Discourse functions of the two non-active voices in Indonesian: Based on the web corpus data in MALINDO Conc

Asako SHIOHARA, Yuta SAKON and Hiroki NOMOTO

Tokyo University of Foreign Studies
asako@aa.tufs.ac.jp, sakon.yuta.n0@tufs.ac.jp, nomoto@tufs.ac.jp

This study examines the function of non-active voices observed in recent written Indonesian texts included in the web corpus in MALINDO Conc, with special focus on their sub-types with third person agent pronouns, for which the distribution of all the voice types may overlap. We investigated the environment where the six most frequent transitive verb stems occur in non-actor voices. Almost all the clauses with the stem *miliki* 'possess' occur in non-active voices when non-active voice is syntactically required, that is, where the patient is a syntactic pivot. The other verb stems may do so to varying degrees when non-active voices are not syntactically required: The stem *lakukan* 'do' often takes a demonstrative subject that refers to a proposition introduced in the preceding discourse, about which some circumstantial information is added by the post predicate adverbial clauses, indicating reasons, purposes, or temporal relations. The stems *lihat* 'see', *katakan* 'say', and *buat* 'make' occur very frequently with a complement clause, expressing a perceived event, speech, and caused event, respectively.

1. Introduction

The aim of this research is to investigate the function of non-active voices in recent written texts in the MALINDO Conc Web Corpus. According to Arka and Manning (1998) and Cole et al. (2008), among others, standard Indonesian has three voice constructions, one active voice and two non-active voices, that is, passive voice and objective voice. Sentences (1), (2), and (3) are examples of active voice, passive voice, and objective voice[1], respectively.

(1) *Dia* *belum* *mem-baca* *buku* *sejarah* *ini*.
3 not.yet ACT-read book history this

'He has not read this history book.'

(2) *Buku* *sejarah* *ini* *belum* *di-baca*.
book history this not.yet PASS-read

'(He) has not read this history book.'

(3) *Buku* *sejarah* *ini* *belum* *dia* *baca.*
book history this not.yet 3 read

'(He) has not read this history book.'

[1] Various labels have been given to the objective voice. For example, Dardjowidjojo (1978) and Sneddon et al. (2010: 255ff) use the label 'passive type 2' for this construction, while, Cole et al. (2006) use the traditional label *pasif semu* 'pseudo-passive'.

SHIOHARA, Asako, Yuta SAKON and Hiroki NOMOTO, 2019. 'Discourse functions of the two non-active voices in Indonesian: Based on the web corpus data in MALINDO Conc'. In Hiroki Nomoto and David Moeljadi, eds. *Linguistic studies using large annotated corpora*, NUSA 67, 77–101. Permanent URL: http://repository.tufs.ac.jp/handle/10108/94453] [doi: 10.15026/94453]

In active voice, the verb appears with a derived form with the prefix *meN-*, while in passive voice the verb appears with the prefix *di-*. In objective voice, the verb appears in the bare stem form and the agent pronoun obligatorily occurs immediately preceding the verb. In active voice the agent NP is a grammatical subject, while in the two other non-active voice constructions the patient NP is a grammatical subject (Sneddon et al. 2010).

In passive clauses, the agent is often left unexpressed, as in (2), or expressed by an NP occurring immediately after the verb, or as a PP introduced by the preposition *oleh*, as in (4) and (5). Also, it may be expressed by the third person pronominal clitic *=nya*, as shown in (6).

(4) *Buku sejarah ini belum di-baca (oleh) guru itu.*
book history this not.yet PASS-read by teacher that
'The teacher has not read this history book.'

(5) *Buku sejarah ini belum di-baca (oleh) dia.*
book history this not.yet PASS-read by 3
'He has not read this history book.'

(6) *Buku sejarah ini belum di-baca=nya*[2].
book history this not.yet PASS-read=3
'He has not read this history book.' or 'They have not read this history book.'

The two non-active voice types exhibit a nearly complementary distribution; passive voice is canonically used only when the agent is third person, while objective voice is used only when the agent is expressed by a pronoun or pronoun substitutes such as kinship terms (Sneddon et al. 2010: 255). When the agent is expressed by a third-person pronoun, however, both the voice types can be used. We will discuss this type of construction through a syntactic and discursive comparison of the two non-active clauses in Section 4. Table 1 summarises the distribution of voice types mentioned above; the target of our research is shown by shading.

Table 1. Canonical distribution of the three voices

	Agent forms		
	First- and second-person pronouns and pronoun substitutes	Third-person pronouns	Third-person lexical nouns
meN- active	+	+	+
objective voice	+	+	−
di- passive	−	+	+

Table 2 shows the constructions we investigate in the following sections.

[2] The enclitic *=nya* may refer to either singular or plural referents (Sneddon et al. 2010: 171).

Table 2. Target clause types in this study[3]

Subtype		Marking	Position of agent	Example sentences
Objective voice		Ø	pre-adjacent to V	*Tini akan dia jemput.* 'He will meet Tini.' *Tini akan mereka jemput.* 'They will meet Tini.'
Di- passive	DP type	*di-*	post-adjacent to V	*Tini di-jemput=nya.* 'He/she/they meet(s) Tini.' *Tini di-jemput mereka.* 'They meet Tini.'
	Oleh type	*di-*	introduced by *oleh* 'by'	*Tini di-jemput oleh=nya.* 'He/she/they meet(s) Tini.' *Tini di-jemput oleh mereka.* 'They meet Tini.'

There is a wide consensus on the discourse property of the two non-active voice types that the referent of the subject, that is, the patient, expresses a topic of the clause (Sneddon et al 2010: 255). The precise discourse functions of the two non-active voice types, however, have not been fully discussed. First, we will investigate whether there is any functional difference between the two non-active voice types which semantically express the same proposition as exemplified in Table 2.

The constructions in Table 2 exhibit a discourse function which is distinct from cross-linguistic passive voices in that both the agent and patient exhibit high topicality. The passive voice is often characterised as indicating non-topical agent status (cf. Givón 1994: 9) or the defocusing of the agent (Shibatani 1984: 832), but the construction above the agent is explicitly marked by a pronoun, indicating its high topicality. Thus, the second question arises: in which pragmatic environment do the two 'topical' NPs co-exist in these voice types?

Another issue to be considered is the status of the so-called 'event focus' function of *di-* clauses. As discussed in Hopper (1984) and Kaswanti Purwo (1988), among others, *di-* passive clauses, especially when they take VS order, may be used irrespective of the topicality of the agent or patient in narrative as an indicator of 'sequenced events which pertain to the main line of the discourse' (Hopper 1984: 84). This function, however, has become less common in written Indonesian since the 1980s, as suggested in Cumming (1984), who discusses the functional changes of several Indonesian syntactic

[3] The classification of Indonesian *di-* passive voices and the terms for their subtypes are based on Nomoto and Kartini (2014), who examined the frequencies of different passive subtypes in relative clauses using Malaysian Malay corpora. See Nomoto (2015, 2018) for discussions of the syntax of different subtypes and relations between them.

structures, including this type of *di-* passive clause[4]. Thus the question also arises as to whether this functional change has progressed in contemporary Indonesian.

Data in the MALINDO Conc web corpus would enable us to address the questions above based on its substantial amount of naturalistic written data. A technical problem, however, arises in identifying objective voice clauses in MALINDO Conc data. As suggested in Cole et al. (2006) and Sneddon (2006: 43ff), among others, in current colloquial Indonesian, clauses with a bare transitive stem may appear in a voice type similar to active voice in standard Indonesian. Consider sentences (7a) and (7b) given in Cole et al (2006: 43); (7b) shows a bare-active voice syntactically similar to (7a), the canonical active sentence in standard Indonesian[5].

