Selecting the UD v2 Morphological Features for Indonesian Dependency Treebank

Ika Alfina
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
ika.alfina@cs.ui.ac.id

Daniel Zeman
Faculty of Mathematics and Physics
Charles University
Prague, Czechia
zeman@ufal.mff.cuni.cz

Arawinda Dinakaramani
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
arawinda@cs.ui.ac.id

Indra Budi
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
indra@cs.ui.ac.id

Heru Suhartanto
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
heru@cs.ui.ac.id

***Abstract*—The objectives of our work are to propose the relevant Universal Dependencies (UD) morphological features for Indonesian dependency treebank and to apply the proposed features to an existing treebank. We propose the use of 14 UD v2 features and the corresponding 27 feature-value tags. To evaluate the quality of the resulting treebank, we built models for lemmatization, POS tagging, morphological features analysis, and dependency parsing using UDPipe, a trainable pipeline for tokenization, tagging, lemmatization, and dependency parsing of CoNLL-U files. For lemmatization, POS tagging, and morphological features analysis tasks, the resulting models have F1-score of more than 93% that shows that the consistency of annotations for the columns LEMMA, UPOS, and FEATS in the treebank is already good. However, the accuracy of the Indonesian dependency parser built is still only 82.59% for UAS and 79.83% for LAS. The experiments also show that morphological features information has no or little impact on improving the quality of lemmatization, POS tagging, and dependency parsing models for Indonesian.**

***Index Terms*—annotation guidelines, dependency treebank, morphological features, Universal Dependencies**

I. Introduction

Morphology is "*the study of the way words are built up from smaller meaning-bearing units, morphemes*" [1]. Morphological parsing can produce information about the words like the stem/lemma, affixes, part-of-speech (POS), and additional information called morphological features [1]. Example of morphological features are singular/plural indicator for nouns or active/passive voice for verbs.

Universal Dependencies (UD) is a framework for grammar annotation for various languages, from parts of speech, morphological features, and syntactic dependencies. The first version of the annotation guidelines was called UD v1 [2] and current version is called UD v2 [3]. The recent release 2.6 of the UD dataset consist of 163 treebanks from 92 languages.

Indonesian, also known as *Bahasa Indonesia*, is an Austronesian language with over 260 million speakers in 2020. In [4], Indonesian is addressed as Malay-Indonesian, since the indigenous name of the language is *Bahasa Melayu* (the Malay language). Indonesian is similar with the Malay language used in Malaysia, Brunei and Singapore.

There are already two Indonesian treebanks in the UD dataset v2.6, named Indonesian GSD [5] and Indonesian PUD [6]. In 2019, [7] conducted analysis to both treebanks and suggested that their quality still needed improvements since many aspects in tokenization, POS tagging and syntactic annotation were not fully aligned to the Indonesian grammar. They proposed improvements and revised the Indonesian PUD treebank. We found out that the revised Indonesian PUD treebank produced by [7] also still need improvements since they have not filled the lemma and features columns in that treebank.

Previous works had worked on Indonesian morphological features. In 2008, [8] built a morphological analyzer that produce information about the active/passive voice of Indonesian verbs. MorphInd [9] produces more features like singular/plural information for nouns. Some UD Indonesian treebanks also have applied UD features to the dataset. However, we found many differences among those works about what features are relevant for Indonesian.

The objectives of our work are:

- To propose the relevant UD v2 features for Indonesian grammar.
- To revise an Indonesian dependency treebank by applying the proposed features.
- To investigate the impact of using morphological features in building models for lemmatization, POS tagging, and dependency parsing for Indonesian.

The contributions of our work are a list of 14 proposed relevant UD v2 features along with 27 feature-value tags for Indonesian and a new version of the Indonesian PUD treebank that has been made public[1] for Indonesian NLP community.

