Expanding Universal Dependencies for Polysynthetic Languages: A Case of St. Lawrence Island Yupik

Hyunji Hayley Park
Department of Linguistics
University of Illinois
hpark129@illinois.edu

Lane Schwartz
Department of Linguistics
University of Illinois
lanes@illinois.edu

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

Abstract

This paper describes the development of the first Universal Dependencies (UD, Nivre et al., 2016, 2020) treebank for St. Lawrence Island Yupik, an endangered language spoken in the Bering Strait region. While the UD guidelines provided a general framework for our annotations, language-specific decisions were made necessary by the rich morphology of the polysynthetic language. Most notably, we annotated a corpus at the morpheme level as well as the word level. The morpheme level annotation was conducted using an existing morphological analyzer (Chen et al., 2020) and manual disambiguation. By comparing the two resulting annotation schemes, we argue that morpheme-level annotation is essential for polysynthetic languages like St. Lawrence Island Yupik. Word-level annotation results in *degenerate* trees for some Yupik sentences and often fails to capture syntactic relations that can be manifested at the morpheme level. Dependency parsing experiments provide further support for morpheme-level annotation. Implications for UD annotation of other polysynthetic languages are discussed.

1 Introduction

The Universal Dependencies (UD) project (Nivre et al., 2016, 2020) provides a cross-lingual syntactic dependency annotation scheme for many languages. The most recent release of the UD treebanks (version 2.7) contains 183 treebanks in 104 languages. However, polysynthetic languages, known for words synthesizing multiple morphemes, are still much under-represented in the UD treebanks. To our knowledge, Abaza[1] and Chukchi (Tyers and Mishchenkova, 2020), are the only polysynthetic languages included in UD version 2.7.

In this paper, we describe how we annotated a corpus of St. Lawrence Island Yupik (also known as Central Siberian Yupik), a polysynthetic language spoken in parts of Alaska and Chukotka, Russia, within the framework of the UD guidelines. While UD is a framework for word-level annotations, we argue that morpheme-level annotations are more meaningful for polysynthetic languages. We provide morpheme-level annotations for Yupik in addition to word-level annotations.[2] We believe that subword-level annotations can help better capture morphosyntactic relations for polysynthetic languages and assist further dependency annotations and morphosyntactic research for polysynthetic languages.

Previously Tyers and Mishchenkova (2020) called for the need to annotate parts of words in regard to noun incorporation in Chukchi. They proposed annotating a noun incorporated into a verb via morphology as a separate token available in the enhanced dependency structure. While our approach is motivated by a similar need to annotate subword units for another polysynthetic language, our paper focuses on morpheme-level annotations, which may be applied to other types of multi-morphemic words than just noun incorporation.

In what follows, we describe the characteristics of the Yupik language (§2) and show how we annotated a corpus at the morpheme level as well as the word level (§3 and §4). Then we present some language-specific decisions we made for morpheme-level annotations and illustrate Yupik constructs captured by the new annotation scheme (§5 and §6). We also compare the performance of the two annotation schemes in automatic parsing experiments (§7). Based on our findings, we conclude that the morpheme-level annotation is essential and effective for polysynthetic languages and discuss implications of the study for other polysyn-

[1]The Abaza treebank, as released in UD v2.7, contains 33 sentences and does not provide any language-specific documentation.

[2]The UD_Yupik-SLI treebank is scheduled to be released in UD v2.8 on May 15, 2021. See https://universaldependencies.org for details.

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 131–142

thetic languages and the UD framework (§8 and §9).

2 St. Lawrence Island Yupik

St. Lawrence Island Yupik (ISO 639-3 *ess*; Yupik hereafter) is a polysynthetic language in the Inuit-Yupik language family, spoken in parts of Alaska and Chukotka, Russia. Like other polysynthetic languages, Yupik is characterized by its rich morphology. Jacobson (2001) provides the most thorough descriptions of the Yupik grammar with an emphasis on the morphology. Yupik is strictly suffixing with the exception of one prefix. Yupik words typically have the following form:

root (+ derivational morphemes)*
+ inflectional morpheme (+ enclitic)

That is, a typical Yupik word has a root, followed by zero or more derivational morphemes (thus forming a stem), followed by obligatory inflectional morpheme(s), finally followed by an optional enclitic. Most roots are nominal or verbal, such as *mangteghagh-* 'house' and *negh-* 'to eat' respectively. The language also includes a set of non-inflecting particles, such as *quunpeng* 'always' or *unaami* 'tomorrow'.

Yupik derivational morphology is highly productive; words with up to seven derivational morphemes have been attested (de Reuse, 1994, p.53), and words with 1-3 derivational morphemes are very common. The Badten et al. (2008) Yupik-English dictionary and the Chen et al. (2020) Yupik finite-state morphological analyzer document about 400 derivational suffixes:

- 81 noun-elaborating suffixes (N→N) that attach to nominal roots and yield nominal bases
- 61 verbalizing suffixes (N→V) that attach to nominal roots and yield verbal bases
- 218 verb-elaborating suffixes (V→V) that attach to verbal roots and yield verbal bases
- 36 nominalizing suffixes (V→N) that attach to verbal roots and yield nominal bases

We now provide two example Yupik sentences involving the Yupik nominal base *mangteghagh-* 'house'.