(7) a. *Siti* *menulis* *surat* *itu*
Siti ACT-write letter that

'Siti write the letter.'

b. *Siti* *tulis* *surat* *itu*
Siti write letter that

'Siti write the letter.'

Because of the presence of the bare-active voice, the voice type of a bare stem clause may be indeterminate. When the patient NP occurs after the predicate, as in (7b) above, the clause can be considered to be the bare active voice, whereas when auxiliaries occur before the pronoun and verb, as in (8) (= (3)) below, the clause is considered to be in objective voice. In other cases, however, the status of the clauses is indeterminate.

(8) *Buku* *sejarah* *ini* *belum* *dia* *baca.*
book history this not.yet 3 read

'He has not read this history book.'

Sentences (9) and (10) are examples of such indeterminate clauses from the MALINDO Conc corpus.

(9) *Buku-buku* *tersebut* ***ia*** ***baca*** *satu* *per* *satu*.
books above-mentioned 3 read one by one

'He read the books one by one.'

http://arimateapusat.wordpress.com/

[4]For details of the 'event focus' function, see Kroeger (2016: 21ff), who gives a survey of this issue.

[5] Although Cole et al. (2006) call the construction bare active, the syntactic behaviour does not perfectly overlap with active voice in standard Indonesian in that it permits object relativisation, as Cole et al. (2006) suggest in the same study.

(10) *Karakter* *individu* *di-dasarkan* *pada* *orang-orang* *yang*
character individual PASS-base to people REL

di-kenal *oleh* *Bramantyo* *atau* *yang* ***ia*** ***baca***
PASS-know by Bramantyo or REL 3 read

tentang *orang* *tersebut.*
about person above-mentioned

'Individual characters are based on people who are known by Bramantyo or who he read about that person.'

http://id.wikipedia.org/wiki/?_(film)

In standard Indonesian, only the subject argument can be relativised, while, as discussed in Cole et al. (2006) and Nasanius et al. (2016), bare active voice clauses may relativise either the agent NP or patient NP. Thus, the voice of the relative clause is indeterminate if no auxiliary precedes the agent pronoun and verb, as in (10).

In this study, we include the indeterminant clauses exemplified above in our investigation as potential objective voice clauses. The bare active voice is a non-canonical construction observed in colloquial varieties, such as Jakarta colloquial Indonesian. The MALINDO Conc corpus we employ for our study is a web-corpus that comprises of written specimens. We thus inferred that the frequency of bare active clauses is not very high, which would justify the decision.

In Section 2, we outline the nature of the MALINDO Conc corpus data. In Section 3, we examine the frequency of each voice subtype shown in Table 2 in MALINDO Conc. In Section 4, in order to determine the precise environment in which non-active type clauses occur, we focus on the six most frequent transitive stems, namely *miliki* 'possess, own', *lakukan* 'conduct, do', *buat* 'make', *lihat* 'see', *gunakan* 'use', and *katakan* 'say', and examine the discourse function of the two non-active voices.

2. The frequencies of transitive *meN-* verbs and *di-* verbs

2.1 The size of the corpus and the rough frequency of *meN-* verbs and *di-* verbs

Before examining the discourse function of each voice type, we introduce the nature of the corpus data we employ and show the frequencies of *di-* passive sentences and *meN-* active sentences in the corpus in this section.

We used the following three Indonesian subcorpora of the Leipzig Corpora Collection (Goldhahn, Eckhart & Quastohoff 2012) that are included in the MALINDO Conc online concordance system (Nomoto, Akasegawa & Shiohara 2018a)[6] in April 2019: ind_mixed_2012, ind_web_2012, and ind_wikipedia_2016. These subcorpora are the results of reclassification of the original Malay and Indonesian data into standard Malay (ISO693-3 zsm) and standard Indonesian (ISO693-3 ind) by Nomoto, Akasegawa and Shiohara (2018b). They are thus more reliable in terms of language identification than the original version, though language identification errors still remain. Table 3 summarises the data and sizes of the three subcorpora.

[6] https://malindo.aa-ken.jp/conc/

Table 3. The data and sizes of the three Indonesian subcorpora

Subcorpus		ind_mixed_2012	ind_web_2012	ind_wikipedia_2016
Data		mixed (e.g. blog)	mostly online news	Wikipedia articles
Size	(sentence)	300,000	300,000	300,000
	(token)	5,428,067	5,540,573	5,634,138

The corpora in MALINDO Conc, including the three Indonesian subcorpora used for the present study, are all annotated morphologically. Specifically, each word contains information about the root, affixes (prefixes/proclitics, suffixes/enclitics, circumfixes), and reduplication type (full, partial, rhythmic). The annotated corpora are in XML format. An excerpt from the ind_mixed_2012 subcorpus is shown in Figure 1. The attributes in the `w`(ord) elements specify the morphology information of the word in question: `rt` for root, `p1` and `p2` for prefixes/proclitics[7], `s1` and `s2` for suffixes/enclitics, `c1` and `c2` for circumfixes[8], and `r` for reduplication type. The `pu` tag is used for punctuation symbols. The morphological analysis is based on version 20181125 of the morphological dictionary MALINDO Morph (Nomoto et al. 2018)[9]. For example, the word (`w`) *sedikit* in line 5 consists of the root (`r`) *dikit* and the prefix (`p1`) *se-*.

It must be noted that the morphological annotation in the corpus is not free from errors. The following errors can affect the results of the present study. The first error is related to the words that are morphologically ambiguous. Most of them have been disambiguated manually, except for extremely frequent words such as *mereka*, for which we intended to give the most frequent morphological analysis (e.g. *mereka* as opposed to *meN-* + *reka*) as default. In the process of annotation, however, the string `Mereka` with a capital *M* was always analysed as *meN-* + *reka* 'to devise' (transitive verb), but not as *mereka* 'they' (pronoun), which is monomorphemic; the string `mereka` with a lowercase *m*, on the other hand, is always analysed as the latter. The string `Mereka` is extremely frequent in the corpus (9,432 instances) and almost all of its occurrences are as the pronoun *mereka*. The second type of annotation errors results from errors in MALINDO Morph. Since no principled way exists to catch such errors exhaustively, we decided to ignore these types of errors in this study.

[7] Prepositions wrongly combined with the following word are also included here. For instance, the *di* in *dikampus* in Figure 1 is not the prefix *di-* but the preposition *di*.

[8] The numbers 1 and 2 after 'p', 's', and 'c' refer to the first and second occurrences.