[1]https://github.com/ialfina/revised-id-pud

978-1-7281-7689-5/20/$31.00

TABLE I
THE UD V2 UPOS TAGSET

UPOS Tag	Class Word	UPOS Tag	Class Word
ADJ	adjective	PART	particle
ADP	adposition	PRON	pronoun
ADV	adverb	PROPN	proper noun
AUX	auxiliary	PUNCT	punctuation
CCONJ	coordinating conjunction	SCONJ	subordinating conjunction
DET	determiner	SYM	symbol
INTJ	interjection	VERB	verb
NOUN	noun	X	other
NUM	numeral		

TABLE II
THE UD V2 MORPHOLOGICAL FEATURES

Abbr	Foreign	Polite
Animacy	Gender	Poss
Aspect	Mood	PronType
Case	NounClass	Reflex
Clusivity	Number	Tense
Definite	NumType	Typo
Degree	Person	VerbForm
Evident	Polarity	Voice

The rest of this paper is organized as follows. Section II discusses the related work, our proposed relevant UD v2 features are explained in Section III. Section IV describes our approach in conducting the revision and the statistics of the revised treebank. Section V presents the experiment results and discussion. Finally, we discuss the conclusions and future work in Section VI.

II. RELATED WORK

In this section, we discuss the previous works on morphological features. First, we discuss Universal Dependencies morphological features and then we present the previous works on Indonesian morphological features.

A. Universal Dependencies Morphological Features

Universal Dependencies (UD) uses Universal Part-of-Speech (UPOS) for core part-of-speech categories. UD v2 annotation guidelines defined 17 UPOS tags as shown in Table I. Tags from a secondary, language-specific tagset can be put in the XPOS column of the treebank. For a UD treebank, UPOS is mandatory while XPOS is optional.

Furthermore, for fine-grained part-of-speech and grammatical categories, UD defines morphological features. In UD v2 annotation guidelines, 24 features are defined. Table II shows the complete list of UD v2 morphological features.

B. Indonesian Morphological Features

In 2008, Pisceldo et al. [8] built an Indonesian morphological analyzer tool that provides a detailed analysis of the affixation process and reduplication. This tool can parse the word into the affixes, lemma, and additional information like part-of-speech (POS) and active/passive voice information for verbs.

In 2011, Larasati et al. built MorphInd, an Indonesian morphological analyzer [9] that can produce lemma and part-of-speech (POS) tag of words. Moreover, this tool also provide additional morphological features information as follows:

- singular/plural tags for noun, proper nouns, personal pronouns, and verb
- first/second/third person tags for personal pronouns
- cardinal/ordinal/collective tags for numbers
- positive/superlative tags for adjective
- gender with feminine/masculine/non-specified tags for noun and proper nouns
- active/passive voice tags for verbs

Furthermore, Indonesian dependency treebank named Indonesian GSD built by [5] in 2013 applied 12 UD v1 morphological features to that treebank: *Clusivity, Degree, Gender, Number, Number[psor], NumType, Person, Person[psor], Polarity, Polite, PronType*, and *Voice*. Another Indonesian dependency treebank named Indonesian PUD built in 2018 by [6] applied only four features of 24 UD v2 features: *Foreign, Number, Polarity* and *PronType*. We can see that there are various set of morphological features employed by these works for Indonesian.

III. SELECTING THE UD V2 FEATURES FOR INDONESIAN

In this section we discuss the 24 morphological features of UD v2 and their relevance to Indonesian grammar.

A. Selection Procedure

The selection process begins by conducting the literature study to UD v2 annotation guidelines and Indonesian reference grammar [10], [11]. If a feature was discussed in at least one of the two Indonesian reference grammar then it will be considered relevant. For each possible feature, there are also one or more possible feature values. We also conducted some analysis to decide which values are relevant.

B. The Proposed Relevant Features

We propose the use of 14 UD v2 morphological features that we consider relevant to Indonesian grammar. Table III shows the 14 selected features, along with their relevant feature-value tags. In total, we propose the use of 27 UD v2 feature-value tags for Indonesian dependency treebank.

Among the 14 features, 3 features are universal and not specific to Indonesian grammar: *Abbr, Foreign*, and *Typo*. Feature *Abbr* is used for abbreviation words, feature *Foreign* is used for foreign words, and feature *Typo* is used for the misspelled word. We will discuss the remaining 11 features in the following paragraphs.

