MALINDO Morph: Morphological dictionary and analyser for Malay/Indonesian

Hiroki Nomoto⋆, Hannah Choi◦, David Moeljadi◦, Francis Bond◦
⋆Tokyo University of Foreign Studies
3-11-1 Asahi-cho, Fuchu, Tokyo 183-8534 Japan
nomoto@tufs.ac.jp
◦Nanyang Technological University
14 Nanyang Drive, Singapore 637332
YUNJUNG001@e.ntu.edu.sg, davidmoeljadi@gmail.com, bond@ieee.org

Abstract

Malay/Indonesian lacked an open wide-coverage dictionary that can be used for both NLP tasks and non-NLP purposes. The MALINDO Morph morphological dictionary is the first such dictionary. It provides morphological information (root, prefix, suffix, circumfix, reduplication) for roughly 232K surface forms. The entry forms are those found in the authoritative dictionaries in Malaysia (*Kamus Dewan*[4]) and Indonesia (*Kamus Besar Bahasa Indonesia*[5]) (core dictionary) as well as frequent words in the Leipzig Corpora Collection (Goldhahn et al., 2012) (expanded dictionary). The morphological analyses were checked by hand for all surface forms, except for (i) basic and *di-* forms in the expanded dictionary whose existence is predicted from the corresponding *meN-* active forms in the core dictionary and (ii) the case variants of the items in the core dictionary. This paper also discusses the morphological analyser that we developed to create our morphological dictionary. Our morphological analyser is more linguistically rigorous than previous morphological analysers and stemmers/lemmatizers such as MorphInd (Larasati et al., 2011) because it takes into account circumfixes, which have previously been neglected, largely due to a misunderstanding among NLP researchers that circumfixes are no more than combinations of a prefix and a suffix.

Keywords: Malay/Indonesian, morphological dictionary, morphological analyser

1. Introduction

A good dictionary with wide coverage is crucial to the success of a robust morphological analysis, which in turn becomes the basis for higher-level tasks such as syntactic parsing. While open dictionaries such as the NAIST Japanese Dictionary[1] and UniDic[2] are available for Japanese, nothing comparable exists for Malay/Indonesian. Hence, we created a morphological dictionary for Malay/Indonesian. This paper describes our dictionary and the morphological analyser that we developed for its creation. Both the dictionary and morphological analyser will be made publicly available at `https://github.com/matbahasa/MALINDO_Morph`, licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

This paper is organized as follows. First, we present a brief overview of the Malay and Indonesian languages (section 2) and their morphology (section 3). Section 4 summarizes previous work on dictionaries for NLP tasks and on stemmers, lemmatizers and morphological analysers for Malay/Indonesian. The tools that we have developed, the MALINDO Morph morphological dictionary and morphological analyser, are described in section 5. Section 6 concludes the paper and discusses ways of using the MALINDO Morph dictionary for NLP and non-NLP purposes. It also suggests ways in which the MALINDO Morph dictionary can be enriched in the future.

2. Malay and Indonesian

The "Malay" language (ISO693-3 msa), from the Austronesian language family, is the official language of four Southeast Asian countries in different parts of the Malay Archipelago. There are two regional varieties of the same language, namely Malay in the narrow sense (ISO693-3 zsm), used in Malaysia, Brunei and Singapore, and Indonesian (ISO693-3 ind), used in Indonesia. In this paper, we refer to the Malay language in the narrow sense simply as "Malay."

Many tools and resources are available that have been independently developed in each region, including standard dictionaries and language resources. In addition, some collaboration has occurred, such as the Majlis Bahasa Brunei-Indonesia-Malaysia (Language Council of Brunei-Indonesia-Malaysia) or MABBIM, a regional language organization whose role is to plan and monitor the development of the Malay/Indonesian language in the region, with Singapore as an observer. While some variations exist between the two languages, they are mutually intelligible, with only about 10% of lexical difference (Asmah, 2001). The two languages also share the same set of affixes. As such, a morphological dictionary can be developed that covers both Malay and Indonesian.

3. Malay/Indonesian Morphology

Malay/Indonesian is an agglutinating language whose morphology involves the use of affixation, reduplication and cliticization.[3] It has productive prefixes, suffixes and cir-

[1] `https://ja.osdn.net/projects/naist-jdic/`

[2] `http://pj.ninjal.ac.jp/corpus_center/unidic/`

[3] A comprehensive description of these processes can be found, among others, in Abdullah (1974), Asmah (2009) and Sneddon et al. (2010).

cumfixes, which can be either derivational or inflectional. It also has infixes, but they are no longer productive. Productive reduplication is achieved through full reduplication of stems (e.g. *kucing* 'cat' → *kucing-kucing* 'cats'). Its semi-productive morphological processes include rhythmic reduplication, which involves vowel and/or consonant alternation (e.g. *gunung* 'mountain' → *gunug-ganang* 'mountain range'). Partial reduplication, which adds the base-initial consonant plus *e* to the base, is semi-productive at best in Indonesian but somewhat productive in Colloquial Malay (e.g. *mula* 'to start' → *memula* 'at first' (= *mula-mula*)). The clitics consist of proclitics (e.g. *ku*= 'I') and enclitics (e.g. =*ku* 'me/my').

The interaction of different morphological processes can give rise to quite a complex word structure. For example, *keterbatasan-keterbatasan* 'limitations' is derived from the root *batas* 'limit', as shown in Figure 1. Notice that the relative order between affixation and reduplication is not fixed. The reverse order is also possible, as illustrated by *keanak-anakan* 'childishness' in the same figure.