[9] https://github.com/matbahasa/MALINDO_Morph

```
1   <doc url="http://0502198800.blogspot.com/"
       subcorpus="ind_mixed_2012.4.300K">
2      <s id="ind_mixed_2012.4.300K.0000088"
          subcorpus="ind_mixed_2012.4.300K" date="2012-05-08"
          url="http://0502198800.blogspot.com/">
3          <w rt="itu">Itu</w>
4          <w rt="hanya">hanya</w>
5          <w rt="dikit" p1="se-">sedikit</w>
6          <w rt="contoh">contoh</w>
7          <w rt="dari">dari</w>
8          <w rt="kian" p1="se-">sekian</w>
9          <w rt="banyak">banyak</w>
10         <w rt="hal" r="R-penuh">hal-hal</w>
11         <w rt="buruk">buruk</w>
12         <w rt="yang">yang</w>
13         <w rt="ada">ada</w>
14         <w rt="kampus" p1="di">dikampus</w>
15         <w rt="kena" c1="ber--an">berkenaan</w>
16         <w rt="tentang">tentang</w>
17         <w rt="tingkah">tingkah</w>
18         <w rt="laku">laku</w>
19         <w rt="mereka">mereka</w>
20         <w rt="yang">yang</w>
21         <w rt="jauh">jauh</w>
22         <w rt="tidak">tidak</w>
23         <w rt="cermin" p1="meN-" s1="-kan">mencerminkan</w>
24         <w rt="ciri">ciri</w>
25         <w rt="orang">orang</w>
26         <w rt="didik" p1="ber-" c1="peN--an">berpendidikan
              </w>
27         <pu>.</pu>
28      </s>
29  </doc>
```

Figure 1. Morphological annotation in the XML format

Morphological annotation enables one to search a corpus with queries in terms of morphological properties of the target words such as 'all words that contain the prefixes *meN-* and *ter-*' [10]. Morphological searches can be conducted using the online concordancer MALINDO Conc. However, searches involving words whose frequencies exceed 100,000 per subcorpus are impossible with MALINDO Conc because the maximum number of result lines for a search is set at 100,000. Because some of the target words of this study occur highly frequently, we used the actual corpus files rather than the online interface provided by MALINDO Conc[11] and wrote Python scripts to process and to search the corpus files.

[10] Sixty-three such tokens were found in the Indonesian subcorpora available in MALINDO Conc: *menertawakan* (41), *mentertawakan* (6), *menertawai* (5), *mentertawai* (2) 'to laugh at' (2), and *mentertawakannya* 'to laugh at him/her' (7).

[11] Readers interested in the relevant files can contact the last author of this paper.

2.2 Methodology for selecting target clauses

We counted the numbers of transitive *meN-* verbs and *di-* verbs as a way of counting the occurrences of `"meN-"` and `"di-"` as `p1` or `p2` as follows[12]. It is straightforward to obtain the number of *di-* verbs in the corpus by simply counting the number of words that contain the prefix *di-*. In the XML format in Figure 1, such words are `w` elements that have either `p1="di-"` or `p2="di-"` as their attributes.

It is trickier to search for transitive *meN-* verbs. Not all `w` elements that have either `p1="meN-"` or `p2="meN-"` as their attributes are exclusively transitive; some are intransitive and others can be either transitive or intransitive. Given the large size of our corpus, we decided to regard those that can be either transitive or intransitive simply as transitive, and just exclude those that are exclusively intransitive. In order to distinguish the transitive verbs from the (exclusively) intransitive verbs, we identified the former by choosing those that belong to the categories Type 1 and Type 2 below, which are, with high probability, transitive.

- Type 1: *meN-* verbs that have corresponding *di-* passive forms
- Type 2: *meN-* verbs containing the transitive suffix *-kan* or *-i* and/or a pronominal enclitic (e.g., *=ku* '1SG'), which serves as the object of a *meN-* verb

No comparable diagnostics exist for intransitive verbs and annotation errors. Hence, the best we can do is to identify *meN-* verbs belonging to the above two types from the set of all *meN-* verbs.

To identify *meN-* verbs of the two types above, we used the MALINDO Morph morphological dictionary. A MALINDO Morph line (version 20190129) consists of the following eight elements: ID, root, surface form, prefix/proclitic, suffix/enclitic, circumfix, reduplication type, and source. An example line is given in (11).

(11) `ex-11302 besar diperbesarnya di-+per- -nya 0 0 Leipzig`

To obtain Type 1 words, one can first create a set of *meN-* verbs by replacing the prefix *di-* of *di-* verbs. For example, the *meN-* active counterpart of *diperbesarnya* 'to be enlarged by him/her/it' is *memperbesarnya* 'to enlarge him/her/it', which can be obtained by searching MALINDO Morph for a line such that the root is *besar*, the prefix/proclitic is *meN-+per-*, the suffix/enclitic is *=nya*, and the circumfix and reduplication type are both 0. This finds 5,090 such *meN-* verb types. For Type 2, one can search for lines such that the prefix/proclitic contains `meN-` and the suffix/enclitic contains the relevant suffixes and enclitics; 5,036 such *meN-* verb types were found. Among the Type 2 verbs, some that occur with the prefix *meN-* and suffix *-kan* may be adjectives. We excluded such cases manually and then obtained the set of clearly transitive verbs.

The study obtained 636 *meN-* verbs, whose transitivity is unclear and are subject to annotation errors, by subtracting the clearly transitive verbs mentioned above. We manually coded these 636 verbs as (i) transitive, (ii) transitive/intransitive, (iii)

[12] The prefixes *meN-* and *di-* are expected to occur in the `p1` position by default. However, we also count the occurrences in the `p2` position, as it may occur as `p2`, as in *meN-* in the words such as *dimengerti* (`rt="erti" p1="di-" p2="meN-"`) or *memeratakan* (`rt="rata" p1="meN-" p2="meN-"`), though the result of the counting shows that such cases are infrequent in number and, therefore, ignorable.

intransitive, or (iv) other (= errors in annotation). 414 *meN-* verbs were classified into categories (iii) and (iv) and are thus irrelevant for our study.

Finally, we counted the number of the `w` elements that have either `p1="meN-"` or `p2="meN-"` as their attributes apart from these 414 words. We also excluded `Mereka` 'they', `dimengerti` 'to be understood', `belajar-mengajar` 'study-teach', and `kemelaratan` 'misery, poverty'. The first of these is a known annotation error while the rest were caught by examining the list of words with `p2="meN-"`, that is, words in which another affix occurs outside the prefix *meN-*.

Table 4 summarises the frequencies of the transitive *meN-* verbs and *di-* verbs occurring in the three Indonesian subcorpora.

Table 4. Frequencies of transitive *meN-* verbs and *di-* verbs

Subcorpus	*meN-* verbs		*di-* verbs	
	token	type	token	type
ind_mixed_2012	296,766	5,694	123,575	4,062
ind_web_2012	302,247	5,727	127,897	3,946
ind_wikipedia_2016	279,215	4,261	138,613	3,449

3. Voice subtypes

3.1 Methodology

As mentioned in the Introduction, our primary focus is the non-active voice clauses with the third person pronouns, that is, the subtypes summarised in Table 2 in the Introduction, which will be repeated as Table 5. In this section, however, we show the frequencies of all the voice subtypes with the agent pronouns included in the MALINDO Conc corpus in order to capture the size of the clause types that will be the focus of Section 4.

We restrict our target active clauses to those where the agent is pre-adjacent to the verb for ease of search, though the agent can occur at other pre-verbal positions too. Thus, the actual number of target active clauses may be larger than the figure given in Tables 6 and 7. This applies to the number of *meN-* active clauses with the third person pronouns shown in Tables 8 and 9 in Section 4.

We applied the following methodology to extract the target clauses from the MALINDO Conc corpus. First of all, we determined a set of common stems based on which the different voice subtypes in Table 5 are formed. The common stems were chosen from the set of all *di-* verbs in the corpus that also exist in MALINDO Morph by removing the prefix *di-* and enclitics (e.g. *=nya* 'him/her/it'), if any (see section 2.2 on how to obtain *di-* verbs from the annotated corpus). Thus, 3,944 target stems were obtained[13].