Clusivity is a feature of first-person plural personal pronouns that indicates whether the other party in conversation is included. This feature has two possible values: *Ex* and *In*, *Ex* for exclusive and *In* for inclusive. In Indonesian grammar, there are pronouns *kami* and *kita* that both translated to "we" in English. The pronoun *kami* refers to the speaker and other people but the listeners are not included, this word will be

TABLE III
THE PROPOSED RELEVANT UD V2 FEATURES FOR INDONESIAN

Feature	Feature-value	Description
Abbr	Abbr=Yes	abbreviation words
Clusivity	Clusivity=Ex	exclusive plural pronoun
	Clusivity=In	inclusive plural pronoun
Degree	Degree=Sup	superlative adjective
Foreign	Foreign=Yes	foreign word
Mood	Mood=Imp	imperative mood
	Mood=Ind	indicative mood
Number	Number=Plur	plural
	Number=Sing	singular
NumType	NumType=Card	cardinal number
	NumType=Ord	ordinal number
Person	Person=1	first person
	Person=2	second person
	Person=3	third person
Polarity	Polarity=Neg	for negation or negative response
Poss	Poss=Yes	possessive pronoun
PronType	PronType=Dem	demonstrative pronoun
	PronType=Emp	emphasis determiner
	PronType=Ind	indefinite pronoun/determiner/adverb
	PronType=Int	interrogative pronoun/adverb
	PronType=Prs	personal pronoun
	PronType=Rel	relative pronoun/adverb
	PronType=Tot	total pronoun/determiner/adverb
Reflex	Reflex=Yes	reflexive pronoun
Typo	Typo=Yes	misspelled words
Voice	Voice=Act	active verb
	Voice=Pass	passive verb

labeled with *Clusivity=Ex*, while the pronoun *kita* refers to the speker and the listener, and will be labeled with *Clusivity=In*.

Degree is a feature for degree of comparison for some adjectives or adverbs. This feature has five possible values, but for Indonesian only one value, *Sup* (for superlative adjectives), we consider relevant. Indonesian superlative adjective has special marking, i.e. adjective with prefix *ter-*, such as *terbaik* "the best" and *tercantik* "the most beautiful".

Mood is a feature for verbs that expresses modality. This feature has 12 possible values, among them are *Imp* (imperative), *Ind* (indicative), and *Cnd* (conditional). For Indonesian, we suggest only *Mood=Imp* and *Mood=Ind* are relevant. Some Indonesian imperative verbs have special marking with suffix *-kan*, for example *Cerita**kan** padaku!* "Tell me!". We use *Mood=Ind* as the default value for this feature.

Number is a feature to indicate the quantity of a noun. This feature has 11 possible values. For Indonesian, we propose the use of only two values: *Sing* for singular and *Plur* for plural. There are three part-of-speeches (POS) that can be tagged with *Number*:

- Nouns. Indonesian mainly uses reduplication to make a singular noun became plural. For example, anak "child" is a singular noun, the plural one is *anak-anak* "children".
- Pronouns. *Number=Sing* can be applied to *saya/aku* "I", while *Number=Plur* can be applied to *kami/kita* "we".
- Determiners. Plural noun can also be formed by adding certain determiners like *para/beberapa* "several/some/-many" to a singular noun, such as in *beberapa mahasiswa* "some students".

NumType is a feature to describe numeral type. This feature has seven possible values. For Indonesian, we propose the use of only two values: *Card* for cardinal numbers and *Ord* for ordinal numbers.

- *NumType=Card*, for *satu* "one", *juta* "million"
- *NumType=Ord*, for *pertama* "first", *ke-7* "7th"

Person is a feature of personal and possessive pronouns/determiners, and of verbs. For Indonesian, it can be applied to all personal and possessive pronouns. This feature has five possible values: 0, 1, 2, 3, and 4, but only three values are relevant to Indonesian:

- Person=1, for the first person, such as *aku, saya* "I", *-ku* "me/my", *kami/kita* "we".
- Person=2, for second person, such as *kamu/anda/kalian* "you", *-mu* "you/your".
- Person=3, for third person, such as *dia/ia* "he/she/it", *-nya* "him/her/it", *mereka* "them".

Polarity is a feature for polar indicator with two possible values: *Pos* and *Neg*. *Polarity=Neg* is used for negating words or also can be used for negative response. For Indonesian, *Polarity=Neg* can be applied to:

- Negating words, such as *tidak/tak* "not", *belum* "not yet", *jangan* "do not".
- Negative response, such as *tidak* "no" in ***Tidak**, terima kasih* "**No**, thank you".

