*Corpora* • Volume 16, Issue 2, Pages 287 - 299 • August 2021

**Document type**
Article

**Source type**
Journal

**ISSN**
17495032

**DOI**
10.3366/COR.2021.0221

**Publisher**
Edinburgh University Press

**Original language**
English

View less

# An evaluation of MorphInd's morphological annotation scheme for Indonesian

Prihantoro

Save all to author list

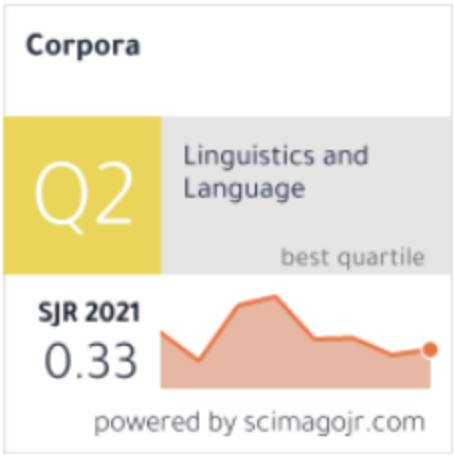
[a] Centre for CASS (Corpus Approaches to Social Science) Office, Lancaster University, Lancaster, LA1 4YW, United Kingdom



Abstract

## Abstract

MorphInd2 (Larasati et al. 2011) is a state-of-the-art morphological analyser for Indonesian. To




← Show this widget in your own website

Just copy the code below and paste within your html code:

<a href="https://www.scimag

# Source details



## Corpora

Scopus coverage years: from 2006 to 2022
Publisher: Edinburgh University Press
ISSN: 1749-5032 E-ISSN: 1755-1676
Subject area: Arts and Humanities: Language and Linguistics | Social Sciences: Linguistics and Language
Source type: Journal

View all documents > | Set document alert | Save to source list

CiteScore 2021
1.6

SJR 2021
0.331

SNIP 2021
1.377

CiteScore | CiteScore rank & trend | Scopus content coverage

**Improved CiteScore methodology**

CiteScore 2021 counts the citations received in 2018-2021 to articles, reviews, conference papers, book chapters and data papers published in 2018-2021, and divides this by the number of publications published in 2018-2021. Learn more >

**CiteScore 2021**

$$1.6 = \frac{108 \text{ Citations } 2018 - 2021}{68 \text{ Documents } 2018 - 2021}$$

Calculated on 05 May, 2022

**CiteScoreTracker 2022**

$$1.4 = \frac{108 \text{ Citations to date}}{75 \text{ Documents to date}}$$

Last updated on 05 April, 2023 • Updated monthly

CiteScore rank 2021

Randi Reppen, Northern Arizona University, USA

## Editorial Board

Published in Corpora 2021 Vol. 16(3)
https://euppublishing.com/doi/abs/10.3366/cor.2021.0221
Print ISSN: 1749-5032 ; Online ISSN: 1755-1676

# An Evaluation of MorphInd's Morphological Annotation Scheme for Indonesian

Prihantoro ♠♢ [ORCID ID: 0000-0001-7708-9785]

♠ Lancaster University, Lancaster, United Kingdom
♢ Universitas Diponegoro, Semarang, Indonesia
prihantoro2001@yahoo.com

## Abstract

MorphInd[1] (Larasati et al., 2011) is a state-of-the-art morphological analyser for Indonesian. To date, there has not been any comprehensive evaluation of the morphological annotation scheme which MorphInd implements. My evaluation of this annotation scheme reveals a number of significant drawbacks. Some analytic features encoded in MorphInd's tagset seem not to reflect features actually present in Indonesian morphology, while certain common features in the analysis of Indonesian are absent. Likewise, the Part of Speech (POS) hierarchy in the MorphInd tagset does not reflect the usual POS hierarchy used by Indonesian reference grammars. Moreover, the MorphInd output does not link morphological tags to the corresponding morpheme. Finally a number of issues which might problematise text/corpus querying in the annotation's layout are observable, particularly relating to affixes, reduplication, and the affix-reduplication interface.