[13] In fact, these target stems also contain such words as *kota* 'city' and *keesokan* 'next day'. This happened because the unchecked part of MALINDO Morph, which was used to annotate the corpus, contains errors. For example, *dikota* is analysed as the prefix *di-* + *kota*, when it is in fact the result of a common mistake of writing the preposition *di* 'at' together with the following noun without an

Table 5. Voice subtypes examined in this section

Subtype		Voice marking on V	Position of Agent
Bare clause		Ø	pre-adjacent to V
Di- passive	DP type	*di-*	post-adjacent to V
	Oleh type	*di-*	introduced by *oleh* 'by'
MeN- active		*meN-*	pre-verbal

Next, we created target *di-* and *meN-* verbs as follows. First, we edited the morphological information of a stem to make it to a hypothetical *di-* or *meN-* form. Note that these forms were mechanically created, and there is no guarantee that a given created form actually exists. The relevant form is thus hypothetical at this point.

(12) Format: (root, prefix/proclitic, suffix/enclitic, circumfix, reduplication type)

a. Stem (*edarkan* 'to distribute')
`('edar', '0', '-kan', '0', '0')`

b. Hypothetical *di-* form (substitution of `di-` for `0`)
`('edar', 'di-', '-kan', '0', '0')`

c. Hypothetical *meN-* form (substitution of `meN-` for `0`)
`('edar', 'meN-', '-kan', '0', '0')`

Next, we searched MALINDO Morph for a surface form matching the morphological information in question. For (12b) and (12c) above, MALINDO Morph contains the following lines:

(13) Format: ID root surface form prefix/proclitic suffix/enclitic circumfix reduplication type source

a. `ex-22166 edar diedarkan di- -kan 0 0 Leipzig`

b. `cc-21031 edar mengedarkan meN- -kan 0 0 Kamus`

Consequently, we add *diedarkan* 'to be distributed' and *mengedarkan* 'to distribute' to our target *di-* and *meN-* verbs. Table 6 summarises the numbers of various target forms. Note that the number of target *di-* verbs is expected to be larger than that of target *meN-* verbs because the target stems were determined from the set of *di-* verbs. The 'no enclitic' category of *di-* verbs includes two *di-* passive subtypes, namely, (i) the DP type and *oleh* type with a non-enclitic agent NP (e.g. *diambil* (*oleh*) *mereka* 'to be taken by them') and (ii) the *pro* type, in which no overt agent expression occurs (e.g. *diambil* 'to be taken').

intervening space. We are not able to exclude such forms (i.e. the combination of the preposition *di* and noun indicating the location) in this study. Similarly, there may be some cases in which the *di-* prefix is written separately from the stem (e.g. *di milikinya*), which are not included in the data.

Table 6. The numbers of target forms

Form	Stems	*Di-* forms				*MeN-* forms
		No enclitic	*=ku* (1SG)	*=mu* (2SG)	*=nya* (3)	
Number	3,944	3,801	7	5	1,774	4,164
Total	3,944	5,587				4,164

3.2 Results

Table 7 shows the results. Among the pronouns, the forms with extremely low frequency, such as *engkau* '2SG', *beliau* '3' are excluded from the data shown in Table 7. Kinship terms such as *Bapak* 'father' and *Ibu* 'mother' may refer to the first or second person, but such cases are not included, as it is impossible to automatically distinguish those nouns with first- or second-person referents from those with literal meanings.

Table 7. Frequencies of agent pronouns in each voice type clauses

	meN- active clauses	bare clauses	*di-* passive clauses without *oleh* 'by'	*di-* passive clauses with *oleh* 'by'
saya '1SG'	3,994	8,087	51	7
aku '1SG'	1,914	2,563	20	1
=ku '1SG'	-	-	2	3
Anda '2SG'	2,335	3,464	38	9
kamu '2SG'	613	918	5	0
=mu '2SG'	-	-	2	2
dia '3'	6,493	3,072	66	15
ia '3'	16,051	6,441	100	1
=nya '3'	-	-	10,786	143
kita '1PL (INCL)'	4,496	9,643	81	18
kami '1PL (EXCL)'	1,617	2,550	15	6
kalian '2PL'	413	549	3	2
mereka '3PL'	5,996	6,124	231	142
Total	832,426	1,337,661	305,936	43,935

The results show the following points:

(a) In bare construction and *meN-* active clauses, we cannot see any clear difference in frequency among the three persons. On the contrary, the first- and second-person pronouns rarely occur in passive clauses in general, though they do occur in a small number of clauses. This confirms the observation of Sneddon et al. (2010: 258) that the use of *di-* passive clauses with first- or second-person agent is not canonical in standard

Indonesian, but may be used (A similar situation is reported in Malaysian Malay by Nomoto and Kartini 2014).

(b) The occurrence of pronouns in *oleh* phrases of *di-* passive clauses are generally very rare, though the third-person clitic *=nya* and the third-person plural *mereka* occur more frequently than the other pronouns, exhibiting a considerable number of occurrences. This shows that *oleh* phrases have a tendency not to be used to introduce the salient entity in discourse.

4. Discourse functions of non-active clauses with the six most frequent stems

4.1 Determining the target clauses

In this section, we observe the voice type distribution of the six most frequent transitive stems with the third person pronoun, namely, *miliki* 'possess', *lakukan* 'do', *buat* 'make', *lihat* 'look', *gunakan* 'use', and *katakan* 'say'. The six stems are selected from the 11 most frequent stems occurring in the *meN-* form shown in Table 8. The remaining five stems are excluded, namely, *jadi* 'become', *rupakan* 'constitute', *turut* 'follow', *lalui* 'go through', and *reka* 'invent', as the corresponding passive forms with the third person pronoun do not occur or only very rarely occur in the corpus[14].

Table 9 shows the frequencies of each of these verb forms with the third person pronoun, including its enclitic form *=nya*. As mentioned above, it is impossible to extract all of the *meN-* active clauses with the third person pronoun using only the concordancer provided by MALINDO Conc. Here, we restrict our target to those in which the agent is pre-adjacent to the verb, though the agent can occur at other pre-verbal positions as well. Thus, the actual frequencies of the *meN-* active clauses are expected to be higher than shown in Table 9. Except for the verbs *lakukan* 'do', the occurrence number of the *meN-* active clauses is higher than that of the clauses of other types.

[14] The form *di-jadi* (No. 1) is not attested as a word. It is included in the data wrongly, as a part of the sequence of *dijadi-kan*, where the hyphen '-' is inserted before a new line in example (i) below. Substantial amount of occurrence for the form meN+reka (mereka) are considered to be wrongly included because of the annotation error mentioned in the previous section, that is, the sequence mereka are always analysed as meN+reka.

(i) *Ak- hirnya, Ratu Pembayun dijadikan triman (dikeluarkan dari Keraton, diberikan kepada seseorang untuk diasuh atau dijadi- kan istri), dan diserahkan kepada Ki Ageng Karanglo*.

http://adbmcadangan.wordpress.com/wisata-adbm/2/index0334.html

Table 8. Frequencies of *meN-* active clauses

	Stem	*MeN-* (all)	Bare (all)	*Di-* (all)
1	*jadi*	16,029	2,846	1
2	*rupakan*	8,321	2	1
3	***miliki***	**7,395**	**272**	**613**
4	***lakukan***	**5,956**	**1,379**	**5,034**
5	***buat***	**5,202**	**1,138**	**1,112**
6	*turut*	4,222	353	1
7	***lihat***	**3,970**	**956**	**1,039**
8	*lalui*	3,953	39	97
9	***gunakan***	**3,695**	**484**	**2,924**
10	*reka*	3,107	4	1
11	***katakan***	**2,906**	**377**	**1,061**

Table 9. Frequencies of each clause type

	Third-person pronoun +*meN-* (rough estimate)	Third-person pronoun + bare form	*di-*V + third-person pronoun
miliki	1,240	200	405
lakukan	515	673	543
buat	413	115	156
lihat	832	98	271
gunakan	295	165	94
katakan	593	88	145

Table 10 shows the frequencies of the bare stem clauses with each verb stem. As mentioned in the Introduction, bare active clauses followed by the patient free NP are excluded from consideration in the remainder of this paper. The form *=nya* may be attached to the *di-* prefixed verb to derive a nominalised verb (Grange 2008). We exclude such nominalised *di-V=nya* forms from the present consideration[15].