Poss is a feature that indicates whether the word is possessive, with only one possible value: *Yes*. For Indonesian, this feature can be applied to possessive pronouns. Note that all Indonesian personal pronouns can serve as possessive pronouns.

PronType is a feature that describes the type of pronouns, pronominal adjectives (determiners), pronominal numerals (quantifiers), and pronominal adverbs. Among the 11 possible values for this feature, we propose the use of only seven values, as follows:

- *PronType=Dem*, for demonstrative pronoun (e.g. *ini* "this"), determiner (e.g. *tersebut* "the"), and adverb (e.g. *sana* "there").
- *PronType=Emp*, for emphatic determiner. In Indonesian, the word that qualifies as empathic determiner is *sendiri* "self", such as in *Kamu harus percaya pada diri **sendiri*** "You have to believe in yourself". Word *sendiri* in this case is a determiner that emphasize the pronoun *diri*.
- *PronType=Ind*, for indefinite pronoun (e.g. *seseorang* "someone", sesuatu "something"), determiner (e.g. *seorang* "a person", sebuah "a"), numeral (e.g. *banyak/beberapa* "many/some"), and adverb (e.g. *kadang-kadang* "sometimes").
- *PronType=Int*, for interrogative pronoun (e.g. *apa* "what", *siapa* "who"), and adverb (e.g. *mengapa* "why", *bagaimana* "how", *kapan* "when", *di mana* "where")
- *PronType=Prs*, for personal or possessive personal pronoun or determiner. For Indonesian, all personal pronouns can also serve as possessive personal pronoun. All words

that qualify for feature *Person* are also qualify for this feature.

- *PronType=Rel*, for relative pronoun (e.g. *apa* "what", *siapa* "who", *yang* "which"), and adverb (e.g. *mengapa* "why", *bagaimana* "how", *kapan/saat/ketika* "when", *di mana* "where")
- *PronType=Tot*, for total (collective) pronoun (e.g. *semua* "all"), determiner (e.g. *seluruh/semua/segala* "all", *setiap/masing-masing* "each"), and adverb (e.g. *selalu* "always")

Reflex is a feature to indicate whether a word is reflexive, with only one possible value: *Yes*. A word is reflexive if it refers to the subject of its clause. In Indonesian, according to [11], only one word qualifies as the reflexive word: the reflexive pronoun *diri* "self". For example, in *Ia mencalonkan **diri** menjadi presiden* "He is running for president", the word *diri* refers to the subject *Ia* "he/she".

Voice is a feature of verbs that indicates the type of the subject. In UD v2, this feature has 10 possible values, but only two values are relevant to Indonesian: *Act* for active verbs where the subject is the actor and *Pass* for passive verbs where the subject is the patient. Indonesian grammar uses inflection to differentiate between active and passive verbs, as follows:

- Active verb can be formed either without affix or with one of the prefixes *me-/ber-*, such as *pergi* "go", ***me**manggil* "call", ***be**kerja* "work".
- Passive verb can be formed with one of the prefixes *di-/ter-* or the circumfix *ke-an*, such as ***di**beli* "be bought", ***ter**jatuh* "fell", and ***ke**curi**an*** "be stolen".

C. Irrelevant Features

There are 10 UD v2 features that we consider irrelevant to Indonesian grammar: *Animacy, Aspect, Case, Definite, Evident, Gender, NounClass, Polite, Tense*, and *VerbForm*.

The features *Case, Animacy, Gender* and *NounClass* are usually the inflectional features that apply to nominal words like nouns, pronouns, or determiners and are used to mark agreement between nouns and other parts of speech. In Indonesian grammar, nouns have no grammatical gender and there is no requirement that the subject agrees with other parts of the clause.

Tense and *Aspect* are verb features. *Tense* is a feature that specifies the time when the action happened. This feature is irrelevant since Indonesian verbs have the same forms for all tenses. For example, the verb *pergi* "go" always has the same form in present, past and future tenses. *Aspect* is a feature that specifies duration of the action in time. Indonesian verbs also do not inflect for this feature.

Definite is typically a feature of nouns, adjectives and articles with five possible values. Indonesian does not have articles, and as far as we know, Indonesian nouns or adjectives do not have different forms reflecting definiteness.

Evident is the morphological marking of a speaker's source of information. Based on Indonesian reference grammars that we used, there are no Indonesian words that have this feature.