Keywords: MorphInd, Indonesian, morphology, morphosyntactic, annotation

## 1. A brief description of Indonesian and MorphInd

Indonesian (ISO 639-3 Ind), or *Bahasa Indonesia* (its autonym), is one of the standardised varieties of Malay or *Bahasa Melayu* (its autonym). Indonesian is the sole official language, as well as the national language, of the Republic of Indonesia, spoken by more than 200 million speakers (Lewis, 2009) either as their first or second language. Morphologically, the majority of polymorphemic words in Indonesian are built by means of affixation, reduplication and compounding (Mueller, 2007:1208-1215).

MorphInd (Larasati et al. 2011) is a Morphological Analyser, or MA, for Indonesian. In this paper[2], I evaluate MorphInd's annotation scheme. This is Larasati et al.'s (2011) morphological annotation scheme, abbreviated here as LM. I argue that a comprehensive evaluation of LM is important for two reasons. First, MorphInd is considered the state-of-the-art MA for Indonesian (see section 2), but there has to date not been any detailed evaluation of LM. Second, no matter how excellently the MorphInd program performs, its performance is simply measured by how successful MorphInd implements LM. If linguistically incorrect analyses come from the flaws in LM, standard evaluation of MorphInd would treat these as successes. While

[1] https://septinalarasati.com/morphind/
[2] I would like to thank Andrew Hardie for useful feedback on the initial draft of this paper.

Published in Corpora 2021 Vol. 16(3)
https://euppublishing.com/doi/abs/10.3366/cor.2021.0221
Print ISSN: 1749-5032 ; Online ISSN: 1755-1676

not incorrect, as MorphInd is designed to implement LM 'as is', this may give a misleading view of the utility of the system.

MorphInd outputs an analysis by creating all possible analysis of each word according to following LM. If a word has only one analysis in terms of tokenisation and tagging, this analysis will be supplied in the final output. If a word in has more than one analysis, MorphInd selects a single analysis by statistical disambiguation. The detail of how this works is beyond the scope of this paper as it is a matter of the MorphInd software, not the annotation scheme. However, MorphInd's production of single-analysis output via this disambiguation means that LM as an annotation does not include provision for ambiguous annotation.

## 2. Why MorphInd is the state-of-the-art morphological analyser for Indonesian

There exist two MAs for Indonesian. The first was built by Pisceldo et al., (2008) which I refer to as PMA (for *Pisceldo et al.'s Morphological Analyser*). The second is MorphInd[3] (Larasati et al., 2011), which, I argue, is presently the state-of-the-art MA for Indonesian[4]. MorphInd is presented as an advance on PMA by Larasati et al. (2011:120-121).

Why should MorphInd be considered state-of-the-art? First, MorphInd's annotation scheme (LM) represents an improvement relative to PMA in terms of tokenisation and tagging (Larasati et al., 2011:120-121). In addition to affixation and reduplication, also covered by PMA, LM can additionally represent cliticisation analysis. LM's fine-grained tagset is richer than PMA's comparatively underspecified tagset. LM annotation is also more robust than PMA as all morphemes are presented.

Second, MorphInd is functional and in relatively continuous development. This stands in contrast to PMA, which due to various technical issues does not run on current systems. This explains why MorphInd, rather than PMA, is used by other Indonesian NLP systems, as the following non-exhaustive survey illustrates: Dinakaramani et al. (2014) used MorphInd to build a rule-based POS tagger for Indonesian. Green et al. (2012) used MorphInd to build an Indonesian dependency treebank. Rahutomo et al. (2018) used MorphInd as a sub-system of an Indonesian automatic grammar checker. MorphInd has also been widely used to annotate Indonesian corpora such as the IDENTIC Corpus (Larasati, 2012), the TUFS Asian Language Parallel Corpus or TALPco (Nomoto et al., 2018), and Malindo CONC[5] (Nomoto et al., 2018).