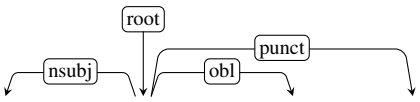
Taghnughhaat *aanut*
Taghnughha-at aan-u-t
child-ABS.PL to.go.out-IND.INTR-3PL

(1) ***mangteghameng***
mangtegha-meng
house-ABL_MOD.SG
'The children went out of the house.'
(Jacobson, 2001, p.22)

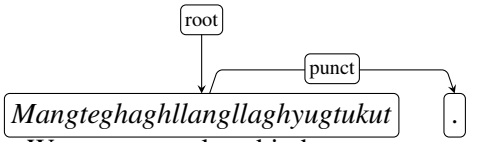
In (1), the Yupik nominal base *mangteghagh-* 'house' forms the word *mangteghameng* 'from the house' by taking the inflectional suffix *-meng* to mark ablative-modalis case.

(2) ***Mangteghaghllangllaghyugtukut.***
Mangtegha-ghlla-ngllagh-yug-tu-kut
house-big-to.make-to.want.to-IND.INTR-1PL
'We want to make a big house.'
(Jacobson, 2001, p.47)

In (2), the same nominal base takes multiple derivational morphemes, forming the sentence-length word *Mangteghaghllangllaghyugtukut*. To form this multi-morphemic word, the nominal base *mangteghagh-* first combines with the noun-elaborating derivational suffix *-ghlla-* (N→N), yielding an extended nominal base *mangteghaghlla-* 'big house'. This extended nominal base then combines with the verbalizing derivational suffix *-ngllagh-* (N→V) to create an extended verbal base *mangteghaghllangllagh-* 'to make a big house'. Next, this extended verbal base combines with the verb-elaborating suffix *-yug-* (V→V) to yield the extended verbal stem *mangteghaghllangllaghyug-* 'to want to build a big house'. Finally, the inflectional suffix *-tu-* attaches to the extended verbal stem to mark the verb's valency as intransitive and its mood as indicative, while the inflectional suffix *-kut* marks the person and number of the verb's subject as first person plural; the final result is the fully inflected word *mangteghaghllangllaghyugtukut* 'we want to make a big house'.

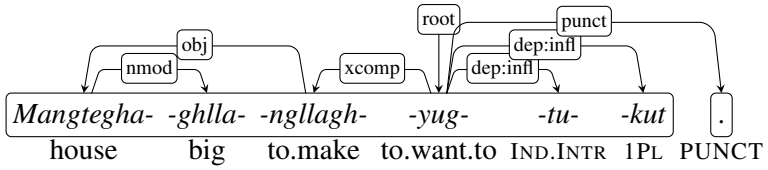
(3) 

Taghnughhaat *aanut* *mangteghameng* .

(4) 

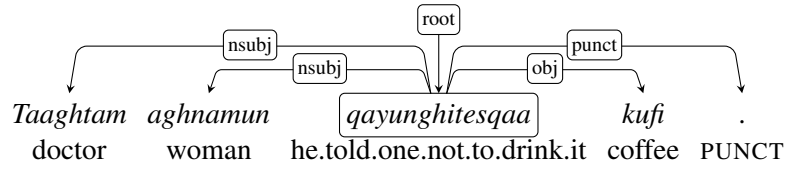
We want to make a big house. PUNCT

(5)

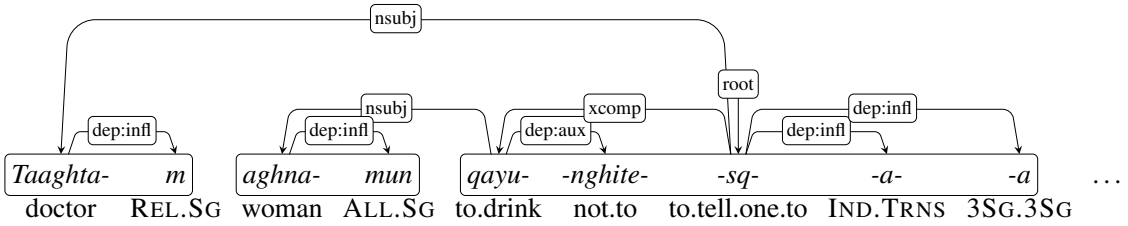


(6) Taaghta-m aghna-mun qayu-nghite-sq-a-a kufi-∅
doctor-REL.SG woman-ALL.SG to.drink-not.to-to.tell.one.to-IND.TRNS-3SG.3SG coffee-ABS.SG
'The doctor prevented the woman from drinking the coffee.' (Jacobson, 2001, p.67)

(7)



(8)



3 Morpheme-level dependency relations

The UD annotation guidelines are lexicalist (Chomsky, 1970; Bresnan and Mchombo, 1995) in nature, specifying that syntax dependencies should be annotated at the word level, such that both the head and the child of each dependency relation are words (Nivre et al., 2016).

In (3), we see the Yupik sentence from (1) with dependency relations annotated at the word level, following the UD guidelines. The resulting dependency tree successfully depicts the core syntactic information in the Yupik sentence, with the intransitive verb *aanut* at the root of the dependency tree, with a nominal subject and an oblique argument as children. However, when we annotate the single-word Yupik sentence from (2) according to the UD annotation guidelines, the result is a degenerate tree that completely fails to capture any syntactic information about the Yupik sentence.