4. Existing Tools and Their Problems

4.1. Morphological Dictionary

No large dictionary file is publicly available in an accessible format. The Malay tokenizer/lemmatizer described in Baldwin and Su'ad (2006) has a small dictionary file, which consists of word-lemma-POS (part of speech) triples for 2,499 words.[4]

One can create a larger dictionary by using the data from online dictionaries (not specifically for NLP) such as *Dr. Bahnot's Malay-English Cyber-Dictionary*[5] and *Kateglo ∼ Kamus, tesaurus, dan glosarium bahasa Indonesia*.[6] The latter takes most of its data from the third edition of *Kamus Besar Bahasa Indonesia* and provides an API to access its structured data under a CC BY-NC-SA 3.0 license. However, to the best of our knowledge, no existing dictionary contains the kinds of morphological information that our dictionary offers: affixes (prefixes, suffixes, circumfixes), clitics (proclitics, enclitics) and reduplication types.

4.2. Morphological Analyser

Much work has been done in the past on stemmers/lemmatizers for Malay/Indonesian (see, for example, Baldwin and Su'ad (2006), Adriani et al. (2007), Larasati et al. (2011), Mohamad Nizam et al. (2016) and the studies cited therein). Not mentioned in these papers but notable is the Sastrawi stemmer,[7] which uses Kateglo (see section 4.1) as its dictionary and is offered in multiple languages, namely PHP, Java, C, Python, Go and Ruby.

Morphological analysers analyse the non-stem/lemma strings of a word in addition to identifying the stem/lemma. Currently, MorphInd (Larasati et al., 2011) seems to be the most sophisticated morphological analyser for Malay/Indonesian. It identifies morpheme boundaries and assigns two POS tags to a token: one for the lemma ('lemma tag') and another for the entire token ('morphological tag'). For instance, the verb *mengirim* 'to deliver', which is derived from the root *kirim* by attaching the prefix *meN-*, is analysed as `meN+kirim<v>_VSA`. `<v>` is the lemma tag for verbs, whereas `_VSA` is the morphological tag indicating that the entire token is a singular active verb.[8]

There is a common misunderstanding among NLP researchers about Malay/Indonesian morphology, specifically concerning the notion of the 'circumfix' (also called 'confix'). Circumfixes are incorrectly thought of as a combination of a prefix and a suffix. However, a circumfix is in fact a single morpheme that surrounds a stem. It is true that *meN-X-kan* contains the prefix *meN-* and the suffix *-kan*, but one must not describe *meN- -kan* as a circumfix, as a circumfix encodes syntactic and semantic information that cannot be ascribed to the component parts. The meaning of *ke- -an*, which is a genuine circumfix, cannot be obtained by combining the meanings of *ke-* and *-an*.

Presumably due to this misunderstanding, MorphIndo analyses the non-lemma strings, but it does not specify what they are, that is, whether they are a prefix, suffix or circumfix. For example, *pengiriman* (= *kirim* + circumfix *peN- -an*) 'delivery' is analysed as `^peN+kirim<v>+an_NSD$`. From this output, it is not obvious whether *peN* and *an* are a combination of two morphemes (prefix *peN-* and suffix *-an*) or a single morpheme (circumfix *peN- -an*). In fact, the correct identification of circumfixes presents a major challenge to morphological analysis in Malay/Indonesian.[9] This is because the strings appearing in circumfixes constitute a proper subset of those appearing in prefixes and suffixes. A correct circumfix cannot be identified by just looking at the two strings at the left and right edges of a token. Thus, *berakhiran* 'suffixed' can be segmented as `ber-akhir-an`, but the word does not contain the circumfix *ber- -an*. The word is derived from the root *akhir* by attaching the suffix *-an* to derive *akhiran* 'suffix' and then attaching the prefix *ber-* to this derived form. Likewise, *berperadaban* 'civilized', which is segmented as `ber-per-adab-an`, has a circumfix, but it is not *ber- -an* but *per- -an*.

5. MALINDO Morph

5.1. Morphological Dictionary

Size and format The MALINDO Morph morphological dictionary currently has a total of 232,550 lines, with each containing an analysis for one (case-sensitive) token. These 232,550 tokens are based on 78,750 distinct roots. Each line is made up of the following six items, separated by tabs:

[4] https://github.com/averykhoo/malay-toklem/blob/master/lexicons/word-lemma-pos

[5] http://dictionary.bhanot.net/

[6] http://kateglo.com/

[7] https://github.com/sastrawi/sastrawi

[8] Since Malay/Indonesian does not have subject-verb agreement, the number information should in fact be unspecified. Moreover, roots may not need POS tags because roots, unlike surface forms, are abstract entities. This point is clear in languages like Arabic, in which roots are not used as surface forms (e.g. root *k-t-b* 'having to do with writing' → surface forms *kataba* '(he) wrote' (verb), *kitab* 'book' (noun), etc.).