## 3. Evaluation of LM

The two tagsets utilised in MorphInd, namely the 'lemma' tagset and the 'morphological' tagset, constitute LM as an annotation scheme. The structure of the annotation layout, which I

[3] http://septinalarasati.com/MorphInd/
[4] http://bahasa.cs.ui.ac.id/resources.php (accessed 06/12/2019)
[5] https://malindoconc.lagoinst.info/concordance/ind/

Published in Corpora 2021 Vol. 16(3)
https://euppublishing.com/doi/abs/10.3366/cor.2021.0221
Print ISSN: 1749-5032 ; Online ISSN: 1755-1676

also evaluate, are exemplified in this paper using a layout adapted from the current version of MorphInd's output (v.1.4).

### 3.1 'Lemma' tagset

In linguistics, the term lemma is closely related to inflection. While Prentice (1976:33) believes that a small number of Indonesian affixes are inflectional, Musgrave (2001:5) argues that Indonesian morphology is exclusively derivational. Instead of categorising an affix as inflectional or derivational, most Indonesian scholars (e.g. Kridalaksana, 1989; Alwi et al., 1998) typically analyse perform morphological analysis by dividing polymorphemic words into roots and affixes and categorise the affixes in terms of their formal morphological criteria, and their functions.

While the title 'lemma' seems to acknowledge the presence of inflections in Indonesian, LM's further details seems to agree with the view of most Indonesian scholars. In LM, the Indonesian words in Table 1 are considered to have the same 'lemma', *cuci* 'to wash'. This is linguistically inaccurate; these words have the same root *cuci* 'to wash', not the same lemma. Examples 1 and 2 in Table 1 are considered to be within the same lemma, because the prefixes *men-* and *di-* are considered inflectional in Indonesian (Prentice 1976:193). But the other affixes (if we follow Prentice's view), in 3 example and 4, are derivational, and thus should each be treated as creating a new lemma in the lexicon. I suspect the term 'lemma' is inaccurately used in LM.

| | |
|---|---|
| 1 | *men-cuci*<br>PFX.ACV-wash<br>'wash (v)' |
| 2 | *di-cuci*<br>'PFX.PASS-wash'<br>'be washed (v)' |
| 3 | *cuci-kan*<br>wash-SFX.APPL-CAUS<br>'have something washed for someone (v)' |
| 4 | *pen-cuci-an*<br>CFX.NOMZR-wash-CFX<br>'laundry (n)' |

Table 1. Words which LM treats as part of a lemma 'cuci'

I argue that LM's 'lemma' tags are better seen as root tags, and will refer to them as such. The organisation of the root tags (see table 2) does not fully reflect the usual organisation of the POS categories of Indonesian lexicon. All 17 categories (excluding foreign word, unknown, and punctuation) are treated as major categories even though some of them are obviously

Published in Corpora 2021 Vol. 16(3)
https://euppublishing.com/doi/abs/10.3366/cor.2021.0221
Print ISSN: 1749-5032 ; Online ISSN: 1755-1676

subcategories, as evidenced by accounts including Alwi et al.[6] (1998) or Kridalaksana (2007). This includes the interrogative and personal pronouns (subcategory of pronoun), subordinating and coordinating conjunctions (subcategories of conjunction), and modal and negation (subcategories of adverb).

| | |
|---|---|
| noun (n)<br>personal pronoun (p)<br>verb (v)<br>numeral (c)<br>adjective (q)<br>coordinating conjunction (h)<br>subordinating conjunction (s)<br>foreign word (f)<br>preposition (r) | modal (m)<br>determiner (b)<br>adverb (d)<br>particle (t)<br>negation (g)<br>interjection (i)<br>copula (o)<br>question (q)<br>unknown (x)<br>punctuation (z) |

Table 2. LM's root tagset (adapted from https://septinalarasati.com/morphind/)

### 3.2 'Morphological' tagset

Several features of LM that are claimed to be 'fine grained morphological analyses' are in fact word-level or morphosyntactic analyses. This is evident from the encoding of the analyses as well as overall tagset content. I therefore refer to this as the morphosyntactic tagset.