In order to adequately represent the syntactic relations in (2), it is necessary to discard the lexicalist hypothesis and annotate relations between morphemes rather than between words. When we contrast (4) with (5), we observe that annotating relations at the morpheme level results in a meaningful linguistic analysis for this Yupik sentence. It is clear from these two dependency trees that treating morphemes as the basic unit of syntactic dependency relations is necessary in order to adequately encode the syntax of the Yupik sentence in (2). By doing so, we move from a degenerate tree devoid of syntactic information to a tree that successfully encodes a main verb *-yug-* ('to want to') with a complement *-ngllagh-* ('to make'), and an object *mangtegha-* ('house') with a nominal modifier *-ghlla-* ('big'); the inflectional suffixes encode the number and person of the subject (1PL, 'we') and the main verb's mood and valency (IND.INTR).

In (6) we observe a more complex Yupik sentence; we see the sentence *Taaghtam aghnamun qayunghitesqaa kufi* ('The doctor prevented the woman from drinking the coffee') annotated in (7) with dependency relations between words. The resulting dependency tree fails to illustrate the complex verbal structure of the multi-morphemic third word *qayunghitesqaa* ('he told one not to drink it'); it is only in (8) when we annotate (6) with syntactic relations between morphemes that we are able to observe that *aghnamun* ('the woman') is the subject of the embedded verb *qayu-* ('to drink') while *Taaghtam* ('the doctor') is the subject of the main verb *-sq-* ('to tell'). That is, parts of the Yupik word, the main verb *-sq-* ('to tell') and the embedded verb *qayu-* ('to drink'), participate in different syntactic relations, which cannot be annotated at the word level. The necessity for this type of sub-word annotation is not unique to Yupik; see Çöltekin (2016)

for a discussion of subword syntactic units in Turkish.

If sentences that required morpheme-level dependency relations were rare, it might be reasonable to accept the inclusion of a few degenerate and under-annotated trees such as (4) and (7) in a Yupik dependency treebank. However, Yupik is polysynthetic, and multi-morphemic words involving complex derivation are very common; the same is true of all of the languages in the Inuit-Yupik language family. For the polysynthetic languages in this language family, there are simply too many sentences that require morpheme-level dependency annotations to annotate only dependency relations between words. In particular, essentially all words formed with derivational suffixes require morpheme-level dependency relations in order to satisfactorily encode the syntax of the sentence.

In annotating Yupik sentences with dependency relations, we therefore treat each Yupik morpheme as a token rather than treating each Yupik word as a token. This necessarily requires that Yupik words be analyzed and segmented into morphemes prior to dependency annotation; this task was performed using the existing Yupik finite-state morphological analyzer (Chen et al., 2020). In cases of ambiguity when the analyzer provided multiple possible analyses for a given word, we selected the gold analysis via manual disambiguation.

We chose to represent all Yupik morphemes as independent syntactic tokens, including inflectional morphemes. An alternative approach would be to instead not tokenize inflectional morphemes, but rather annotate inflectional information using feature values. A major benefit of our choice is greater compatibility with the existing Yupik morphological analyzer (Chen et al., 2020), which treats inflectional morphemes as independent tokens in the underlying lexical form.

Because the UD annotation guidelines were not designed for morpheme-level annotation, some minor adaptations were required; we discuss these adaptations in §5 and §6 as we discuss the POS tags and dependency relations used in our corpus along with sample sentences. In order to enable the use of morphemes as tokens, we adapted the existing "multiword expressions" annotation mechanism. The UD annotation guidelines recognize that syntactic words do not always align perfectly with orthographic word boundaries; this can occur even in analytic languages such as English, for example, in words involving a clitic or a contraction. For example, in Spanish, the word *dámelo* ('give it to me') may be broken down into *dá me lo* ('give me it') for the purpose of UD annotations; the annotation scheme records that the single orthographic token (*dámelo*) is annotated as multiple syntactic words, and that information can be used to collapse the annotations to the single orthographic token when needed. In our case, we treat each multi-morphemic Yupik word as a UD "multiword expression," with Yupik morphemes serving as the tokens within the "multiword expression."

Recognizing the UD project's lexicalist view of syntax, we provide a script to convert our morpheme-level annotations into word-level annotations. This script deterministically merges each multi-morphemic word into a single word token using Udapi (Popel et al., 2017). Because our morpheme-level annotation does not strictly follow the entirety of the UD guidelines, a small number of sentences had to be manually corrected after the conversion. We plan to release our morpheme-level annotation in UD version 2.8 along with descriptions of the conversion process from the morpheme-level annotations to the word-level annotations.

Unit	Word-level	Morph-level
Sentences	309	309
Words	1,221	1,221
Segments	1,221	2,568
Fused	–	773

Table 1: Number of annotations per annotation level for the Jacobson corpus. **Words** mean the number of word tokens while **Segments** count any sub-word tokens instead of word tokens if applicable. **Fused** counts the number of word tokens that are split into subword units.

4 Corpus

The annotated corpus is comprised of exercise sentences from the Yupik reference grammar (Jacobson, 2001, as released in Schwartz et al., 2021). The grammar book, designed to teach Yupik at the college level, provides end-of-chapter exercises with sample Yupik sentences. Morphological segmentation and analyses were performed using the Chen et al. (2020) Yupik morphological analyzer and manually verified when needed.