[9] The difficulty involved in distinguishing circumfixes from combinations of a prefix and a suffix has also been noted by Ranaivo-Malançon (2004).

batas	'limit'	*anak*	'child'
↓	(affixation: *ter-*)	↓	(reduplication)
terbatas	'limited'	*anak-anak*	'children'
↓	(affixation: *ke- -an*)	↓	(affixation: *ke- -an*)
keterbatasan	'limitation'	*keanak-anakan*	'childishness'
↓	(reduplication)		
keterbatasan-keterbatasan	'limitations'		

Figure 1: The derivations of *keterbatasan-keterbatasan* 'limitations' and *keanak-anakan* 'childishness'

- Root
- Surface form
- Prefix(es), proclitic: *meN-*, *N-* (Indonesian), *di-*, *per-*, *ber-*, *ter-*, *peN-*, *pe-*, *ke-*, *se-*; *ku=*, *kau=*[10]
- Suffix(es), enclitic(s): *-kan*, *-i*, *-in* (Indonesian), *-an*, *-nya*; *=ku*, *=mu*, *=kau*, *=nya*, *=lah*, *=kah*
- Circumfix(es): *ber- -an*, *ber- -kan*, *ke- -an*, *peN- -an*, *pe- -an*, *per- -an*, *se- -nya*
- Reduplication: Full, Partial, Rhythmic

Some sample lines are shown in Figure 2.
Our dictionary was built in two steps. First, we built a core dictionary with entries from the authoritative dictionaries in Malaysia and Indonesia, respectively: *Kamus Dewan*[4] (KD) and *Kamus Besar Bahasa Indonesia*[5] (KBBI). Then, we created an expanded dictionary for other tokens that are not listed in KD or KBBI. The source of the expanded dictionary was the reclassified version of the Leipzig Corpora Collection (LCC; Goldhahn et al. (2012); Nomoto et al. (under review)). Tables 1 and 2 summarize the current sizes of the two dictionaries and the frequencies of different morphological processes found in them, respectively.

Dictionary	Checked	Unchecked	Total
Core	84,402	0	84,402
Expanded	47,399	100,749	148,148
Total	131,801	100,749	232,550

Table 1: Sizes of the MALINDO Morph dictionaries (unit: line)

Morphology	Core	Expanded	Total
PREFIXES AND PROCLITICS			
meN-	12,336	8,939	21,275
N-	2	147	149
di-	167	10,787	10,954
per-	634	1,146	1,780
ber-	4,936	3,514	8,450
ter-	2,600	2,190	4,790
peN-	2,127	2,345	4,472
pe-	177	313	490
ke-	123	524	647
se-	1,038	1,621	2,659
ku=	0	1,258	1,258
kau=	0	84	84
SUFFIXES AND ENCLITICS			
-kan	4,423	11,565	15,988
-i	1,390	4,151	5,541
-in	0	25	25
-an	2,718	4,785	7,503
-nya/=nya	70	28,716	28,786
=ku	0	4,209	4,209
=mu	0	2,862	2,862
=kau	0	45	45
=lah	13	6,513	6,526
=kah	5	999	1,004
CIRCUMFIXES			
ber- -an	624	239	863
ber- -kan	227	57	284
ke- -an	2,431	4,146	6,577
peN- -an	2,387	3,040	5,427
pe- -an	76	93	169
per- -an	694	1,357	2,051
se- -nya	92	139	231
REDUPLICATION			
Full	3,693	4,908	8,601
Partial	230	105	335
Rhythmic	735	232	967
NO MORPHOLOGY (ROOT = SURFACE FORM)			
None	50,350	71,264	121,614

Table 2: Morphological profile of the MALINDO Morph dictionaries (unit: token)

Core dictionary We extracted all of the roots and their derived forms from KD and KBBI. We did this manually for KD.[11] As for KBBI, we extracted root and surface forms from a database built by a team that includes the third author of the present paper as a member (Moeljadi et al., 2017).

[10]This slot may also include other items occurring before the root, such as the preposition *ke* 'to' as in *mengebumikan* 'to bury' or the negator *tidak* 'not' as in *ketidakcukupan* 'insufficiency'.

[11]In fact, the KD database has been commercialized, and we could have purchased the necessary information from a company. However, we did not do so because the price offered was more expensive than the cost of manual work.

Root	Surface form	Prefix	Suffix	Circumfix	Reduplication
perlu	perlu	0	0	0	0
perlu	seperlunya	0	0	se- -nya	0
perlu	memerlukan	meN-	-kan	0	0
perlu	perlu-memerlukan	meN-	-kan	0	R-full
perlu	keperluan	0	0	ke- -an	0

Figure 2: Morphological analysis for the root *perlu* 'necessary' and its derivatives

The morphological analyses were conducted using Microsoft Excel functions. The results were manually checked by Japanese undergraduate students who had learnt Malay or Indonesian as their major for more than three years, Indonesian research students and the first and second authors of the present paper.

For some words, KD and KBBI assume different morphological analyses. There are also cases in which their analyses are good enough for practical purposes but not very precise as linguistic analyses. In such cases, we adopted our own analyses which we think are adequate linguistically. For example, both KD and KBBI list *anai-anai* 'termite' as a head word of its own, whereas the MALINDO Morph morphological dictionary lists the word as a derivative of the root *anai*. KD and KBBI make a practically reasonable choice, given that the form *anai* is a bound root and is not used by itself. However, for a rigorous linguistic analysis, *anai-anai* should be treated as a fully reduplicated form of the bound root *anai*. With departures like this, our core dictionary is not identical to either KD or KBBI.

Expanded dictionary Sixteen 300K subset files (Malay 3, Indonesian 13) of LCC were used as sources of additional data to expand our dictionary. Each 300K file consists of 300,000 sentences. 1,005,007 word types (case-sensitive) were not found in the core dictionary. They include genuine Malay/Indonesian words, proper names, abbreviations, spelling variants/errors, foreign words and non-alphabets. Out of these words, only frequent ones that occurred at least ten times in one of the sixteen subset files were subjected to further processes.