In LM's morphosyntactic tags, each element of the analysis is represented by a letter; the full tags are decomposable strings in which one letter marks one analysis. For instance, the tag VSA is the decomposable tag for Verb Singular Active. However, some tags are shorter than three letters; see table 3.

| **First letter** | **Second letter** | **Third letter** |
|---|---|---|
| Noun (N) | Plural (P)<br>Singular (S) | Feminine (F)<br>Masculine (M)<br>Non specified (D) |
| Personal pronoun (P) | Plural (P)<br>Singular (S) | First person (1)<br>Second person (2)<br>Third person (3) |

[6] Alwi et al.'s (1998) work is a reference grammar of Indonesian by *Badan Bahasa*, a formal government institution for language development in Indonesia

Published in Corpora 2021 Vol. 16(3)
https://euppublishing.com/doi/abs/10.3366/cor.2021.0221
Print ISSN: 1749-5032 ; Online ISSN: 1755-1676

| | | |
|---|---|---|
| Verb (V) | Plural (P)<br>Singular (S) | Active (A)<br>Passive (P) |
| Numeral (C) | Cardinal (C)<br>Ordinal (O)<br>Collective (D) | None |
| Adjective (A) | Plural (P)<br>Singular (S) | Positive (P)<br>Superlative (S) |
| Coordinating conjunction (H)<br>Subordinating conjunction (S)<br>Foreign word (F)<br>preposition (R)<br>modal (M)<br>determiner (B)<br>adverb (D)<br>particle (T)<br>negation (G)<br>interjection (I)<br>copula (O)<br>question (Q)<br>unknown (X)<br>punctuation (Z) | None | None |

Table 3. LM's morphosyntactic tagset (adapted from https://septinalarasati.com/morphind/)

The analysis encoded by the first letter uses the same categories as the root POS tag, but whereas in the root tagset for the POS is that of the root morphemes, POS tags in the morphosyntactic tagset apply to complete words, which may or may not share the POS of their root. My evaluation of the POS tags (encoded in the first letter) is therefore not distinct from my evaluation for POS tags in root tagset. What, though, of the analyses that the other letters (claimed to be fine-grained) encode?

Several analyses encoded by second and third letters are linguistically inaccurate. For instance, LM includes the features of number (singular/plural) and person in pronoun tags. While this is proper practice for languages such as Finnish (Koskenniemi, 1983), Arabic (Ryding, 2005), or Turkish (Goksel, 2004), in which because person and number affect verb agreement among other features of morphosyntax, in Indonesian, person and number are lexical, not a grammatical, features, as they have no effect on inflection and are not expressed affixally..

Correspondingly, PM also treats number as a property of verbs. This is equally inaccurate because, in contrast to the languages mentioned above, Indonesian verbs lack number agreement completely.

The 'non-specified' value for the third letter of noun tag refers to gender. Just as with person and number, there is no productive grammatical gender in Indonesian, unlike languages

Published in Corpora 2021 Vol. 16(3)
https://euppublishing.com/doi/abs/10.3366/cor.2021.0221
Print ISSN: 1749-5032 ; Online ISSN: 1755-1676

like French or German. Two Sanskrit loan-suffixes, *-wan* and *-wati*, were used for masculine and feminine in earlier periods, but are no longer productive. In *Kamus Besar Bahasa Indonesia* or KBBI, whic the standard dictionary of Indonesian[7], words with these suffixes, e.g. *wartawan* and *wartawati* (see table 4), are treated as monomorphemic.

| Gloss | *warta-wan*<br>news-AgntNomzr.Male<br>'male reporter' | *warta-wati*<br>news-AgntNomzr.Female<br>'female reporter' |
|---|---|---|
| Input | *wartawan* | *wartawati* |
| LM output | warta<n>+wan_NSM | warta<n>+wati_NSF |

Table 4. Analysis of *wartawan* and *wartawati*

One major problem with LM's morphosyntactic analysis is that only a handful of common functional features of Indonesian morphemes are captured. The first is voice, but only active and passive are available in the tagset; the second is adjective degree, limited to superlative. Morphemes that mark equative degree, reciprocal voice, repetitive mood, and agentive and instrumental nominalisation are commonly found in Indonesian (Alwi et al., 1998; Sneddon et al., 2010), but these features are absent from LM.