The number of annotations for the final version of the Yupik treebank is summarized in Table 1. A total of 309 sentences with 1,221 word tokens

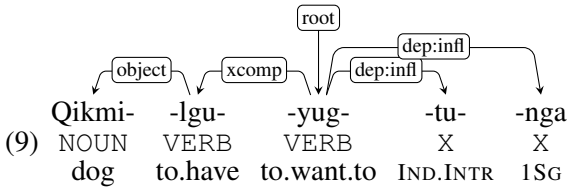
UPOS	Word-level	Morph-level
ADV	62	65
CCONJ	-	4
DET	5	5
NOUN	426	486
NUM	1	1
PART	16	16
PRON	19	23
PUNCT	310	310
VERB	382	556
X	-	1,102

Table 2: Frequencies of Part of Speech (POS) tags in the word-level and morpheme-level annotations for the Jacobson corpus.

were annotated. For the morpheme-level annotation, about 63% of the words (773 words) were further analyzed into the subword units, with a total of 2,568 segments (i.e. morphemes, particles and punctuation marks) annotated.

5 POS Tags

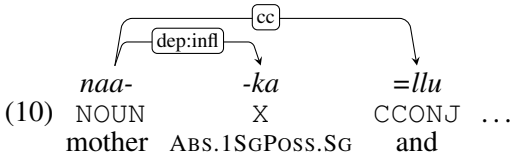
We annotated our Yupik corpus using the tags shown in Table 2.[3] Our morpheme-level annotations make use of ten POS tags; when these annotations are converted into word-level annotations, only eight POS tags are utilized.



We tagged nominals and nominal bases as NOUN and verbals and verbal bases as VERB. We tagged derivational suffixes that yield nominal stems (N→N, V→N) as NOUN and those that yield verbal stems (N→V, V→V) as VERB. For example, (9) shows the morpheme-level annotation for the word *Qikmilguyugtunga* 'I want to have a dog'. In the annotation, the nominal root *Qikmi-* 'dog' combines with a verbalizing derivational suffix (*-lgu-* 'to have', N→V) to yield a verbal base (*Qikmilgu-* 'to have a dog'). Then this extended base combines with the verb-elaborating suffix (*-yug-* 'to want to', V→V) to yield a complex verbal stem (*Qikmilguyug-* 'to want to have a dog'), which is followed by inflection. The two verb-yielding derivational suffixes are tagged as VERB.

Uninflected words or particles were given the particle tag (PART). Many Yupik particles are borrowed from Chukchi, a geographically neighboring language, and are mostly adverbial or connective in meaning (de Reuse, 1994, p.14). Examples include *ighivgaq* 'yesterday' and *qayughllak* 'because'.

The two additional POS tags available only at the morpheme level were X and CCONJ . The POS tag X is reserved for words that are outside of POS tags defined within the UD framework. We used the X tag for inflectional suffixes such as *-tu-* and *-nga* as in (9). Coordinating conjunctions (CCONJ) were only found at the morpheme level because they are only expressed as an enclitic in the language: *=llu* 'and' as in (10).



6 Dependency relations

Our morpheme annotation scheme makes use of 25 types of dependency relations while our word annotation scheme makes use of 14 dependency relations. In general, we followed the UD annotation guidelines, except in cases where polysynthetic nature of Yupik made divergence from the guidelines necessary. The full documentation on POS tags, morphological features, and dependency relations used in the treebank is available at the language's UD documentation page.[4]

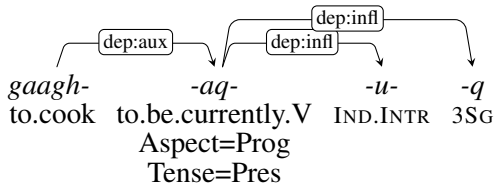
The most notable difference between the two annotation schemes is the dep relation. Within the UD framework, the dep relation is reserved for unspecified relations. Because morpheme-level annotations require multiple dependency relations specified for subword units, we created a few dependency relations under the dep relation for the morpheme-level annotation only. Note that some relations that are commonly annotated at the word level for other languages (e.g. auxiliary, copula) are only available at the morpheme level in Yupik. When we can, we expanded existing relations, defined at the word level, to morphemes (e.g. nmod

[3]The primary descriptions of Yupik are de Reuse (1994), which provides a description of Yupik syntax within the framework of autolexical syntax, and Jacobson (2001), which provides a description of Yupik grammar focusing on morphology and phonology in the context of a college-level Yupik class.

[4]More details are provided in Appendix A. See https://universaldependencies.org/ess/index.html for the full documentation.

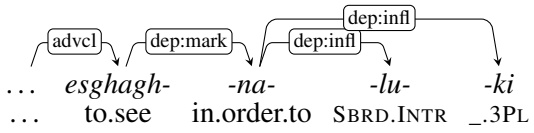
for nominal modifier). Whenever that was not possible, we created a version of the corresponding dependency relation in our morpheme annotation scheme.

For example, we used `dep:aux` for verb-elaborating (V→V) derivational morphemes that modify the base verb's tense and aspect information. For example, the V→V derivational morpheme (as manifested as *-aq-* in the context) adds the present tense and progressive aspect to the base *gaagh-* 'to cook' in (11).