[15]Sentence (i) below is one of the examples of *di-V=nya* nominalisation obtained from the MALINDO Conc corpus.

(i) *Satu-satunya syarat yang diperlukan untuk menempuh raja yoga ini adalah dimilikinya suatu dugaan kuat bahwa diri manusia sebenarnya jauh lebih mengagumkan dari yang kita sadari saat ini*. 'The only condition needed to take this yoga king is to have a strong suspicion that the human self is actually far more amazing than we realise today.'

(http://id.wikipedia.org/wiki/Atman, IND WKP2016 subcorpus of the Leipzig Corpora Collection (Goldhahn, Eckart & Uwe Quasthoff 2012) using MALINDO Conc (Nomoto, Akasegawa & Shiohara 2018).

Table 10. Frequencies of bare active clauses and target clauses

	miliki	*lakukan*	*buat*	*lihat*	*gunakan*	*katakan*
Total	200	673	115	98	165	88
Bare active voice	0	18	9	3	8	2
Target clauses	200	655	106	95	157	86

The target clauses shown in Table 10 can be grouped into (i) clauses that clearly have objective voice, or (ii) clauses that are indeterminate as to whether they have bare active voice or objective voice. Two of the verbs, *lihat* 'see' and *katakan* 'say', may take a complement clause indicating a perceived situation or the utterance content. As the syntactic status of such complement clauses in the matrix clauses is not clear at the present stage of research, we simply place clauses with a complement clause in an independent category.

Table 11. Syntactic properties of the target bare stem clauses

	miliki	*lakukan*	*buat*	*lihat*	*gunakan*	*katakan*
Objective voice	12	199	18	13	20	12
Indeterminate	188	456	88	79	137	51
With a complement clause	0	0	0	3	0	23
Total	200	655	106	95	157	86

Table 12 shows the frequencies of the *di-* passive verbs with the third-person enclitic *=nya* and the third-person plural pronoun *mereka*.

Table 12. Frequencies of *di-* passive clauses with the third-person pronouns

	miliki	*lakukan*	*buat*	*lihat*	*gunakan*	*katakan*
di-V=nya	401	526	152	269	85	141
di-V mereka	4	17	4	2	9	4
Target clauses	405	543	156	271	94	145

In the remaining part of this section, we will investigate the environment in which the two non-active voice types occur.

4.2 Syntactically required use of non-active clauses

In some subordinate clauses, voice types are syntactically determined. A great majority of such cases are relative clauses introduced by the relativiser *yang*. The other types of subordinate clauses include purpose clauses introduced by the conjunctions *untuk* 'in order to' and *supaya* 'in order to' and the manner clause *sebagaimana* 'in such a way', among others. The voice type of those clauses is determined by the semantic relation between the antecedents and predicate (Sneddon et al. 2010: 302, 310). Therefore, when the antecedents have a semantic role of patient, the non-active voices are required. Sentences (14) and (15) are examples of relative clauses and manner clauses observed in the MALINDO Conc corpus.

(14) *Bahasa **yang di-gunakan mereka** adalah bahasa Arab.*

'The language **which they used** was Arabic'

https://id.wikipedia.org/wiki/Alawi

(15) *Seperti yang dikatakannya, maka ia akan melepaskan diri dengan cara **sebagaimana di-katakan=nya** itu.*

'As he said, then he will break away **in the way he said** it.'

http://adbmcadangan.wordpress.com/buku-III-79/

Tables 13 and 14 show the frequency of the syntactically required objective voice clauses and *di-* passive clauses with a third-person pronoun (=*nya* '3' and *mereka* '3PL'), respectively.

Table 13 shows the high frequency of the syntactically required objective voice for all six stems, among which the stem *miliki* exhibits a much higher frequency than the other stems.

Table 13. Frequencies of the syntactically required bare stem clauses

	miliki	*lakukan*	*buat*	*lihat*	*gunakan*	*katakan*
Syntactically required objective voice clauses	198 (99.0%)	552 (84.3%)	98 (92.4%)	87 (91.6%)	118 (75.2%)	59 (68.6%)
Not syntactically required objective voice clauses	2 (1.0%)	103 (15.7%)	8 (7.6%)	8 (8.4%)	39 (24.8%)	27 (31.4%)
Total	200	655	106	95	157	86

As shown in Table 14, the frequency of the syntactically required *di-* passive clauses varies across stems. The verb stem *miliki* exhibits a similar tendency as the bare stem clauses; the frequencies of the syntactically conditioned clauses are extremely high. This, at least partially, may be explained by the semantics of the stem that expresses the state of possession. The remaining stems can be grouped into two categories. The stems *lakukan* 'do' and *gunakan* 'use' exhibit a relatively high frequency of syntactically conditioned *di-* passive clauses, while the stems *buat* 'make', *katakan* 'say', and *lihat* 'look' exhibit relatively low frequencies of syntactically conditioned *di-* passive clauses. This difference may be explained by the difference in the structure that predicates with these stems may take, that is, they take a complement clause. We will examine this point in Section 4.3.

Table 14. Frequencies of the syntactically required *di-* passive clauses

	miliki	*lakukan*	*buat*	*lihat*	*gunakan*	*katakan*
Syntactically conditioned *di-* passive clauses	402 (99.3%)	447 (82.3%)	97 (63.0%)	113 (41.7%)	80 (85.1%)	75 (51.7%)
Not syntactically conditioned *di-* passive clauses	3 (0.7%)	96 (17.7%)	57 (37.0%)	158 (58.3%)	14 (14.9%)	70 (48.3%)
Total	405	543	154	271	94	145

4.3 Environments in which not syntactically conditioned clauses occur

In this section, we will examine the environment in which non-active clauses that are not syntactically required occur. Table 15 shows the syntactic structures of bare stem clauses. In a majority of the clauses, the patient argument occurs in the pre-predicate position. With the stem of the utterance verb *katakan* 'say', a structure with a complement clause also occurs frequently. The stem *lihat* 'see' also takes a similar structure, though the occurrence number of this stem in total is quite small.

Table 15. Syntactic structures of the pragmatically motivated bare stem clauses

Verb stems	*miliki*	*lakukan*	*buat*	*lihat*	*gunakan*	*katakan*	Total
P in the pre-predicate position	2	90	8	2	27	4	133
P left unexpressed	0	13	0	3	12	0	28
Complement clause in post-predicate position	0	0	0	3	0	23	26
Total	2	103	8	8	39	27	187

Table 16 shows the syntactic structure of the non-syntactically motivated *di-* passive clauses. The stems *lakukan* 'do', *miliki* 'have', and *gunakan* 'use' often take the patient argument in the pre-predicate position, as with bare stem clauses. Contrastively, in clauses with the stems *katakan* 'say', *lihat* 'see', and *buat* 'make', a structure with a complement clause occurs frequently; the stem *buat* is used as a causative verb in this environment.