It seems that LM has a single tag for the element *=nya* or *–nya*, namely PS3 (personal pronoun 3rd person). This only accommodates one of two distinct forms, that is, *=nya* as the clitic form of the 3rd personal pronoun *dia*, but not *–nya* as a definite suffix (Mueller, 2007:1212). Thus, under LM, MorphInd could only tag both *=nya* and *–nya* with PS3, even though this is inaccurate for *–nya*.

Reduplication was analysed as a feature of form in the initial version of LM, but in the most recent version (v.1.4), this has been replaced by a functional analysis, so that reduplicated nouns are tagged as plural (e.g. *orang* 'person' > *orang-orang* 'people'). I argue that analysing Indonesian reduplication according to formal morphological criteria is more reasonable because not all reduplications (even those with the same phonetic pattern) mark plurality; reduplication also can indicate similarity, variation, or reciprocality (Alwi et al., 1998: 132 ; Sneddon et al., 2010:18-25). Taking it as read that all reduplication indicates the plural is liable to lead to an incorrect analysis.

In contrast, analysing reduplication by form – simply tagging reduplicated elements *as* 'reduplicated' – is a safer option, but one not present in LM. Tags for formal analyses (e.g. prefix, suffix, circumfix, infix, proclitic, enclitic) are of crucial importance to morphological annotation, as users of the annotation might wish to utilise queries based on such formal morphological criteria. However, this kind of tag is absent from LM's morphosyntactic tagset.

[7] KBBI is built by Badan Bahasa, a government institution for language development in Indonesia https://kbbi.kemdikbud.go.id/

Published in Corpora 2021 Vol. 16(3)
https://euppublishing.com/doi/abs/10.3366/cor.2021.0221
Print ISSN: 1749-5032 ; Online ISSN: 1755-1676

### 3.3 Output

LM's layout is unique. It presents all morphemes within polymorphemic words (roots, affixes, and cliticised pronouns/particles) in their canonical or citation form, with plus symbols for morpheme breaks, as shown in Table 5. Root morpheme tags are given as single lowercase root tags surrounded by angle brackets after the root (e.g. *kirim*<v>). The morphosyntactic tag is given following an underscore symbol (e.g. VSA) after the complete chain of the morpheme(s) that form the word. Pronominal clitics are considered as separate words by LM, but *not* split from their base; for this reason, in Table 5, two morphosyntactic tags are given for the clitic pronouns (PS1, PS3), and one for the active voice verb (VSA).

| | |
|---|---|
| Gloss | *ku=meng-(k)irim-kan=nya*<br>1p=ACV-send-APPL=3s<br>'I send him/her (something)' |
| Input | *kumengirimkannya* |
| Output | aku<p>_PS1+meN+kirim<v>+kan_VSA+dia<p>_PS3 |

Table 5. Analysis of *kumengirimkannya*

The most fundamental concern here for morpheme-level analysis is that LM leaves all affixes untagged. This is a disadvantage, as affixation is the most productive word formation process in Indonesian. A few affixes are accommodated by the morphological tagset (as shown in section 3.2). However, these analyses cannot be linked to the affixes which instantiate them. In Table 5, for instance, 'A' in the tag VSA analyses the verb as having active voice. The active voice is encoded specifically by prefix *meng-*, but the analysis 'A' is merged within the morphosyntactic tag VSA for *mengirimkan*, and is not linked to the prefix *meN-*.

LM annotation includes morphemes in their citation form, not the actual orthographic form (allomorph). In *kumengirimkannya*, four out of five morphemes (*ku=*, *meng-*, *irim*, and *=nya*) are present in the output in citation form only (*aku, meN, kirim* and *dia* respectively). Only one, *-kan*, has a citation form identical to its orthographic form in this word. The morphophonological processes behind these alternations are less important than the fact that the orthographic forms omitted by LM's layout thus cannot be used as criteria in queries.

One fundamental concern regarding LM's morpheme segmentation is that it does not distinguish prefix-suffix combinations from circumfixes. This distinction is crucial in Indonesian (Alwi et al., 1998:31; Sneddon et al., 2010:xxi; Chaer, 2008:23; Kridalaksana, 1989:28). In LM's layout, *kejatuhan* 'fall (n)' (Table 6) is segmented in exactly the same way as *mengirimkan* 'to send something' (the form in Table 5 minus clitic pronouns). However, *ke--an* is a circumfix, while *meng-* and *-kan* are a prefix-suffix combination; prefix and suffix are together in this word, but need not always be.