VerbForm is a feature that indicates the form of a verb with respect to its distinct morphological and distributional behavior. There are 8 possible values, such as *Fin* (finite), *Inf* (infinitive), and *Part* (participle). We consider *VerbForm* irrelevant since there are no such distinctions in the reference grammars that we used.

Polite is a feature to indicate politeness and has four possible values: *Infm* (informal), *Form* (formal), *Elev* (referent elevating), and *Humb* (humbling). There are words that are considered informal and formal in Indonesian. However, since we limited our work on formal Indonesian, we decided to hold off on using this feature and will discuss it on future work.

IV. Revising the Indonesian PUD Treebank

In this section, we describe how we revised the Indonesian PUD treebank that had been revised by [7]. The main revision task was to add the morphological features that had been defined in Section III.

A. Annotation Procedure for Lemma and Features

The treebank revised by [7] has empty value for lemma and features. To fill these columns, initially we assigned the values automatically using an Indonesian morphological analyzer tool named Aksara [12] that built as a parallel project with our work. Currently, Aksara only implemented 18 of 27 feature-value tags proposed by our work.

The manual corrections were done by four annotators. The treebank consists of 1,000 sentences. We divided it into two parts, each consists of 500 sentences. Each part was manually corrected by two annotators. We evaluated the corrections iteratively. A series of meetings that involving all team members of five persons were conducted to compare the differences between the two annotators and resolved the problems.

B. Conducting minor revisions to word segmentation, POS tagging and syntactic annotations

For word segmentation, we adopted the proposed word segmentation by [7], except for one case. In [7], a verb nominalized with the suffix *-nya* was tagged NOUN and the suffix was kept part of the word. For example, the token *meningkatnya* "the increase" that consist of the verb *meningkat* "to increase" and *-nya* "the" was not split into two tokens in the previous treebank, and it was tagged as a NOUN. Since we will not find the nominalized form in an Indonesian dictionary, and since *-nya* has other functions in which it is separated as a clitic (which makes tokenization and POS tagging more difficult), we decided to split that kind of tokens in our proposed revision.

We also conducted minor corrections for POS tagging especially for compound words and possessive determiners so that the POS tagging is more aligned to UD v2 annotation guidelines. For syntactic annotation, we changed some dependency relations used by removing and adding subtypes. The details of why we changed the subtypes are beyond the scope of this paper.

TABLE IV
THE COMPARISON OF THE PREVIOUS AND THE PROPOSED REVISION OF INDONESIAN PUD

Description	Previous	Proposed
Word count	19,401	19,440
Average sentence length	19.40	19.44
UPOS tag count	17	17
Feature count	**0**	**14**
Feature-value tag count	**0**	**26**
Deprel count	47	46
Language specific deprel	15	14

TABLE V
THE DISTRIBUTION OF POS TAGS IN THE PROPOSED REVISION

POS tag	Freq.	POS tag	Freq.	POS tag	Freq.
ADJ	1032	INTJ	4	PUNCT	2384
ADP	1906	NOUN	4681	SCONJ	439
ADV	646	NUM	515	SYM	38
AUX	405	PART	225	VERB	2392
CCONJ	591	PRON	1331	X	41
DET	723	PROPN	2087		

C. Statistics of the Revised Treebank

Table IV shows the comparison of general statistics of the previous version of Indonesian PUD [7] and our proposed revision. Since we conducted minor revision for word segmentation and syntactic annotation, the number of words is changed and also the number of dependency relations used. Major changes are shown in number of features and feature-value tags since in the previous version, there are no features at all. Note that the number of feature-value tags in that table is only 26 since one tag is not represented in this treebank, *Mood=Imp*.

Table V shows the distribution of POS tags in the proposed revision of Indonesian PUD treebank. We can see that all 17 POS tags defined in UD v2 are represented, with NOUN as the most frequent POS tag with 4681 counts and INTJ (interjection word) as the least frequent one with only 4 occurrences.

Table VI shows the distribution of feature-value tags. The most frequently occurred feature-value tag is *Number=Sing* with 5089 occurrences, followed by *Voice=Act* with 1889 counts. Feature *Number* is related to NOUN, PRON and DET tags. Since the occurrence of these three POS tags are 6735 in total, we can expect that feature Number will be dominant, with feature-value tag *Number=Sing* as the most frequent one, compared to *Number=Plur*.