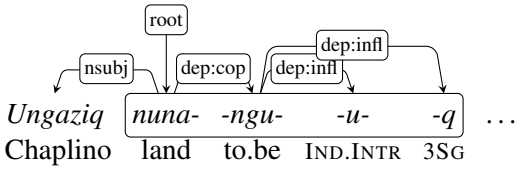
(11)


This relation would fit the descriptions of the auxiliary (`aux`) relation if it were annotated at the word level. We created a new relation as `dep:aux` to describe the dependency relation at the morpheme level because there were UD limitations to applying the existing `aux` relation to morphemes. First, the `aux` relation requires a short list of possible word forms while morphemes with the `dep:aux` relation may take many different forms depending on the context as they undergo morphophonological processes. Second, the word with the `aux` relation cannot have any children while corresponding morphemes often have inflections as their children.

Similarly, we included the `dep:mark` relation to represent the marker (`mark`) relation at the morpheme level. In (12) we observe a word that acts as a subordinate clause in a sentence and is roughly translated as 'in order to see them'. The second morpheme of the word *-na-* marks the word as a subordinate clause to the main verb, a `mark` relation in the word level UD annotation.[5] Again, because of some limitations of using this relation at the morpheme level, we created the `dep:mark` relation for morpheme-level anntoations.

(12)


On a similar note, the `dep:cop` relation was added to represent the copula (`cop`) relation at the morpheme level. In (13), the verbalizing (N→V) derivational suffix *-ngu-* acts as a copula, turning the nominal base as a verbal stem, which combines with the inflection to form a verbal word meaning 'it is a land' in the sentence meaning 'Chaplino is a land'.

(13)


The `dep:infl` was used for the relation between the stem and its inflectional suffix as shown in (13). Because all Yupik words other than particles require one or more inflectional morphemes, the `dep:infl` relation was the most frequently used in the morpheme-level annotation.

In general, morpheme-level annotation was needed to capture some of important morphosyntactic relations present in Yupik words. The `aux` and `cop` relations are only available at the morpheme level in Yupik. While a small number of particles act as marker, the `mark` relation was also primarily attributed to derivational suffixes. When annotating Yupik sentences at the word level, such dependency relations are lost. Only when we annotate at the morpheme level can we find such constructions, which may be invaluable in subsequent linguistic inquiries or computational applications alike.

7 Parsing experiments

In order to investigate the practical usage of the annotations, we conducted automatic parsing experiments using UDPipe 1.2 (Straka and Straková, 2017) and UDPipe 2.0 (Straka, 2018). The UDPipe project[6] provides a trainable pipeline for any UD treebanks in the CoNLL-U format.

7.1 Data

We made use of two sets of data: the Jacobson corpus and a separate test corpus annotated using the same word-level and morpheme-level annotation schemes. A text extracted from Nagai (2001) was annotated to provide an out-of-domain test set. The Nagai corpus was smaller than the entire Jacobson corpus with 360 word tokens or 834 tokens when including morphemes. The Nagai corpus is quite distinct from the Jacobson corpus. The former is a collection of an elder Yupik speaker's speech while the latter is a college-level grammar book. Therefore, the former has more disfluencies, repetitions,

[5] The inflection also shows that the word is in subordinative mood, where the subject of the verb is the same as the subject of the main verb.

[6] https://ufal.mff.cuni.cz/udpipe

	Word-level (Automatic segmentation)		Morph-level (Automatic segmentation)		Morph-level (Gold segmentation)	
Corpus	Jacobson (2001)	Nagai (2001)	Jacobson (2001)	Nagai (2001)	Jacobson (2001)	Nagai (2001)
Words	100	100	100	100	100	100
Segments	100	100	71.56 ± 3.68	42.39	100	100
UPOS	93.01 ± 2.08	71.59	69.82 ± 3.69	34.16	97.22 ± 1.40	80.79
Lemmas	71.47 ± 3.14	40.39	71.05 ± 3.67	39.51	99.19 ± 0.79	92.32
Features	78.17 ± 3.32	46.24	67.14 ± 2.65	34.02	94.17 ± 2.29	78.03
UAS	88.86 ± 1.64	60.72	45.82 ± 7.77	9.33	91.82 ± 2.98	67.95
LAS	81.52 ± 2.91	43.45	45.13 ± 7.69	9.33	89.30 ± 3.06	61.46

Table 3: Automatic parsing results using UDPipe 2.0 (Straka, 2018) for the word-level and morpheme-level annotation schemes. A test set was either 1) automatically segmented or 2) manually verified to have gold segmentation. The annotations on Jacobson (2001) was trained and tested using ten-fold cross validation. A sample text from Nagai (2001) was annotated to provide an out-of-domain test set. The columns show F_1 score: **Words** word tokenization; **Segments** splitting words into morphemes when applicable; **Lemmas** lemmatization; **UPOS** universal part-of-speech tags; **Feats** morphological features; **UAS** unlabelled attachment score (dependency heads); **LAS** labelled attachment score (dependency heads and relations).

and some code-switching with English words while the latter contains sample sentences in the literary language without any foreign words.[7]

7.2 Tokenization

At annotation time, the process of tokenizing sentences into syntactic tokens is performed manually as part of the annotation process. When annotating relations between morphemes, each morpheme serves as a token. When annotating relations between words, each word (delimited by whitespace or punctuation) serves as a token.

At test time, it is also necessary to tokenize each sentence. In our experiments, we consider three mechanisms for doing so.