It is not possible to distinguish circumfixes from prefix-suffix combinations by reference to the position of the morpheme breaks, because in both cases there is one break directly before

Published in Corpora 2021 Vol. 16(3)
https://euppublishing.com/doi/abs/10.3366/cor.2021.0221
Print ISSN: 1749-5032 ; Online ISSN: 1755-1676

and one break directly after the root. If an annotation scheme explicitly classifies *ke--an* as a circumfix and *meng-* and *-kan* as a prefix and suffix, the issue is avoided. However, as mentioned in 3.2, this approach has no place in LM, which lacks formal morphological tags.

| | |
|---|---|
| Gloss | *ke-jatuh-an*<br>CFX.Nomzr-to fall-CFX.Nomzr<br>'the fall' |
| Input | *kejatuhan* |
| Output | ke+jatuh<v>+an_NSD |

Table 6. Analysis of *kejatuhan*

In Indonesian, the parts of a reduplication are orthographically linked by a hyphen (see Table 7). In LM layout, such reduplicated forms are presented as a single root form, omitting one of the parts. This would be disadvantageous for users searching for reduplicated words by their orthographic form. Users would have to query, for instance, *buku-buku* 'books' by searching for *buku* with a plural tag (NPS). A query for orthographic *buku-buku* would not yield any results from a text or corpus annotated using the LM layout.

| | |
|---|---|
| Gloss | *buku-buku*<br>book-RED.pl<br>'books' |
| Input | *buku-buku* |
| Output | buku<v> _NPD |

Table 7. Analysis of *buku-buku*

The reduplication-affixation interface is not yet fully accommodated yet in LM. The existing LM annotation produced by MorphInd, illustrated in Table 8, lays out the word *melempar-lemparkan* 'to throw something repeatedly' (reduplicated root plus prefix and suffix marking voice) as two separate word tokens, which, as the affixation pattern shows, is definitely not the case (affixation affects word bases, not two-word sequences). In this case, the error is expected but it is not caused by MorphInd program. Rather, it comes from LM, the scheme that MorphInd implements.

| | |
|---|---|
| Gloss | *me-lempar-lempar-kan*<br>PFX.Acv-throw-RED.Itrv-SFX<br>'to throw something repetitively' |
| Input | *melempar-lemparkan* |
| Output | me+lempar<v>_VSA lempar<v>+kan_VSA |

Published in Corpora 2021 Vol. 16(3)
https://euppublishing.com/doi/abs/10.3366/cor.2021.0221
Print ISSN: 1749-5032 ; Online ISSN: 1755-1676

Table 8. Analysis of *melempar-lemparkan*

## 4. Overall evaluation

The advantages and drawbacks of LM are summarised in table 9 and 10. Not all of these points require further comment, but those that do are addressed in the discussion that follows. That LM morphological analysis tags, including word POS labels, are linked to the whole word token (see section 3.2), is characteristic of a morphosyntactic or POS tagger. This has the advantage that MorphInd can be used for POS tagging, even though it is a morphological analyser.

MorphInd and LM tokenise Indonesian words into a variety of morphemes (root, affix, clitic, particle) as shown in point 2 of Table 9. The previous state-of-the-art Indonesian morphological analyser built by Pisceldo et al. (2008) does not handle clitics and particles (section 2). The improved tokenisation scheme in LM allows searches to directly reference a wider variety of morphological tokens.

| | Evaluation | Implication |
|---|---|---|
| 1 | Morphological tags are not linked to specific morphemes but merged as one tag for the whole word token (3.3). | Enables morphosyntactic searches at word level |
| 2 | Words are tokenised into affixes, roots, clitics, and particles (3.3). | User can identify (citation forms of) affixes and roots |
| 3 | Roots are POS tagged separately from word tokens (2) | Enables POS tag searches at root level (and allows lemmatisation to be drawn from LM annotation). |
| 4 | All analyses are unambiguous (1) | Annotation accuracy is easy to assess |

Table 9. Advantages of LM, cross-referenced to foregoing discussion

Despite the above-mentioned positive evaluations, however, and even assuming flawless implementation of the analysis scheme, LM cannot capture information necessary to serve a number of needs which we may anticipate users of morphological tagging to have, as presented in Table 10.

