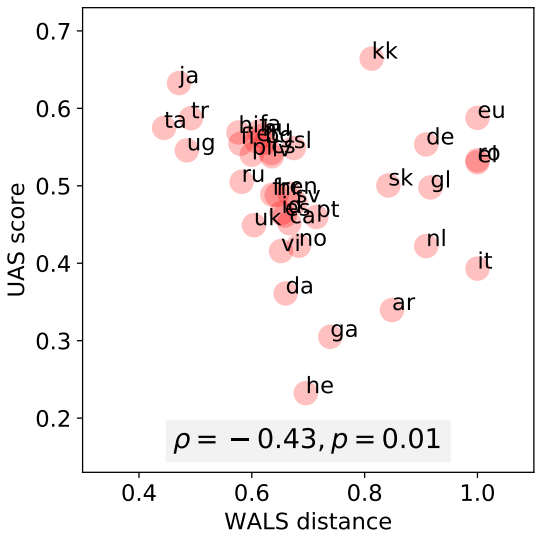

Figure 3: Cross-lingual parsing UAS scores correlated with source language WALS vector Hamming distance to Shipibo-Konibo. The correlation coefficient is Spearman's ρ.

PREL. Yet, to avoid any dependency label inconsistencies since our treebank is small, we evaluated for UAS only. We excluded all multiword tokens from the experiment, while retaining their respective syntactic words. A single delexicalized parser was trained for each UD v2.0 source treebank and applied on the Shipibo-Konibo test data.

Table 11 presents the results of the transfer parsing experiment. We achieve by far the best parsing results via the Kazakh delexicalized parser (66% UAS), closely followed by Japanese (63%), Basque and Turkish (ca 59%), and then Tamil, Persian, and Hindi (57%). Specifically, Kazakh presents morphosyntactic features similar to Shipibo-Konibo, such as SOV word order, high presence of agglutinative suffixes and head-final directionality (Mukhamedova, 2015). Moreover, the results are interesting as the top-performing cluster of sources for Shipibo-Konibo comprises languages that mainly feature as outliers in most cross-lingual parsing research, owing to the strong mainstream bias towards experimenting with resource-rich languages, as argued by Agić et al. (2016).

To further support our findings, we correlate the cross-lingual parsing UAS scores with language similarity of UD v2.0 source languages to Shipibo-Konibo. We express language similarity as pairwise Hamming distance between WALS vectors (Dryer and Haspelmath, 2013) for Shipibo-Konibo and the respective UD v2.0 source languages similar to Agić (2017). We depict this set of results in Figure 3, where we show a moderate negative correlation (Spearman's $\rho = -0.43$) between UAS and WALS distance, that is unlikely to be random at $p < 0.05$. In other words, the source languages that are more similar to Shipibo-Konibo in terms of WALS are more likely to provide Shipibo-Konibo with good delexicalized parsers. That said, some of the best source parsers are outliers in the figure: Kazakh and Basque yield good parsers for Shipibo-Konibo, but their WALS distance to it is large. This is due to the sparsity of WALS features for these languages: for example, 183 of 202 WALS features are null for Kazakh, and 188 for Basque, but only 41 for Japanese. Fixing these WALS feature deficiencies would in turn arguably strengthen the correlations to further support our findings. Besides, this analysis could be complemented by using a subset of WALS features that are generally available, as well as by inferring empty Kazakh features from related languages in the Kypchak group.

6 Conclusion and Future Work

We've presented Shipibo-Konibo from the Amazon region of Peru and our ample progress in building a treebank conforming to Universal Dependencies v2.0. We argued for segmenting syntactic words (versus tokens) along phrasal clitic boundaries and provided parse examples of this.

While our treebank is still a work in progress with 407 sentences, we've learned much already

about what distinguishes us from other languages and treebanks. Segmenting on phrasal clitics and POS tagging as PART resulted in a phenomenal 14% of clitics tagged as PART in our treebank, following only PUNCT, NOUN, VERB in popularity.

Several morphological features were added to account for past and future verb tenses, And and Ven aspects, Chez case, and Nomcl, Spcl, and Lfcl clitics. Each of these additions matters in the meaningful annotation of Shipibo-Konibo.

We considered two new dependency relation subtypes: aux:valid and compound:onom. The aux:valid relation occurred 176 times (5.6% of words and almost half of sentences). This high use evidentiality function invites further linguistic study.

By segmenting on phrasal clitics Shipibo-Konibo stands out in its use of multiword tokens (MWTs) including both two and three word MWTs. The Spcl clitic usually projects to the verbal head, but since it succeeds other clitics, projectivity is preserved. Shipibo-Konibo has a huge fives times as many MWTs (∼15% versus ∼3% for Turkish) versus other (agglutinative) languages.

Word order of Shipibo-Konibo versus Spanish reveals dramatic differences, which informs our work on machine translation between them. We largely confirmed WALS word order features for Shipibo-Konibo, except for our finding that adjective precedes noun is *dominant* as opposed to *no dominant order* as reported in WALS.

Results on a monolingual parser show promise with better than the language average performance for gold POS tags. Delexicalized cross-lingual parsing using parsers trained on all UD v2.0 treebanks, showed a maximum 66% unlabeled attachment score (UAS) for Kazakh, a language with similar morphosyntactic features, followed closely by Japanese at 63%. A plot of UAS versus Hamming distance from WALS vectors reveals the expected inverse correlation between WALS distance and UAS (lesser WALS distance related to higher UAS). Japanese showed a low WALS distance and a high UAS, but Kazakh showed both high WALS distance and high UAS (seemingly an outlier).