The least occurred feature-value tags is *Mood=Imp* with zero count, followed by *Typo=Yes* and *PronType=Int*. Feature *Mood=Imp* indicates imperative sentences and *PronType=Int* indicates interrogative sentences in the treebank. For this treebank of 1,000 sentences, unfortunately, there is no imperative sentence and only 12 interrogative ones.

V. EXPERIMENTS AND RESULTS

We used UDPipe [13], a tool that can be used to train models of tagger and parser using supervised learning approach. For tagging, we built lemmatization, POS tagging, and morphological feature analysis models, and for parsing, we built a dependency parser model using the revised treebank discussed in Section IV. Ten-fold cross-validation method was conducted to evaluate the quality of the resulting models.

TABLE VI
THE DISTRIBUTION OF FEATURE-VALUE TAGS

Feature-value tag	Freq.	Feature-value tag	Freq.
Abbr=Yes	104	Polarity=Neg	177
Clusivity=Ex	26	Poss=Yes	268
Clusivity=In	19	PronType=Dem	465
Degree=Sup	58	PronType=Emp	17
Foreign=Yes	32	PronType=Ind	215
Mood=Imp	0	PronType=Int	12
Mood=Ind	2392	PronType=Prs	602
Number=Plur	376	PronType=Rel	611
Number=Sing	5089	PronType=Tot	58
NumType=Card	510	Reflex=Yes	14
NumType=Ord	76	Typo=Yes	6
Person=1	103	Voice=Act	1889
Person=2	16	Voice=Pass	503
Person=3	469		

TABLE VII
F1-SCORES OF LEMMATIZATION, POS TAGGING, AND MORPHOLOGICAL FEATURES ANALYSIS (%)

Scenario	Lemma	POS	Features
Single task	**95.64**	**94.78**	93.98
Two tasks: Lemma-POS	93.49	94.40	-
Two tasks: POS-Features	-	94.75	**94.85**
Two tasks: Lemma-Features	94.22	-	93.64
Three tasks	93.53	94.53	94.38

A. Building Lemmatization, POS Tagging and Morphological Feature Analysis Model

To build models for lemmatization, POS tagging and features analysis, we used the tagging module of UDPipe. The main data for lemmatization task is the lemma column of the treebank, the main data for POS tagging is the UPOS column, and the main data for feature analysis is the FEATS column. To build these models, we conducted three scenarios for these three columns. The first scenario was to build a model that performed only one task, where we just used the main data to build each model. For the second scenario, we trained two tasks for one model, resulting in three variations of models: lemmatization-POS tagging, lemmatization-feature analysis, and POS tagging-feature analysis. Finally, in the third scenario, we built one model that can perform all three tasks together.

Table VII shows the results of the tagging experiments for the three scenarios. We can see that, for lemmatization and POS tagging tasks, the best F1-score was achieved when the models were built using the first scenario, while for the feature analysis task, the best model was achieved using the second scenario, that used the UPOS and FEATS columns.

For the lemmatization task, additional information of POS tag and features have negative impact in building the lemmatization model. For the POS tagging task, the differences between the best scores and others are less than 0.38%. We can say that actually additional information of lemma and features

TABLE VIII
PERFOMANCE OF DEPENDENCY PARSER (%)

Model	UAS	LAS
Without features	82.40	79.49
With features	82.59	79.83

are also not needed to build the POS tagger model. Finally, for the feature analysis task, the difference between the best and the worst scores is 1.21%, where the best score achieved when UPOS was included in building the model and the worst score achieved otherwise. We suggest that since the FEATS column is not mandatory and in the treebank many forms have empty features, additional non-empty data (here the UPOS tags) is needed to build the feature model.

B. Building Dependency Parser

We used two scenarios in building the Indonesian dependency parser using the revised treebank. To build a dependency parser, minimal requirements are using form, UPOS, head, and deprel columns. For the first scenario, we included the FEATS column that contains morphological features, and for the second scenario we built the model without the FEATS column. Table VIII shows the results for these two scenarios. We can see that the differences between them is very small, only 0.19% for UAS and 0.34% for LAS.