In the first experimental condition, we follow standard dependency parsing practice and rely on the dependency parser to tokenize each sentence into word tokens. To do so, we used a UDPipe 1.2 (Straka and Straková, 2017) model to automatically tokenize each test sentence into word tokens. In Table 3, we refer to this tokenization method as *Word-level (Automatic segmentation)*.

In the second experimental condition, we used a UDPipe 1.2 (Straka and Straková, 2017) model to automatically tokenize each test sentence into morpheme tokens. In Table 3, we refer to this tokenization method as *Morpheme-level (Automatic segmentation)*.

In the third experimental condition, we assume that tokenization of words into morphemes is handled as a separate pre-process (for example, by a finite-state morphological analyzer). In this condition, we provide a test file in which words have already been correctly segmented into morpheme tokens. In Table 3, we refer to this tokenization method as *Morpheme-level (Gold segmentation)*.

We observe the results of tokenization in the first two rows of Table 3. The first row shows that all methods were able to identify word boundaries without error. In the second row of Table 3, we observe that using a dependency parser to segment Yupik words into morphemes is only 72% effective. This is problematic, as this places an upper bound on the potential dependency parsing performance of this condition. By definition, the third condition results in perfect morpheme tokenization.

7.3 Methods

We trained separate UDPipe 2.0 (Straka, 2018) parsers for the word-level annotations and the morpheme-level annotations, using the default UDPipe settings. UDPipe 1.2 (Straka and Straková, 2017) models were trained for tokenizing the test sets only, also using the default settings. To test in-domain performance, we trained and tested a parser on the original Jacobson corpus using ten-fold cross validation for each annotation scheme. For out-of-domain performance, we trained a parser on the entire Jacobson corpus and tested it on the Nagai corpus for each annotation scheme. The evaluation was conducted based on the official evaluation script from the *CoNLL 2018 UD Shared Task* (Zeman et al., 2018).

[7] More details about the Nagai corpus are available in Appendix B.

7.4 Results

Parsing results (unlabelled and labelled attachment scores) are shown in the final two rows of Table 3. In all cases, we observe that parsing accuracy for the in-domain data from Jacobson is substantially higher than in the out-of-domain data from Nagai.

When we compare the word-level and morpheme-level parsing given automatically segmented test sets (left and middle columns), the word-level parsing outperforms the morpheme-level parsing due to many segmentation errors present in the latter. Segmentation errors create an effective upper limit for any subsequent parsing efforts at the morpheme level, and all results in the second column are substantially worse than those in the first column.

In contrast, morpheme-level parsing outperforms word-level parsing across the board when correct morpheme tokenization is provided (right-most column). This shows that morpheme-level parsing (the second column) performed poorly on the automatically segmented test set mostly because of the poor quality morpheme segmentation. We observe that the morpheme-level dependency parser (the third column) outperforms the word-level parser (the first column) across the board, and even with the more challenging out-of-domain test set.

The task of analyzing and segmenting a word into its underlying component morphemes is a well-studied task for which robust finite-state solutions are well known. For polysynthetic languages especially, the development of such a finite-state morphological analyzer is nearly always the very first element of language technology developed. It is therefore realistic to assume that tokenization of words into morphemes can be effectively handled by in a pre-processing step prior to dependency parsing.

8 Discussion

The Universal Dependencies project is intended as a de-facto standard for consistent dependency syntax annotations across all of the world's languages (Nivre et al., 2016, 2020). Our attempt to construct a UD corpus of Yupik can be viewed as a kind of stress test for the UD annotation project. If the UD guidelines truly are universal in nature, then it should be possible to construct dependency trees for Yupik while fully following the UD guidelines; to the extent that this is not possible, any such disconnect may serve to illuminate ways in which the UD guidelines might be improved upon in order to be more language universal.

One of the core assumptions of the UD guidelines is lexicalism, the assumption that the fundamental token of syntax should be the word. This assumption has been widely adopted in many syntactic formalisms, including the Lexical-Functional Grammar theory of syntax that UD in part draws upon. It has, however, been widely debated (for a thorough recent critique of lexicalism, see Bruening, 2018), and other theories such as Distributed Morphology (Halle and Marantz, 1993) explicitly reject the lexicalist hypothesis, asserting that large parts of morphology and syntax operate using a common hierarchical mechanism.

The UD guidelines already explicitly recognize that phonological and orthographic boundaries do not always coincide with *syntactic words*. Nivre et al. (2016) recognize that clitics act as words from the viewpoint of syntax, even though phonologically (and orthographically) they must attach to a host word; as such in UD annotations clitics are treated as independent syntactic tokens. Similarly, the UD annotation guidelines recognize that contractions should be treated as the combination of two independent syntactic tokens. Finally, the UD guidelines recognize that some larger units such the English expression *in spite of* act syntactically as a single token.

However, the existing UD guidelines indicate that derivational morphemes should not be treated as syntactic words for the purposes of dependency annotation. For example, in an English dependency tree, the word *dancer* would be treated as a single syntactic token, rather than as two (verbal root *dance-* + nominalizing suffix *-er*). In this paper, we have observed that this approach to derivational morphology fails when applied to Yupik.

The languages in the Inuit-Yupik language family are polysynthetic and rely heavily on productive derivational morphology. St. Lawrence Island Yupik has around 400 derivational suffixes, around half of which are verb-elaborating (V $\rightarrow$ V) derivational suffixes. It is essentially impossible to adequately annotate the syntax of Yupik sentences without recognizing that significant parts of Yupik grammar are handled by Yupik derivational morphology.