There were 282,186 such words, of which 57,633 were English words and 76,638 were non-alphabets and were not included in the MALINDO Morph morphological dictionary. The remaining 147,915 words were analysed using the morphological analyser described below.[12] The results of the morphological analysis were checked by hand, except for the basic and *di-* forms (cf. Table 3) as well as the case variants of the items in the core dictionary.

The expanded dictionary also contains words in the core dictionary that can also be analysed as involving an enclitic. For example, *masalah* 'problem' and *penanya* 'questioner' are listed in the core dictionary as a root and a *peN-* nominal of the root *tanya* 'ask', respectively. However, they can also be analysed as *masa* 'time' + *-lah* (focus particle) and *pena* 'pen' + *-nya* 'his/her', respectively. These and other analyses were done manually and hence were added to the "checked" category of the expanded dictionary.

Limitations Currently, the MALINDO Morph morphological dictionary only targets productive native affixes and reduplication. This is because they play more important roles compared to non-productive and foreign affixes. Borrowed affixes such as *anti-* 'anti-' and *pra-* 'pre-' are thus not analysed unless they occur together with native ones (e.g. *anti-* in *anti-pemerintah* 'anti-government': perintah anti-pemerintah anti-+peN- 0 0 0).

Moreover, no distinction is made between the suffix *-nya* (forming adverbials, nominalizing verbs and adjectives, occurring in exclamatives) and the enclitic *=nya* (third person pronoun, definite marker). Ideally, the latter should be taken from the word during tokenization. However, tokenizing the enclitic *=nya* without overtokenizing the suffix *-nya* seems almost impossible without referring to a dictionary.

5.2. Morphological Analyser

Preparation First, we built a list of roots (rootlist) and a hypothetical dictionary (hyp-dic) consisting of the basic and *di-* passive forms corresponding to the *meN-* verbs in the core dictionary (core-dic). Basic forms are verb stems without the active voice marker *meN-*. They are used in imperative, bare active and bare passive constructions (see Nomoto (2013) for the voice system in Malay/Indonesian). Table 3 illustrates these three verb forms. Malay/Indonesian dictionaries do not list the basic *di-* forms of a verb.[13] However, they can be created automatically by removing the prefix *meN-* from the *meN-* form (basic form) and prefixing *di-* to the resulting form (*di-* form), respectively. The forms thus created are merely hypothetical. Hence, they were added to the expanded dictionary (exp-dic) only if they were found to actually be used in the corpus.

The algorithm Given input W, our morphological analyser works as follows. The 'analysis' in Steps 1–5 is a list of the format $\langle$ affix candidate, root, remaining string before root, remaining string after root, reduplication $\rangle$.

Step 1. If W is a non-alphabet, return analysis $\langle \emptyset, W, \emptyset, \emptyset, \emptyset \rangle$.

Step 2. If W or its lowercase equivalent w is an English word, return analysis $\langle \emptyset, W/w, \emptyset, \emptyset, \emptyset \rangle$.

Step 3. If W/w is in core-dic/hyp-dic, retrieve the line(s) for W/w in core-dic/hyp-dic.

[12] This number is smaller than that reported in Table 1 above because some words are morphologically ambiguous, with two or more possible analyses.

[13] The exceptions include Asakura (1963), Quinn (2001), Nomoto (2016) and Florentina (2017).

Basic (= stem)	*MeN-* active	*Di-* passive	Common morphology
ajar 'to teach'	*mengajar* 'to teach'	*diajar* 'to be taught'	Root *ajar*
ajarkan 'to teach (for)'	*mengajarkan* 'to teach (for)'	*diajarkan* 'to be taught (for)'	Root *ajar* + suffix *-kan*
pelajari 'to learn'	*mempelajari* 'to learn'	*dipelajari* 'to be learnt'	Root *ajar* + prefix *per-* + suffix *-i*

Table 3: Verbal inflection in Malay/Indonesian

Step 4. Strip W/w of clitic strings. If the resulting form r is in core-dic/hyp-dic, retrieve the line(s) for r in core-dic/hyp-dic and add the clitic information.

Step 5. Generate candidate sets $Cand_c$, $Cand_p$ and $Cand_s$, where $Cand_a$ is a set of candidate analyses for token w based on affix/clitic type $a \in \{c$(ircumfix), p(refix/proclitic), s(uffix/enclitic)$\}$.

Step 6. Search the direct product $Cand_c \times Cand_p \times Cand_s$ for members whose elements are mutually compatible.

Step 7. Return $\langle root_c, w, p\text{-}, \text{-}s, c_1\text{-}\ \text{-}c_2, red_c\rangle$ for every such member.

The notion of mutual compatibility among analyses invoked in 6 is defined as follows:

Definition Three lists, $\langle c_1\text{-}\ \text{-}c_2, root_c, start_c, end_c, red_c\rangle$, $\langle p\text{-}, root_p, start_p, end_p, red_p\rangle$ and $\langle \text{-}s, root_s, start_s, end_s, red_s\rangle$, are mutually compatible if and only if all of the conditions below are satisfied:

1. $root_c = root_p = root_s$
2. $red_c = red_p = red_s$
3. $start_c = p$
4. $start_p = c_1$
5. $start_s = c_1 + p$
6. $end_c = s$
7. $end_s = c_2$
8. $end_p = c_2 + s$

Example Let us consider an example from core-dic: *sedianya* 'actually'. The word is made up of the root *sedia* and the suffix *-nya*. Suppose that this form did not exist in core-dic. The word is neither a non-alphabet (Step 1) nor an English word (Step 2. It is not in hyp-dic (Step 3). It contains a clitic string, namely *nya*. So, by Step 4, we remove *nya* and check the dictionaries to determine whether the remaining string *sedia* exists in them. It actually does; core-dic has this line: `sedia sedia 0 0 0 0`. Incorporating the clitic information into this, our morphological analyser will return $\langle$sedia, sedianya, $\emptyset$, -nya, $\emptyset, \emptyset\rangle$. It is a correct result, although the distinction between the suffix *-nya* and the clitic *=nya* has been lost.[14]