The results show that the newly revised treebank can be used to build good lemmatization, POS tagging and feature analysis models with the F1-score of more than 93%. However, the accuracy achieved in building the dependency parser is still around 82% for UAS and 79-80% for LAS. The results also show that morphological features information has very small or no impact on building lemmatization, POS tagging, and dependency parsing.

VI. CONCLUSIONS AND FUTURE WORK

In this work, we have selected 14 of 24 features defined in UD v2 annotation guidelines along with 27 feature-value tags that we consider relevant to Indonesian grammar. The majority of UD v2 features are the inflectional features. Since Indonesian does not have so many inflectional cases, many features are not relevant.

We also have revised an existing Indonesian dependency treebank that initially had not defined the features where the FEATS column in that treebank was empty. Besides added features, we also added lemma, and conducted minor revision for word segmentation, POS tagging, and syntactic annotation.

To evaluate the quality of the resulting treebank, we built lemmatization, POS tagging, morphological analysis, and dependency parsing models using UDPipe. For lemmatization, POS tagging, and morphological features analysis, the resulting models have F1-score of more than 93% that shows that the consistency of annotations for columns LEMMA, UPOS, and FEATS in the treebank is already sufficient. However, for the Indonesian dependency parser built using this revised treebank, the LAS achieved is only 79.83%, which is needed improvement. Experiment results also show that morphological features information has very small or no impact on building lemmatization, POS tagging, and dependency parsing.

For future work, we want to build a bigger Indonesian dependency treebank. Since manual annotation is costly, we will use a semi-supervised approach to enlarge the treebank.

ACKNOWLEDGMENT

This work was supported by the research grant of "Publikasi Terindeks Internasional (PUTI) Prosiding" 2020 Number: NKB-875/UN2.RST/HKP.05.00/2020 from Universitas Indonesia.

The authors also would like to thank Muhammad Yudistira Hanifmuti, Jessica Naraiswari Arwidarasti, and Yogi Lesmana Sulestio that had participated as annotators in revising the Indonesian PUD treebank.

REFERENCES

[1] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2009.

[2] J. Nivre, M.-c. D. M. Filip, G. Yoav, J. Hajič, D. M. Ryan, M. Slav, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman, "Universal Dependencies v1: A Multilingual Treebank Collection," in *LREC 2016*, 2016, pp. 1659–1666.

[3] J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman, "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection," in *Language Resources and Evaluation (LREC)*, 2020, pp. 4034–4043.

[4] D. E. Ager and B. Comrie, *The World's Major Languages*, 1990.

[5] R. McDonald, J. Nivre, Y. Quirmbach-brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, O. Tackstrom, C. Bedini, N. B. Castello, and J. Lee, "Universal Dependency Annotation for Multilingual Parsing," in *the Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013, pp. 92–97.

[6] D. Zeman, J. Hajič, M. Popel, M. Potthast, M. Straka, F. Ginter, J. Nivre, and S. Petrov, "CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies," in *the Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 2018, pp. 1–21.

[7] I. Alfina, A. Dinakaramani, M. I. Fanany, and H. Suhartanto, "A Gold Standard Dependency Treebank for Indonesian," in *the Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33)*, 2019.

[8] F. Pisceldo, R. Mahendra, R. Manurung, and I. W. Arka, "A Two-Level Morphological Analyser for the Indonesian Language," in *the Proceedings of the 2008 Australasian Language Technology Association Workshop (ALTA 2008)*, 2008, pp. 142–150.

[9] S. D. Larasati, V. Kuboň, and D. Zeman, "Indonesian morphology tool (MorphInd): Towards an Indonesian corpus," in *Communications in Computer and Information Science*, vol. 100 CCIS, 2011, pp. 119–129.

[10] H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A. M. Moeliono, *Tata Bahasa Baku Bahasa Indonesia (Formal Indonesian Reference Grammar)*, 1998.

[11] J. N. Sneddon, A. Adelaar, D. N. Djenar, and M. C. Ewing, *Indonesian Reference Grammar*. A&U Academic, 2010.

[12] M. Y. Hanifmuti and I. Alfina, "Aksara: An Indonesian Morphological Analyzer that Conforms to the UD v2 Annotation Guidelines," in *the Proceedings of the 2020 International Conference of Asian Language Processing (IALP)*, 2020.

[13] M. Straka, J. Hajič, and J. Strakov, "UDPipe : Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing," in *LREC 2016*, 2016.