Sentences (16)–(20) are examples of the use of *di-* passive clauses with a complement clause with the verbs *katakan* 'say', *lihat* 'see', and *buat* 'make'.

With the verb stems *katakan* 'say' and *lihat* 'see', the complement clause always occurs after the predicate, as in (16) and (17). Sentences (16)–(19) deviate from a cross-linguistically typical passive clause, as they do not occur with the highly topical patient argument.

Table 16. Syntactic structures of the pragmatically motivated *di-* passive clauses

	miliki	*lakukan*	*buat*	*lihat*	*gunakan*	*katakan*	Total
P in the pre-predicate position	3	89	11	9	8	9	129
P in the post-predicate position	0	4	4	41	2	1	52
P left unexpressed	0	3	1	3	4	1	11
With a complement clause	0	0	41	105	0	59	205
Total	3	96	57	158	14	70	398

(16) *Di-katakan=nya* ***hal ini merupakan kesempatan Indonesia untuk membagi pengalaman kehidupan pluralisme indonesia dengan spanyol***.

'He says **that this is a chance for Indonesia to share the experience of pluralism, which has taken root in Indonesia, with Spain**.'

http://antarajendeladunia.blogspot.com/2010/12/dialog-lintas-agama.html?m=1

(17) *Di-lihat=nya* ***beberapa orang anak muda berdiri di sudut desa***.

'He saw **some children standing on the edge of the village**.'

https://adbmcadangan.wordpress.com/buku-84/6/

The verb stem *buat* exhibits various types of structure; the complement clause may precede, as in (18) and (19), or follow, the predicate, as in (20)[16].

(18) *Waaah senang sekali mendengarnya dan terharu saya* ***di-buat=nya***.

'Wow, (I am) very happy to hear that and **that made** me touched.'

https://akubisadengar.wordpress.com/category/komunitas-tuna-rungu/

(19) *Jika ada hal yang buruk kita benar-benar terhenyak* ***di-buat=nya***.

'If there is a bad thing, **that makes** us really stunned.'

http://hazirahmohdyusof.blogspot.com/2011/04/pemenang-kehidupan.html

(20) *Semua Penonton* ***di-buat=nya*** *terlena dan tetap duduk dan menyaksikan acara ini sampai akhir acara*.

'**That made** all the audience attracted, remain seated, and watch the program until the end of the program.'

http://anaksastra.blogspot.com/2009/01/menggugah-pelestarian-lingkungan.html

[16] Sentence (20) takes a pre-predicate NP, as well as a post-predicate complement clause. In the classification of the structures in Table 16, we put sentences of this type into the category with the patient argument in the pre-predicate position.

A rough observation of the MALINDO Conc data tells us that the active voice of the verb stems *katakan* 'say' and *lihat* 'see' and *buat* also occur with a complement clause to a great extent. The factors that motivate the voice selection of those stems are not clear at this stage of our research.

4.4 Pragmatic status of pre-predicate patient argument

As mentioned in the Introduction, non-active clauses with the pronominal agent marking by default have two topic expressions: the pronominal agent and a pre-predicate NP indicating the patient. To understand the functions in discourse that the two topics play in greater detail, we examined the discourse properties of the patient argument. The results are summarised in Table 17; only a small number of cases of pre-predicate NP have a co-referential NP in the preceding discourse (category (a) in the table). They more often (b) are anchored by an antecedent NP, (c) refer to a proposition or situation previously introduced, or (d) refer to an entity inferred from the preceding discourse.

Table 17. Discourse properties of pragmatically motivated pre-predicate patient NP

Property of the pre-predicate NP	Bare stem clauses	*di-* passive clauses
(a) Having a clear antecedent NP	24 (18.1%)	21 (16.3%)
(b) Modified by an antecedent NP	8 (6.0%)	14 (10.9%)
(c) Referring to a proposition or situation introduced in the preceding discourse	70 (52.6%)	79 (61.2%)
(d) Inferable from the preceding discourse	31 (23.3%)	15 (11.6%)
Total	133 (100.0%)	129 (100.0%)

Example sentences of each category in Table 17 are shown below. To clarify the property of each argument, co-referential NPs are marked with a subscript number. Unlike pre-predicate patient NPs, which in many cases do not have a co-referential NP in the preceding discourse, the agent pronoun always has an antecedent.

(a) Pre-predicate NP with a clear antecedent NP

Bare stem clause

(21) [***Lelaki itu***]$_1$ *pun pergi mencari kayu bakar dan menjualnya.* [***Ia***]$_1$ *pulang dengan membawa* [***hasil sepuluh dirham***]$_2$. [***Uang itu***]$_2$ [***ia***]$_1$ ***gunakan*** *sebagian untuk membeli pakaian dan sebagian lain untuk membeli makanan*.

'[**The man**]$_1$ went to look for timber and sold it. Then [**he**]$_1$ returned with [**ten dirham**]$_2$. [**He**]$_1$ **spent** a part of [**that money**]$_2$ on clothes and the remainder to buy food.'

https://pengusahamuslim.com/75-hukum-jual-beli-jual-beli-yang-diharamkan.html

Di- passive clause

(22) *Pada saat berusia 14 tahun dan masih berada di bangku SMA, sambil bekerja* [***ia***]$_1$ *bisa menghasilkan* [***US$ 1,200***]$_2$, [***uang tersebut***]$_2$ ***di-gunakan=***[***nya***]$_1$ *untuk membeli tanah pertanian seluas 40 ha, setelah itu tanah tersebut ia sewakan kepada petani lokal*.

'When he was 14 years old and still in high school, [**he**]$_1$ earned [**US$ 1,200**]$_2$ from his job, (and) [**he**]$_1$ **spent** [**the money**]$_2$ to buy as much as 40 ha of agricultural land. After that, he rented the land to local farmers.'

http://terakreditasi.blogspot.com/2011/05/kisah-sukses-para-investor-kelas-dunia.html

(b) Pre-predicate NP has an antecedent NP as a modifier

Bare stem clause

(23) *Demikianlah ia akan membuat tercengang banyak bangsa,* [***raja-raja***]$_1$ *akan mengatupkan mulut=*[***nya***]$_1$ *melihat dia; sebab* [***apa yang tidak diceritakan kepada*** [***mereka***]$_1$] *akan* [***mereka***]$_1$ ***lihat****, dan apa yang tidak mereka dengar akan mereka pahami*.

'Like that, he will make many groups surprised. [**Kings**]$_1$ will be overwhelmed (lit. close [**their**]$_1$ mouths) to see him. It is because [**they**]$_1$ will **see** [**what have not been told to** [**them**]$_1$] and they will understand what they have not hear.'

http://airhidup.info/wp/beberapa-kasus-baptisan-dalam-alkitab/

Di- passive clause

(24) *Selama pendidikan di SGA,* [***ia***]$_1$ *pernah mengarang* [***nyanyian untuk ibu***]$_2$. *Kata-kata=*[***nya***]$_2$ *bila disimpulkan, berbunyi: betapa dalam laut, betapa tinggi gunung, tidak dapat melebihi dalam dan tingginya kasih Ibu. Sayang,* [***teks*** [***nyanyian ini***]$_2$] *tidak* ***di-miliki=***[***nya***]$_1$ *lagi, hilang*.