To see how Steps 5–7 work, let us suppose that Step 4 failed for some reason. The candidate sets obtained by Step 5 are:

$$Cand_c = \left\{ \begin{array}{l} \langle \emptyset, \text{sedia}, \emptyset, \text{nya}, \emptyset\rangle, \langle \emptyset, \text{dia}, \text{se}, \text{nya}, \emptyset\rangle, \\ \langle \text{se- -nya}, \text{dia}, \emptyset, \emptyset, \emptyset\rangle \end{array} \right\}$$

$$Cand_p = \left\{ \begin{array}{l} \langle \emptyset, \text{sedia}, \emptyset, \text{nya}, \emptyset\rangle, \langle \emptyset, \text{dia}, \text{se}, \text{nya}, \emptyset\rangle, \\ \langle \text{se-}, \text{dia}, \emptyset, \text{nya}, \emptyset\rangle \end{array} \right\}$$

$$Cand_s = \left\{ \begin{array}{l} \langle \emptyset, \text{sedia}, \emptyset, \text{nya}, \emptyset\rangle, \langle \emptyset, \text{dia}, \text{se}, \text{nya}, \emptyset\rangle, \\ \langle \text{-nya}, \text{sedia}, \emptyset, \emptyset, \emptyset\rangle, \langle \text{-nya}, \text{dia}, \text{se}, \emptyset, \emptyset\rangle \end{array} \right\}$$

Step 6 picks out the following two lists of lists, based on which Step 7 yields the outputs shown to the right of "$\rightarrow$":

1. $\left(\begin{array}{l} \langle \emptyset, \text{sedia}, \emptyset, \text{nya}, \emptyset\rangle, \langle \emptyset, \text{sedia}, \emptyset, \text{nya}, \emptyset\rangle, \\ \langle \text{-nya}, \text{sedia}, \emptyset, \emptyset, \emptyset\rangle \end{array} \right)$
   $\rightarrow \langle \text{sedia}, \text{sedianya}, \emptyset, \text{-nya}, \emptyset\rangle$
2. $\left(\begin{array}{l} \langle \text{se- -nya}, \text{dia}, \emptyset, \emptyset, \emptyset\rangle, \langle \emptyset, \text{dia}, \text{se}, \text{nya}, \emptyset\rangle, \\ \langle \emptyset, \text{dia}, \text{se}, \text{nya}, \emptyset\rangle \end{array} \right)$
   $\rightarrow \langle \text{dia}, \text{sedianya}, \emptyset, \emptyset, \text{se- -nya}, \emptyset\rangle$

Root identification The generation of candidate sets in Step 5 requires root identification. Our morphological analyser contains a root identification function.[15]

Figure 3 shows our root identification algorithm. It is in fact not a simple root identifier; it also identifies reduplication types and pre- and post-root strings. These pieces of information as well as the root information are used for candidate generation in Step 5.

The Hyphen Handler in the algorithm deals with reduplication and other forms with a hyphen, which include hyphenated words (e.g. *Indo-nesia*) and derived words with numeral bases (e.g. *ke-19* '19th', *1990-an* '1990s'). The algorithm of the Hyphen Handler is given in Figure 4.

The "recover root-initial consonant" process recovers a root-initial consonant that does not occur in the surface form as a result of the morphophonological process called 'nasal substitution'. Nasal substitution changes the phone *N* in an affix to a sound that is homorganic to the following consonant, triggering coalescence of the two sounds for native roots starting with a voiceless consonant, as shown in Figure 5.[16] Two outputs are returned: roots with and without a recovered consonant. The Root Well-Formedness Filter filters out items that do not conform to the template for legitimate roots in the language, such as roots with no vowel and roots starting with a complex onset such as *rt*.

[14] We assume that the relevant distinction is handled by a tokenizer.

[15] Alternatively, one can also use existing stemmers/lemmetizers (cf. section 4) for root identification. This is because they normally equate stems/lemmas with roots, even though, strictly speaking, they are distinct units (cf. Table 3).

[16] Coalescence sometimes also occurs with a voiced consonant. See Nomoto (2012) for variations in nasal substitution.