Published in Corpora 2021 Vol. 16(3)
https://euppublishing.com/doi/abs/10.3366/cor.2021.0221
Print ISSN: 1749-5032 ; Online ISSN: 1755-1676

| | Evaluation | Implication |
|---|---|---|
| 1 | Unusual organisation of POS tag hierarchy, and use of analytic categories not present in Indonesian reference grammars (e.g. determiner, copula, gender) (3.2) | Disadvantageous to likely users whose understanding of morphological categories is likely to be based largely or wholly on such Indonesian reference grammars |
| 2 | Many of the morphological analyses encoded in the 2nd and 3rd letters of the morphological tagset are linguistically inaccurate (3.2) | The inaccurate codes does not have any actual use when implemented |
| 3 | A number of common functional categories in Indonesian are absent from the tags (e.g. reciprocal voice, equative degree, etc.) (3.2) | Users cannot undertake searches involving these categories |
| 4 | No formal category tags (e.g. prefix, suffix, circumfix, proclitic, enclitic etc.) (3.2) | |
| 5 | Affixes are left untagged. Some analyses of the affixes are found, but are merged in the morphosyntactic tag (3.3) | Users cannot link affixes to the corresponding analysis. Functional searches must target words instead of the specific morpheme |
| 6 | Some orthographic forms (allomorphs) are not present in the layout of LM analysis (3.3) | Users cannot search the data by orthographic form |
| 7 | No distinction between prefix-suffix combination and circumfix (3.3) | Users cannot create queries which unambiguously include or exclude circumfixes, which are importantly distinct from prefix-suffix pairs in Indonesian |
| 8 | Reduplication without affixation is treated as a single token with the base given in non-reduplicated form (3.3) | Users cannot search for reduplication in the hyphenated orthographic form |
| 9 | Reduplication with affixation is treated as two distinct word tokens. (3.3) | The analysis is both inaccurate and problematic for the composition of queries (no explicit indication that reduplication has taken place) |
| 10 | Analysis is unambiguous (1) | The selected analysis is not necessarily the correct analysis. |

Table 10. Drawbacks of LM, cross-referenced to foregoing discussion

That LM does not reflect the usual POS organisation for Indonesian (point 1) might be because of the influence of the Penn Treebank tagset (Taylor et al., 2003), which Larasati et al., (2011:122) claim to be their inspiration. The presence of features not relevant to Indonesian, such as number, person or gender, (point 2) and absence of other features that are (point 3) is a

Published in Corpora 2021 Vol. 16(3)
https://euppublishing.com/doi/abs/10.3366/cor.2021.0221
Print ISSN: 1749-5032 ; Online ISSN: 1755-1676

major limitation to the utility of LM. Add to that point 4, that no formal morphological features are annotated in LM, and we may conclude that relevant features that users might expect are absent, while some existing features are likely to be of little use. Missing orthographic forms (point 6) means requiring Morphind users to have a correct understanding of the citation forms of all the forms they wish to query. Point 8 in the table is that users cannot search for reduplicated words, but there's another issue, namely, that it is an overgeneralisation to tag all reduplications as plural

Overall, taking all of table 9 and 10 into account, LM is better than earlier morphological annotation scheme for Indonesian (Larasati et al., 2011:120). However, there are a number of substantial issues as I have pointed out in table 10. The most critical issues are the linguistic inaccurateness and the absence of linguistic elements critical for user's query. This means the true nature of Indonesian morphology is not accurately and not completely portrayed.

**Acknowledgment**
I sincerely thank the Indonesia Endowment Fund for Education (*Lembaga Pengelola Dana Pendidikan Indonesia*[8]) for fully sponsoring this work as part of my PhD study (No:20161012119558).