'While (he) was in an educational institution, [**he**]$_1$ made [**a song for his mother**]$_2$. The lyric (lit. of [**it**]$_2$) can be summarised as follows: "No matter how deep the ocean is and no matter how high the mountain is, they are not enough to surpass my mother's love." Unfortunately, [**he**]$_1$ does not **have** [**the original text of** [**this song**]$_2$]. It was lost.'

https://id.wikipedia.org/wiki/Abdullah_Totong_Mahmud

(c) Pre-predicate NP refers to a proposition or situation in the preceding discourse

Bare stem clause

(25) *Setelah wafatnya Abu Salarnah,* [[***para pemuka dari kalangan sahabat***]$_1$ ***bersegera meminang Ummu Salamah***]$_2$. [***Hal ini***]$_2$ [***mereka***]$_1$ ***lakukan*** *sebagai tanda penghormatan terhadapat suaminya dan untuk melindungi diri Ummu Salamah.*

'After the death of Abu Salarnah, [[**the leaders of the friend groups**]$_1$ **rushed to greet Ummu Salamah**]$_2$. [**They**]$_1$ did [**this**]$_2$ in honour of her husband and to protect Ummu Salamah.'

http://alifahsm.blogspot.com/2010/12/ummu-salamah.html

Di- passive clause

(26) *Sebagaimana layaknya kebanyakan orang Papua lainnya,* [***Franklin***]**$_1$** *pun merupakan salah satu orang yang* [***gemar menceritakan kisah lucu***]$_2$. [***Hal itu***]$_2$ ***di-gunakan=***[***nya***]$_1$ *untuk membunuh waktu luang di luar atletik.*

'Just like many other Papuans, [**Franklin**]**$_1$** is also one of those who [**like to tell funny stories**]$_2$. [**He**]**$_1$ used** [**this**]$_2$ to kill time in the intervals of track and field meets.'

https://id.wikipedia.org/wiki/Franklin_Ramses_Burumi

(d) The referent of the Pre-predicate NP is inferable from the preceding discourse[17]

Bare stem clause

(27) *Dan sampai saat ini* [***Fisi***]**$_1$** *tercatat sebagai satu-satunya pembalap yang mampu memenangi lomba lewat pengadilan dan tanpa seremonial podium.* [***Serah terima tropi kemenangan***] [***ia***]**$_1$** ***lakukan*** *bersama Kimi di balapan berikutnya di Imola.*

'And until now, [**Fisi**]**$_1$** is recorded as the only F1 driver to have won a race without having stood atop the podium because he had won a race with a (later) adjustment. [**He**]**$_1$** attended [**the Prize giving ceremony**] with Kimi at the following race in Imola.'

https://id.wikipedia.org/wiki/Giancarlo_Fisichella

Di- passive clause

(28) [***Bong***]**$_1$**, *yang kala itu berusia 11 tahun, berempati atas terpuruknya ekonomi keluarga. Kebutuhan sekolah diusahakan sendiri. Contohnya* [***ia***]**$_1$** *lebih memilih kertas bekas dan memfotokopi buku pelajaran milik temannya ketimbang membeli baru.* [***Beberapa alat tulis***] *juga* ***di-buat=***[***nya***]**$_1$** *sendiri. 'Saya menggunakan karet (gelang) untuk penghapus', tuturnya.*

'When he was 11 years old, [**Bong**]**$_1$** was worried that his family was financially embarrassed. He handled school necessities by himself. For example, [**he**]**$_1$** preferred to pick up used papers and make copies of his friends' textbooks, rather than buy new ones. [**He**]**$_1$** also **made** [**writing materials**] by himself. "I made good use of rubber (bands) as a substitute eraser", he said.'

---

[17]This category often includes those sentences whose patient NPs are nouns such as *cara* 'method', *gaya* 'style', and *upaya* 'effort'. In this case, the previous sentence expresses desire or intention, and these nouns describe the way to reach their goal. In this sense, it can be said that such patient NPs are inferable from the preceding discourse. Moreover, these are sometimes followed by detailed information, as shown in the example below.

Banyaknya arus migrasi [***warga di luar pulau Bali***]$_1$ *untuk datang ke pantai Kuta: sebut saja warga dari pulau Jawa dan pulau Lombok, semata - mata untuk mencari penghidupan yang layak dari lembaran dollar turis asing.* [***Beragam cara***] [***mereka***]$_1$***lakukan*** *untuk mendapatkan uang, seperti:* [***menjual jasa pembuatan tatto, pijat, bahkan menyewakan papan selancar bagi para turis***].

'Quite a number of [**immigrants of people living outside Bali**]$_1$ come to Kuta coast: let us say people from Java island and Lombok island, among others, whose sole purpose is to make their living with money from tourists. [**They**]$_1$ **have tried** [**various methods**] in order to earn money, such as: [**provide services of tattooing, massage, and even lend a surfboard to tourists**.]'

http://andibagus.multiply.com/journal?=20

https://vienmuhadi.com/2010/09/01/hindun-binti-suhail-ummu-salamah-ra-ummirul-mukminin/

The information status of the agent and patient arguments in sentences (21)–(28) above may be characterised using the concept of 'activation', whereby an active referent is one 'that is currently lit up, a concept in a person's focus of consciousness at a particular moment' (Chafe 1987). The agent arguments always have a co-referential NP in the preceding discourse and, therefore, can be seen as an activated referent, while the pre-predicate patient NP, which in most cases does not have a co-referential NP, and its referent tend to be less activated than the agent.

The distinction between the agent and patient arguments can also be seen from the point of the discourse functions they play as a topic. Lambrecht (1996: 165) discusses two different functions of topic, one a referring function, which a lexical NP normally performs, and the other a relational role, which an unaccented pronoun normally plays; the former is a function of referring to, naming, or announcing a new topic, while the latter is a function of coding the relation of the topic referent to the whole proposition. If we apply this distinction to the two topic expressions in non-active clauses in Indonesian, we can say that the pre-predicate NP, which almost always occurs in the form of lexical NP, encodes the referring function, while the agent pronouns, which occur as an enclitic in the *di-* passive clauses or an unaccented pronoun, play the relational role. We could say that non-active clauses with a pre-predicate NP have a similar discourse function to that of so-called detachment constructions, in that they are used to promote the representation of a referent from non-active to active state in the addressee's mind[18] in order to name or announce a new topic (Lambrecht 1996: 181ff.).

4.5 *Di-* passive clauses with post-predicate NP

Table 16 in Section 4.3 shows that the post-predicate patient argument occurs in some of the *di-* clauses; the referent may be brand new and cannot be seen as topical, as seen in sentence (29) below. This shows that the so-called 'event focus' usage of *di-* clauses (Hopper 1984: 84) mentioned in the Introduction is still retained, though to a lesser extent.

(29) *Di-lihat=nya **jam dinding**, 'Sudah jam satu dini hari, ia sanggup bertahan selama itu, oohh hebatnya', batin Bu Henny.*

'She saw **the wall clock**, "It's already one o'clock in the morning, he was able to survive that long, oh great", thought Mrs. Henny.'

https://adityamasiver.wordpress.com/page/38/

The stem *lihat* 'see' occurs relatively frequently with the post-predicate patient argument. In more than half of the cases, the patient argument in this position occurs with the relative clause and semantically plays a similar role to a complement clause; it refers to a situation rather than a simple entity. Thus, we can say that post-predicate NP in this environment means having similar contents as a complement clause.

[18] The pre-predicate NP in non-active clauses functions as a regular argument of the clause, and in that respect, is different syntactically from the detached NP in the detachment construction, which normally has a co-referential resumptive pronoun in the clause and is not considered an argument of the clause itself.