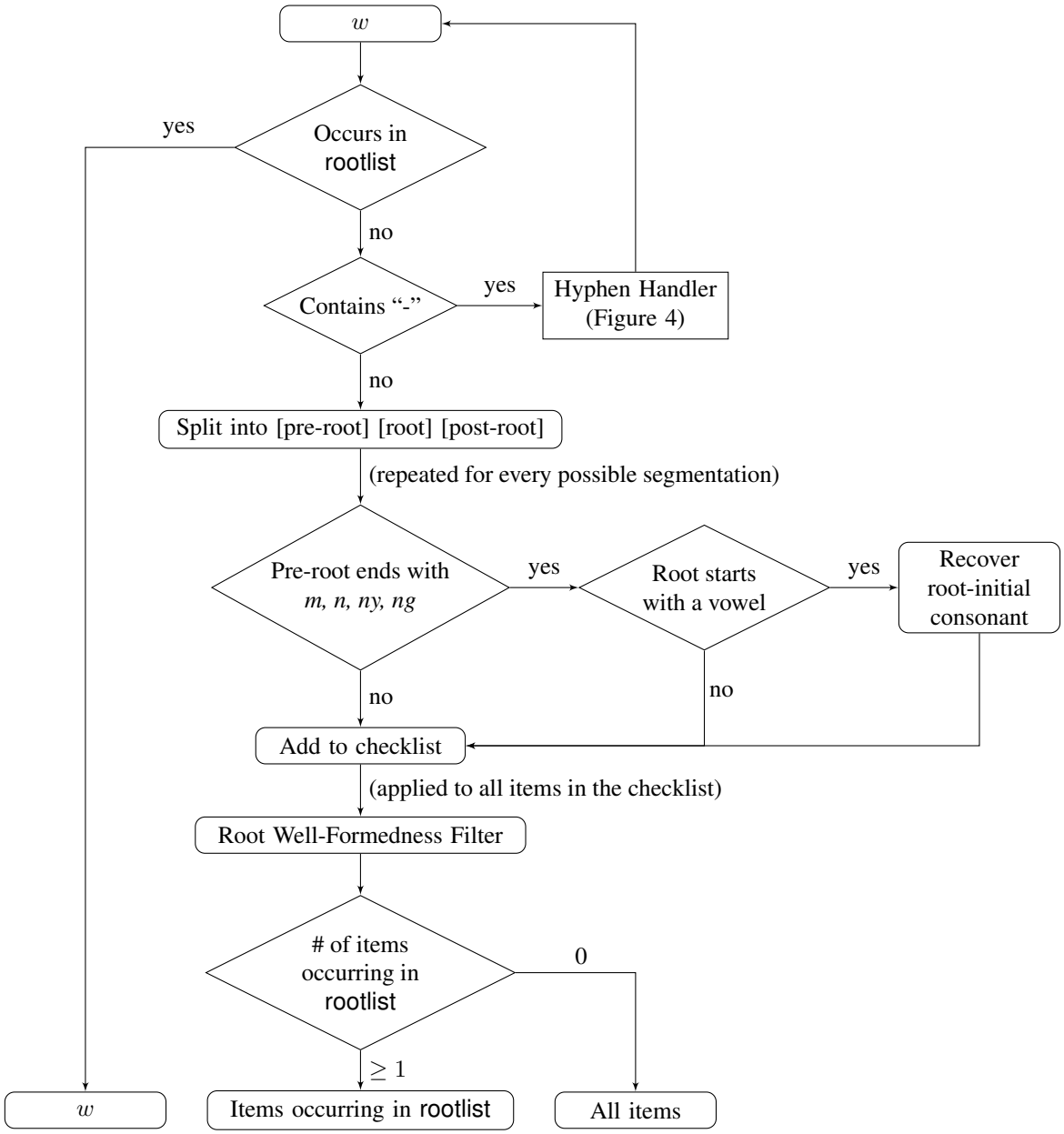

Figure 3: Root identification algorithm

In Figure 6, the root identification algorithm described above is exemplified by the word *mengada-ngadakan* 'to concoct' (= root *ada* + prefix *meN-* + suffix *-kan* + full reduplication). It is a variant of *mengada-adakan* that is not found in either KD or KBBI and hence does not occur in core-dic.

Notice that our root identifier and morphological analyser may return multiple outputs. This is a welcome result because a form can indeed be ambiguous in terms of its morphological composition. For example, a good morphological analyser should be able to come up with the following three analyses for *beruang* in Indonesian:[17]

1. Root *beruang* + no affix 'bear (animal)'
2. Root *ruang* + prefix *ber-* 'to have room'
3. Root *uang* + prefix *ber-* 'to have money'

[17]The third analysis is irrelevant in Malay.

The disambiguation of multiple morphological analyses is only possible when a concrete context is given. Hence, this is a task for a higher level.

6. Conclusions and Future Work

The MALINDO Morph morphological dictionary, with a total of 232,550 lines, will improve the accuracy of stemming/lemmatization in Malay/Indonesian. Stemming/lemmatizing frequent words will become a simple dictionary lookup with an additional disambiguation process for words with ambiguous analyses. The development of stemmers, lemmatizers and root identifiers should then focus on infrequent words. A possible next step to improve them is to incorporate a spell checker, a named entity recognizer and foreign word identifier. As we manually checked the results of the morphological analysis with our morphological analyser (cf. section 5.1), it turned out that infrequent words are often spelling variants/mistakes,

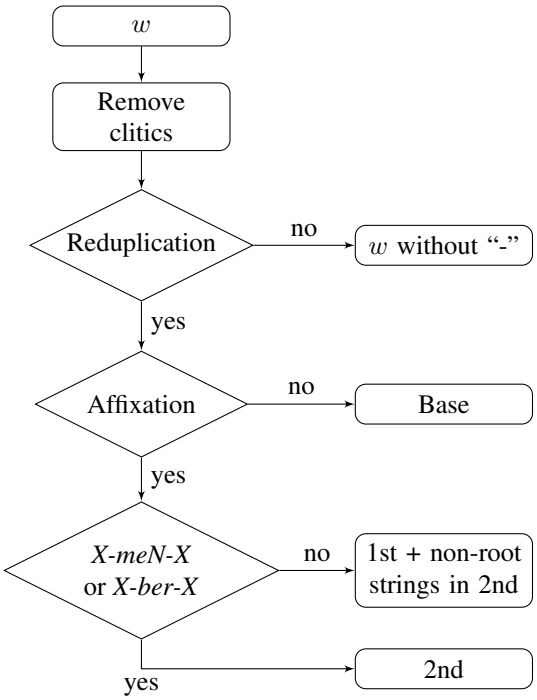

Figure 4: Hyphen Handler algorithm

N-C		Homorganic N		Coalescence
Np	→	mp	→	m
Nb	→	mb	→	m
Nf	→	mf	→	m
Nv	→	mv	→	m
Nt	→	nt	→	n
Nd	→	nd	→	n
Ns(y)	→	ns(y)	→	ny
Nz	→	nz	→	ny
Nc	→	nc	→	ny
Nj	→	nj	→	ny
Nk(h)	→	ngk(h)	→	ng
Ng(h)	→	ngg(h)	→	ng
Nh	→	ngh	→	ng

Figure 5: Nasal substitution

proper names and foreign words, but not a complex combination of productive morphological processes based on known roots.