In this paper, we have chosen to treat every Yupik morpheme (both derivational and inflectional) as a syntactic token. In future work, it may be beneficial

to build upon work by Çöltekin (2016) and treat only some derivational morphemes as syntactic tokens, while not tokenizing other derivational morphemes and perhaps all inflectional morphemes. At a minimum, this work shows that in order to be universal, the UD project must acknowledge that at least some derivational morphemes must be treated as syntactic tokens.

9 Conclusion

This paper presents the first UD treebank for St. Lawrence Island Yupik, the first UD treebank to be annotated at the morpheme level as well as the word level to our knowledge. The polysynthetic language has rich morphology, characterized by a theoretically unlimited number of possible derivations and multimorphemic words. In order to capture the morphosyntactic relations among morphemes, we annotated a corpus (Jacobson, 2001) at the morpheme level and converted the morpheme-level annotations into word-level annotations. While the morpheme-level annotation may require more linguistic resources (e.g. morphological analyzer, morphological segmentation), it provides a deeper insight into the language and better automatic parsing performance. Morpheme-level syntactic dependency annotation may be a better way to represent polysynthetic languages within the framework of UD.

References

Linda Womkon Badten, Vera Oovi Kaneshiro, Marie Oovi, and Christopher Koonooka. 2008. *St. Lawrence Island / Siberian Yupik Eskimo Dictionary*. Alaska Native Language Center, University of Alaska Fairbanks.

Joan Bresnan and Sam A. Mchombo. 1995. The lexical integrity principle: Evidence from Bantu. *Natural Language and Linguistic Theory*, 13(2):181–254.

Benjamin Bruening. 2018. The lexicalist hypothesis: Both wrong and superfluous. *Language*, 94(1):1–42.

Emily Chen, Hyunji Hayley Park, and Lane Schwartz. 2020. Improved finite-state morphological analysis for St. Lawrence Island Yupik using paradigm function morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2676–2684, Marseille, France. European Language Resources Association.

Noam Chomsky. 1970. Remarks on nominalization. In Roderick A. Jacobs and Peter S. Rosenbaum, editors, *Readings in English Transformational Grammar*, pages 184–221. Ginn, Waltham, MA.

Çağrı Çöltekin. 2016. (When) do we need inflectional groups? In *Proceedings of the First International Conference on Turkic Computational Linguistics*.

Willem J. de Reuse. 1994. *Siberian Yupik Eskimo — The Language and Its Contacts with Chukchi*. Studies in Indigenous Languages of the Americas. University of Utah Press, Salt Lake City, Utah.

Morris Halle and Alec Marantz. 1993. Distributed morphology and the pieces of inflection. In Kenneth Hale and S. Jay Keyser, editors, *In The View from Building 20*, pages 111–176. MIT Press, Cambridge, MA.

Steven A. Jacobson. 2001. *A Practical Grammar of the St. Lawrence Island / Siberian Yupik Eskimo Language, Preliminary Edition*, 2nd edition. Alaska Native Language Center, Fairbanks, Alaska.

Kayo Nagai. 2001. *Mrs. Della Waghiyi's St. Lawrence Island Yupik Texts with Grammatical Analysis*. Number A2-006 in Endangered Languages of the Pacific Rim. Nakanishi Printing, Kyoto, Japan.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. 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 Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden. Association for Computational Linguistics.

Lane Schwartz, Emily Chen, Hyunji Hayley Park, Edward Jahn, and Sylvia Schreiner. 2021. A digital corpus of St. Lawrence Island Yupik. *ArXiv*, abs/2101.10496.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

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.

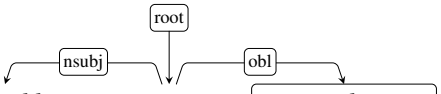
Francis M. Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

A Overview of dependency relations used in the Jacobson treebank

Table 4 summarizes dependency relations used in the word-level and morpheme-level annotations for the Jacobson corpus. In this section, we provide additional descriptions of the dependency relations that we added for Yupik but were not introduced in the main text due to limited space.

We added a sub-relation (`obl:mod`) to the existing `obl` relation to specify a special usage of a noun in ablative-modalis case. The existing `obl` relation is used for an oblique nominal or as a non-core argument of the corresponding verb. For example, a noun in ablative-modalis case is annotated as an oblique nominal (`obl`) when used to express motion away from somewhere as in *mangteghameng* (house-ABL_MOD.SG, 'from the house') in (14).

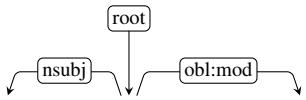


(14) *Taghnughhaat* *aanut* *mangteghameng*
children they.went.out from.house

In contrast, a noun in ablative-modalis case can also be used as "indefinite object" of an intransitive verb (Jacobson, 2001, p.20). For example, *pagunghaghmeng* (crowberry-ABL_MOD.SG) in (15) is understood as the object of an intransitive verb as an indefinite form of the noun (e.g. "crowberries" instead of "the crowberries"). Because an indefinite object in ablative-modalis case is not encoded in the verb, we annotated such nouns as an oblique noun, but distinguished it with the rest of oblique nouns by specifying the sub-relation, `obl:mod`, dedicated to those indefinite objects.