**References**

Alwi, H., Dardjowidjojo, S., Lapoliwa, H., & Moeliono, M. (1998). *Tata Bahasa Baku Bahasa Indonesia (3rd Edition).* Jakarta: Balai Pustaka.

Chaer, A. (2008). *Morfologi Bahasa Indonesia.* Bandung: Rhinneka Cipta.

Dinakaramani, A., Rashel, F., & Luthfy, A. (2014). A rule-based Indonesian POS tagger and manually tagged 250k-word corpus. *Workshop/Hackathon for the Wordnet Bahasa (2014).* Singapore: NTU Press.

Göksel, A., & Kerslake, C. (2004). *Turkish: A comprehensive grammar.* New York: Routledge.

Green, N., Larasati, S.-D., & Žabokrtský, Z. (2012). Indonesian dependency treebank: Annotation and parsing. *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation* (pp. 137-145). Bali: Universitas Indonesia.

Koskenniemi, K. (1983). *Two-Level Model for Morphological Analysis (Thesis).* Helsinki: University of Helsinki.

Kridalaksana, H. (1989). *Pembentukan Kata dalam Bahasa Indonesia.* Jakarta: Gramedia.

Kridalaksana, H. (2007). *Kelas Kata dalam Bahasa Indonesia Edisi ke Lima.* Jakarta: Gramedia Pustaka Utama.

Larasati, S.-D. (2012). IDENTIC CORPUS: Morphologically Enriched Indonesian-English Parallel Corpus. *LREC*, pp. 902-906.

Larasati, S.-D., Kuboň, V., & Zeman, D. (2011). Indonesian Morphology Tool (MorphInd): Towards an Indonesian Corpus. *Systems and Frameworks for Computational Morphology* (pp. 119-129). Zurich, Switzerland: Springer Berlin Heidelberg.

Lewis, M.-P., Simons, G.-F., & Fennig, C.-D. (2009). *Ethnologue: Languages of the world (Vol. 16).* Dallas: SIL International.

[8] https://www.lpdp.kemenkeu.go.id/

Published in Corpora 2021 Vol. 16(3)
https://euppublishing.com/doi/abs/10.3366/cor.2021.0221
Print ISSN: 1749-5032 ; Online ISSN: 1755-1676

Mueller, F. (2007). Indonesian Morphology. In A. Kaye, *Morphology of Asia and Africa* (pp. 1207-1231). Indiana: Eisenbraus.

Musgrave, S. (2001). *Non-Subject Argument in Indonesian.* Melbourne: Ph.D Thesis.

Nomoto, H., Choi, H., Moeljadi, D., & Bond, F. (2018). MALINDO Morph: Morphological dictionary and analyser for Malay/Indonesian. *Proceedings of The 13th Workshop on Asian Language Resources* (pp. 36-43). Miyazaki: LREC.

Nomoto, H., Kenji, O., Moeljadi, D., & Hideo, S. (2018). TUFS Asian Language Parallel Corpus (TALPCo). *Proceedings of the Twenty-Fourth Annual Meeting of the Association for Natural Language Processing* (pp. 436-439.). Tokyo: TUFS.

Pisceldo, F., Mahendra, R., Manurung, R., & Arka, I.-W. (2008). A Two Level Morphological Analyser for the Indonesian Language. *Australasia Technology Association Workshop*, (pp. 142-150).

Prentice, D. (1987). Malay (Indonesian and Malaysian). In B. Comrie, *The Major Languages of East and Southeast Asia* (pp. 185-208). London: Routledge.

Rahutomo, F., Mulyo, A.-S., & Saputra, P.-Y. (2018). Automatic Grammar Checking System For Indonesian. *International Conference on Applied Science and Technology (iCAST)* (pp. 308-313). Sulawesi: IEEE Indonesia Section.

Ryding, K.-C. (2005). *A reference grammar of modern standard Arabic.* Cambridge: Cambridge university press.

Sneddon, J.-N., Adelaar, A., Djenar, D.-N., & Ewing, M.-C. (2010). *Indonesian reference grammar:2nd Edition.* New South Wales: Allen & Unwin.

Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn treebank: an overview. *Treebank*, pp. 5-22.