Furthermore, the MALINDO Morph morphological dictionary provides useful information for other tasks. Parts of speech can be partly predicted from the outermost affix of a word: *meN-* → verb (active), *per- -an* → noun, *se- -nya* → adverb, etc. Specific affixes also provide information about semantics and the argument structure. Words with *peN-* and *peN- -an* are all nouns. However, while the former refers to the external argument (e.g. agent) of the corresponding verb and hence can only take an internal argument (e.g. patient), the latter denotes an action and can take both the external and internal arguments (Nomoto, 2017).

The MALINDO Morph morphological dictionary can also be used for linguistic research. Discoveries of new generalizations regarding the morphological patterns in the language can make a morphological analyser more efficient. The order in which morphological rules are applied is not random. Abdullah (1974) tried to reveal the interaction patterns among them based on 5,153 distinct roots. For instance, he put forward the generalization that "no constructions exceed three layers of affixation" (p. 44). However, no attempt, to our knowledge, has been made to either verify or modify his model based on KD/KBBI-size data using modern computational power.

In the future, the MALINDO Morph dictionary can be enriched by adding more linguistic information. Firstly, the distinction between the suffix *-nya* and the enclitic *=nya* needs to be made in some way (cf. "Limitations" in section 5.1). As a clitic, the latter, by definition, attaches to virtually anything. By contrast, the distribution of the suffix *-nya* is more restricted, though still very wide: (i) it occurs in adverbials (e.g. *biasanya* 'usually' cf. English *-ly*); (ii) it nominalizes verbs and adjectives (e.g. *adanya X* 'the presence of X', cf. Japanese *koto*); (iii) it occurs in exclamatives in Malay (e.g. *sedapnya* 'how delicious(!)', cf. Japanese *koto*). The best way of making this distinction is to flag potential instances of non-suffix *nya*.

Secondly, the information about the variety, i.e. Malay, Indonesian and their dialects, can be added. This can be done by checking whether each word occurs in a corpus of a particular variety. In addition, both KD (Malaysia) and KBBI (Indonesia) contain words that are primarily used in the other variety and indicate such words with special abbreviations. This information can definitely be utilized. However, it must be verified by corpus data before being added to the MALINDO Morph morphological dictionary. The variety information will help to determine a more accurate rate of lexical difference between Malay and Indonesian.

Finally, POSs are another element that a dictionary should provide. As stated above, POSs in Malay/Indonesian can be predicted by morphology, although only partially. Given the large number of lines involved, the POS annotation of the MALINDO Morph morphological dictionary will have to rely on a POS-tagged corpus (ideally the same LCC data that we used) generated by a good POS tagger, the development of which, we believe, should benefit considerably from the MALINDO Morph morphological dictionary.

7. Acknowledgements

The research reported in this paper was conducted under the JSPS grant "Program for Advancing Strategic International Networks to Accelerate the Circulation of Talented Researchers" offered to Tokyo University of Foreign Studies for a project entitled "A Collaborative Network for Usage-Based Research on Less-Studied Languages" as well as the JSPS Grant-in-Aid for Young Scientists (B) (#26770135). We are grateful to JSPS and Nanyang Technological University (NTU) for supporting the first author's stay at NTU.

8. Bibliographical References

KBBI[3] (2001). *Kamus Besar Bahasa Indonesia*. Balai Pustaka, Jakarta, 3rd edn.

KBBI[5] (2016). *Kamus Besar Bahasa Indonesia*. Badan Pengembangan dan Pembinaan Bahasa, Jakarta, 5th edn.

KD[4] (2005). *Kamus Dewan*. Dewan Bahasa dan Pustaka, Kuala Lumpur, 4th edn.