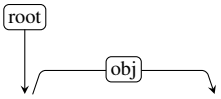
Dependency	Word-level	Morph-level
acl	-	17
advcl	73	73
advmod	73	76
appos	21	21
cc	–	4
conj	2	2
dep:ana	–	7
dep:aux	–	120
dep:cop	–	12
dep:emo	–	1
dep:infl	–	1,087
dep:mark	–	5
dep:pos	–	3
det	5	5
mark	3	3
nmod	46	68
nmod:arg	-	3
nsubj	173	173
nummod	1	1
obj	94	121
obl	67	69
obl:mod	44	44
punct	310	310
root	309	309
xcomp	-	34

Table 4: Frequencies of dependency relations in the word-level and morpheme-level annotations for the Jacobson corpus.



(15) *Afsengaq* *neghtuq* *pagunghaghmeng*
mouse it.ate crowberries

This is different from the `obj` relation for a noun in absolutive case used as the object of a transitive verb. The nominal base (*pagungha-* 'crowberry') takes the absolutive case inflection in (16) when used as the object of a transitive verb.



(16) *Pagunghaat* *aavgii*
crowberries she.divided.them

We also added the `nmod:arg` sub-relation to the existing `nmod` (nominal modifier) relation to specify when a nominal base is used as the argument of a noun-elaborating (N→N) derivational

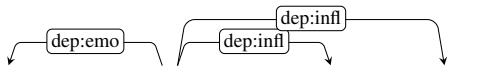
suffix. In (17), the nominal base (*aqavzi-* 'cloudberry') modifies the derivational suffix as the argument (*aqavzileg-* 'the one with cloudberry'). The extended base then combines with the inflection to yield the noun in ablative-modalis case (*aqavzileg-meng* 'from the one with cloudberry').

(17) *aqavzi-* *-leg-* *-meng*
cloudberry one.with.N ABL_MOD.SG

The `dep:pos` relation was used for the relation between a postural root and its postbase. A postural root takes a postbase to yield a verbal stem as in (18). The postural root (*ingagh-* 'lying down') combines with the postbase (*-nga-*) to yield a stative form of the root (*ingaghnga* 'to be lying down'), which combines with the inflection to form the word (*ingaghngaghpek*, 'you are lying down'). A postural root is different from nominal or verbal bases as it can only take one of two postbases that turn the root into a stative or active form to be followed by inflection.

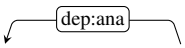
(18) *ingagh-* *-nga-* *-gpek*
lying.down to.be.in.R.posture 2SG

Similarly, the `dep:emo` relation was used for emotional roots. Emotional roots can take one of a select number of postbases to yield nominal or verbal stems. In (19), the emotional root (*qugina-* 'spooked') takes the postbase (*-k-*) to yield a verbal stem (*quginak* 'to be spooked'), which combines with the inflection to form a verbal (*quginakanka* 'I am spooked by them').



(19) *qugina-* *-k-* *-a-* *-nka*
spooked to.feel.R.toward IND.TRNS 1SG.3PL

The `dep:ana` relation is used for the only prefix in Yupik, the anaphoric prefix. In general, the prefix is used for anaphora, emphasis or specificity. The prefix is also used in demonstratives to provide reference to person spoken to or situation spoken about (Jacobson, 2001, p.109).



(20) *taaku-* *-m*
ANAPHOR DEM.PRO.REL.SG

In (20), the anophoric prefix (*taaku-*) combines with the inflection to result in the demonstrative pronoun (*taakum* 'this one').

Unit	Word-level	Morph-level
Sentences	66	66
Words	360	360
Segments	360	834
Fused	–	225

Table 5: Number of annotations in a sample of Nagai (2001). **Words** mean the number of word tokens while **Segments** count any sub-word tokens instead of word tokens if applicable. **Fused** counts the number of word tokens that are split into subword units.

UPOS	Word-level	Morph-level
ADJ	1	1
ADP	1	1
ADV	11	16
NOUN	78	105
NUM	2	2
PART	43	43
PRON	9	9
PUNCT	81	81
VERB	134	214
X	-	362

Table 6: Frequencies of Part of Speech (POS) tags in the word-level and morpheme-level annotations for the Nagai corpus.

B Overview of the Nagai treebank

This section provides additional information about the Nagai annotations, used for the parsing experiments in §7. Table 5 summarizes the number of annotations for the new corpus. As introduced in the main text, this corpus was smaller than the Jacobson corpus, but was bigger than a test set in the ten-fold cross-validation setting.

In general, the new corpus provides a more realistic and challenging test set for an automatic parser. The Nagai corpus records a Yupik elder's speech and presents some code-switching with English words. For example, the Nagai corpus included an English word 'electric beater' inflected in Yupik *electric beater-meng*. For this, we used an additional feature 'Foreign=Yes' in annotating the corpus.

Because of such foreign words, the distribution of the POS tags were slightly different from the Jacobson treebank. Table 6 summarizes the POS tags used to annotate the Nagai corpus, and shows the presence of some tags used only for English words: For example, the Nagai annotations included an adposition (`ADP`), which was an English word, 'on'.

Because the new corpus was smaller than the original treebank, there were some POS tags in the original Jacobson corpus that were missing in the new corpus. No DET or CCONJ tags were used in the new corpus. Similarly, some dependency relations that were present in the Jacobson corpus were not present in the new corpus: cc, dep:emo, and det.