As future work, we will increase the size of the UD treebank, as well as annotate the morphological features in a semi-supervised way. There has been developed an FSM-based morphological analyzer (Cardenas Acosta and Zeman, 2018) that could support the annotation for that purpose. Moreover, as Shipibo-Konibo is one of many in the Panoan linguistic family, the next step would be the definition of the UD tagsets and guidelines for closely related languages, such as Iskonawa or Amawaka. We hope these efforts could extend language technologies development for minority languages in Peru.

Acknowledgments

We gratefully acknowledge the support of the "Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica" (CONCYTEC, Peru) under the contract 225-2015-FONDECYT. Furthermore, we appreciate the detailed feedback of the anonymous reviewers.

References

Željko Agić. 2017. Cross-lingual parser selection for low-resource languages. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 1–10, Gothenburg, Sweden. Association for Computational Linguistics.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Elena Badmaeva. 2016. Universal Dependencies for Buryat. Master's thesis, Universidad del País Vasco/ Euskal Herriko Unibersitate.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97, Beijing, China. Coling 2010 Organizing Committee.

Ronald Cardenas Acosta and Daniel Zeman. 2018. Morphological analyzer for shipibo-konibo. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Rodolfo Cerrón-Palomino. 2008. *Quechumara: Estructuras paralelas del quechua y del aimara*. Universidad Mayor de San Simón, Cochabamba, Bolivia.

Çağrı Çöltekin. 2016. (When) do we need inflectional groups? In *The First International Conference on Turkic Computational Linguistic*.

R. M. W. Dixon and Alexandra Y. Aikhenvald, editors. 2002. *Word: a cross-linguistic typology*, volume 20. Cambridge University Press, Cambridge.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online - World Atlas of Language Structures*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Anna Endresen, Laura A Janda, Robert Reynolds, and Francis M Tyers. 2016. Who needs particles? A challenge to the classification of particles as a part of speech in Russian. *Russian Linguistics*, 40(2):103–132.

Norma Faust. 1973. *Lecciones para el aprendizaje del idioma shipibo-conibo*, volume 1 of *Documento de Trabajo*. Instituto Lingüístico de Verano, Yarinacocha.

David W. Fleck. 2013. *Panoan languages and linguistics*. 99. American Museum of Natural History.

Yoav Goldberg, Reut Tsarfaty, Amir More, and Yuval Pinter. 2017. UD Hebrew HTB. http://universaldependencies.org/treebanks/he_htb/index.html.

Antoine Guillaume. 2009. Les suffixes verbaux de mouvement associé en cavineña. *Faits de Langues : Les Cahiers*, 1:181–204.

Martin Haspelmath. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1):31 –80.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the americas. pages 55–69.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Héctor Martínez Alonso and Daniel Zeman. 2017. UD Spanish AnCora. http://universaldependencies.org/treebanks/es_ancora/index.html.

Rodolfo Mercado, José Pereira, Marco Antonio Sobrevilla Cabezudo, and Arturo Oncevay. 2018. ChAnot: An intelligent annotation tool for indigenous and highly agglutinative languages in Peru. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Thomas Mueller, Helmut Schmid, and Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA. Association for Computational Linguistics.

Raikhangul Mukhamedova. 2015. *Kazakh: A comprehensive grammar*. Routledge.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Marie Candito, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, Jinho Choi, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Linh Hà Mỹ, Dag Haug, Barbora Hladká, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Natalia Kotsyba, Simon Krek, Veronika Laippala, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. Universal dependencies 2.0. http://hdl.handle.net/11234/1-1983. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istan-

bul, Turkey. European Language Resources Association (ELRA).

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal Dependencies for Finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (Nodalida 2015)*, pages 163–172.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara Parser: A fast and accurate dependency parser. *CoRR*, abs/1503.06733.

Annette Rios, Anne Göhring, and Martin Volk. 2008. A Quechua-Spanish parallel treebank. *LOT Occasional Series*, 12:53–64.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Milan Straka, Jan Hajic, Jana Straková, and Jan Hajic jr. 2015. Parsing universal dependency treebanks using neural networks and search-based oracle. In *International Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 208–220.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Umut Sulubacak, Memduh Gokirmak, Francis Tyers, Çağrı Çöltekin, Joakim Nivre, and Gülşen Eryiğit. 2016. Universal Dependencies for Turkish. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3444–3454. The COLING 2016 Organizing Committee.

Francis M. Tyers and Jonathan N. Washington. 2015. Towards a free/open-source universal-dependency treebank for kazakh. In *3rd International Conference on Turkic Languages Processing, (TurkLang 2015)*, pages 276–289.

Universal Dependencies contributors. 2018. Universal dependencies. http://universaldependencies.org.

Pilar Valenzuela. 2003. *Transitivity in Shipibo-Konibo Grammar*. Ph.D. thesis, University of Oregon.

Irina Wagner, Andrew Cowell, and Jena D Hwang. 2016. Applying universal dependency to the arapaho language. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 171–179.

Roberto Zariquiey. 2011. Uchumataqu, the lost language of the Urus of Bolivia: A grammatical description of the language as documented between 1894 and 1952 (hannß). *International Journal of American Linguistics*, 77:316–318.

Roberto Zariquiey. 2015. *Bosquejo gramatical de la lengua iskonawa*. Latinoamericana Editores/ CELACP/ Revista Crítica Literaria Latinoamericana, Lima, Boston.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Arnold M Zwicky. 1977. On clitics. Handout of Indiana University Linguistics Club.

Arnold M Zwicky. 1985. Cilitics and particles. *Working Papers in Linguistics*, 61(2):283–305.