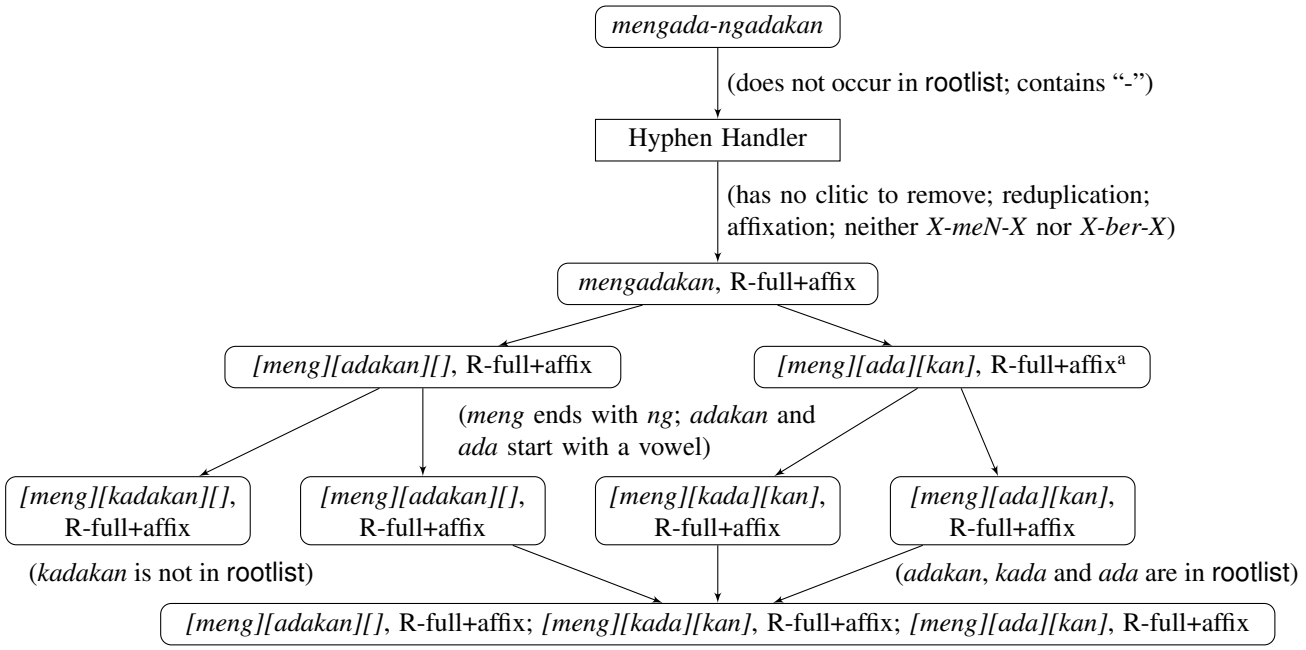

Figure 6: Root identification for *mengada-ngadakan* 'to concoct'
[a] *[][mengadakan][]*, *[][mengada][kan]*, *[me][ngadakan][]* and *[me][ngada][kan]* are also possible but omitted here due to space constraints, as they do not affect the final output.

Abdullah, H. (1974). *The Morphology of Malay*. Dewan Bahasa dan Pustaka, Kuala Lumpur.

Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S. M., and Williams, H. E. (2007). Stemming Indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing (TALIP)*, 6(4):1–33.

Asakura, S. (1963). *Daigakushorin Indoneshiago Shoujiten (Daigakushorin Kamus Bahasa Indonésia-Djepang, Djepang-Indonésia)*. Daigakushorin, Tokyo.

Asmah, H. O. (2001). The Malay language in Malaysia and Indonesia: From lingua franca to national language. *The Aseanists ASIA*, II.

Asmah, H. O. (2009). *Nahu Melayu Mutakhir*. Dewan Bahasa dan Pustaka, Kuala Lumpur, 5th edn.

Baldwin, T. and Su'ad, A. (2006). Open source corpus analysis tools for Malay. In *Proceedings, the 5th International Conference on Language Resources and Evaluation (LREC2006)*, pages 2212–2215.

Florentina, E. (2017). *Pootaburu Nichi-Indoneshia-Ei, Indoneshia-Nichi-Ei Jiten*. Sanshusha, Tokyo.

Goldhahn, D., Eckart, T., and Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Larasati, S. D., Kuboň, V., and Zeman, D. (2011). Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. In Cerstin Mahlow et al., editors, *Systems and Frameworks for Computational Morphology*, pages 119–129. Springer, Verlag.

Moeljadi, D., Kamajaya, I., and Amalia, D. (2017). Building the Kamus Besar Bahasa Indonesia (KBBI) database and its applications. In Hai Xu, editor, *Proceedings of the 11th International Conference of the Asian Association for Lexicography*, pages 64–80.

Mohamad Nizam, K., Mohd Aizaini, M., Anazida, Z., and Amirudin, A. W. (2016). Word stemming challenges in Malay texts: A literature review. In *2016 4th International Conference on Information and Communication Technology (ICoICT)*, pages 1–6, May.

Nomoto, H., Akasegawa, S., and Shiohara, A. (under review). Reclassification of the Leipzig Corpora Collection for similar languages: Malay and Indonesian.

Nomoto, H. (2012). More on Austronesian nasal substitution. In M. Ryan Bochnak, et al., editors, *Proceedings from the 45th Annual Meeting of the Chicago Linguistic Society*, volume 1, pages 503–517, Chicago, IL.

Nomoto, H. (2013). On the optionality of grammatical markers: A case study of voice marking in Malay/Indonesian. In Alexander Adelaar, editor, *Voice Variation in Austronesian Languages of Indonesia*, volume 54 of *NUSA*, pages 121–143.

Nomoto, H. (2016). *Pootaburu Nichi-Maree-Ei, Maree-Nichi-Ei Jiten*. Sanshusha, Tokyo.

Nomoto, H. (2017). Sintaksis nominalisasi bahasa Melayu. [The syntax of Malay nominalization]. In Rogayah Abd. Razak and Radiah Yusoff, editors, *Aspek Teori Sintaksis Bahasa Melayu*, pages 71–117. Dewan Bahasa dan Pustaka, Kuala Lumpur.

Quinn, G. (2001). *The Learner's Dictionary of Today's Indonesian*. Allen & Unwin, Sydney.

Ranaivo-Malançon, B. (2004). Computational analysis of affixed words in Malay language. Paper presented at the 8th International Symposium on Malay/Indonesian Linguistics (ISMIL).

Sneddon, J. N., Adelaar, A. K., Djenar, D. N., and Ewing, M. (2010). *Indonesian: A Comprehensive Grammar*. Routledge, London, 2nd edn.