(30) *Di dalam rombongan kecil itu pula di-lihat=nya* ***seorang yang bersenjatakan tombak pendek****.*

'In the small troupe he saw **a man armed with a short spear**.'

http://adbmcadangan.wordpress.com/424/2/index.html

4.6 Observations on the distribution of *lakukan* 'do'

The verb stem *lakukan* occurs relatively frequently with pragmatically motivated non-active clauses. In that condition, the clauses very often (approximately 90%) occur with a pre-predicate NP. A detailed examination of its occurrence shows that the structure mostly occurs in a specific usage in which the predicate has a pre-predicate NP consisting of demonstrative *ini*/*itu* 'this/that' or a semantically similar NP, such as *hal ini*/*itu* 'this/that point', referring to a proposition introduced in the preceding discourse and followed by an adverbial clause indicating a reason, purpose, or temporal relation. Sentences (31) and (32) are examples of such environments.

(31) *Sebenarnya, jika ia membaca karya orang lain,* ***itu*** *di-lakukan=nya hanya sekedar* ***untuk*** *menemukan pembenaran pendapatnya sendiri.*

'In fact, when he reads another person's work, he does **it** just **in order to** justify his opinion.'

https://alfablitar.wordpress.com/2012/05/17/kisah-beberapa-tokoh-sosiologi/

(32) ***Hal ini*** *di-lakukan=nya* ***karena*** *ketertarikannya yg besar terhadap sepatu.*

'He did **it because of** his strong interest in shoes.'

http://oopittus.blogspot.com/2010/10/ebc-e-business-and-e-commerce.html

From the observation above, we may say that non-active clauses with the stem *lakukan* are used for the specific communicational function of providing additional circumstantial information about an event or situation already introduced in the preceding discourse, and the presence of the structure boosts the frequency of the pragmatically motivated non-active clauses with the verb stem.

5. Summary

In this paper, we attempted to show the function of non-active voice clauses observed in the web corpora data in the MALINDO Conc.

After introducing the nature of the MALINDO Conc web corpora in Section 2, we have clarified the general distribution of *di-* passive clauses from the view of co-occurrence with the agent pronoun in Section 3; *di-* passive clauses are almost exclusively used with third person agent, which verifies the traditional view shown in Sneddon et al. (2010), among others.

In Section 4, we focused on the six most frequent verb stems in the MALINDO Conc corpus, namely *miliki* 'possess', *lakukan* 'do', *buat* 'make', *lihat* 'see', *gunakan* 'use', and *katakan* 'say', occurring with the third person agent pronoun, to determine the functions of the two non-active voice clauses in that condition. For all the stems, the environment in which the non-active voice is syntactically required, typically when the patient arguments are relativised, is one of the most major usages. This applies to the stem *miliki* 'possess' most extremely; almost all the occurrence of this stem is in a relative clause.

The other verb stems may occur in non-syntactically required environments with different frequencies. A substantial number of such non-active clauses occur with a pre-predicate patient NP that expresses the sentence topic. Thus, the structure has two topic arguments, as the agent, which is expressed by the third person pronoun, also has high topicality there. A detailed investigation of the discourse property of the pre-predicate NP shows that the referent of the pre-predicate patient NP is less activated than the agent that is always activated in the preceding discourse; the pre-predicate NP often does not have a co-referential NP in the preceding discourse. The referent is introduced in the form of a proposition or an entity or a situation inferable from the preceding discourse. We might say that the structure has a function similar to that of the detached NP construction in that the pre-predicate NP has a function of referring to or announcing a new topic, while the agent pronoun has the function of indicating the role of the topical agent in the proposition (see Lambrecht (1994) for the function of the detached construction and the two distinctive functions of topics).

The stem *lakukan* very frequently (approximately 90%) occurs with a pre-predicate NP, taking a specific type of structure in which (i) the pre-predicate NP consists of the demonstratives *ini*/*itu* 'this/that', or a semantically similar NP, such as *hal ini*/*itu* 'this/that issue' which refers to the proposition that was introduced in the preceding sentences, and (ii) the predicate is followed by an adverbial clause indicating circumstantial information, such as a reason, purpose, or temporal relation. Thus, this structure in most cases is used for a fixed communicative function of providing circumstantial information additional to that given by the proposition mentioned in the preceding clauses.

In contrast to the stem *lakukan*, the stems *lihat* 'see', *katakan* 'say', and *buat* 'make' occur less frequently with a pre-predicate NP, and very frequently with the post-predicate complement clause, expressing a recognised event, speech, or caused event, respectively.

There is no major functional difference between the two non-active types in the points mentioned above, except that (i) the verb stems *buat* 'make' and *lihat* 'see' occur far more frequently in *di-* passive clauses than in bare stem clauses. We could say that the forms *di-buat=nya* and *di-lihat=nya* are almost exclusively associated with a fixed semantic function, causative for the former and an indication of the recognised situations for the latter, which are rarely expressed by bare stem clauses, and (ii) bare stem clauses, in general, more frequently occur in syntactically required environments than *di-* passive clauses.

Di- clauses were said to have an 'event focus' usage irrespective of the patient's topicality to indicate the main line of discourse (Hopper 1984). Such a usage is observed in the recent texts included in the MALINDO Conc corpus, though to a much lesser extent, and we can say that this type of usage is diminishing but still retained.

Abbreviations

1	first person	2	second person
3	third person	ACT	active
DP	determiner phrase	EXCL	exclusive
INCL	inclusive	NP	noun phrase
P	patient	PL	plural
PASS	passive	REL	relativiser
V	verb	S	subject

References

Arka, I Wayan & Christopher Manning. 1998. Voice and grammatical relations in Indonesian: A new perspective. In Miriam Butt & Tracy King (eds.), *Proceedings of the 1998 International Lexical Functional Grammar conference*. Stanford CA: CSLI Publications. (Also: Peter K. Austin & Simon Musgrave (eds). 2008. *Voice and grammatical functions in Austronesian*. Stanford, CA: CSLI Publications.)

Chafe, Wallace. 1987. Cognitive constraints on information flow. In Russel Tomlin (ed.), *Coherence and grounding in discourse*, 21–51 Amsterdam: John Benjamins.

Cole, Peter, Gabriella Hermon & Yassir Tjung. 2006. Is there pasif semu in Jakarta Indonesian? *Oceanic Linguistics* 45(1). 64–90.

Cumming, Susanna. 1995. *Functional change: The case of Malay constituent order*. Berlin: Mouton de Gruyter.

Dardjowidjojo, Soenjono. 1978. *Sentence patterns in Indonesian*. Honolulu: University of Hawaii Press.

Givón, Talmy. 1994. *Voice and inversion*. Amsterdam: John Benjamins.

Goldhahn, Dirk, Thomas Eckart & Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf

Grange, Philippe. 2008. Verb nominalization and thematization with the suffix *-nya* in Indonesian. Paper presented in the Twelfth International Symposium on Malay/Indonesian Linguistics (ISMIL 12), 26–27 June 2008.

Hopper, Paul. 1984. Ergative, passive, and active in Malay narrative. In Flora Klein-Andreeu (ed.), *Discourse perspectives in syntax*, 67–88. New York: Academic Press.

Kaswanti Purwo, Bambang. 1988. Voice in Indonesian: A discourse study. In Masayoshi Shibatani (ed.), *Passive and voice*, 195–241. Amsterdam: John Benjamins.

Kroeger, Paul. 2014. Passive agents in Malay: The binding properties and discourse functions of agentive =nya. In Siaw-Fong Chung & Hiroki Nomoto (eds.), *Current trends in Malay linguistics*, 5–29. *NUSA* 57. Jakarta and Tokyo: Universitas Katolik Indonesia Atma Jaya and Tokyo University of Foreign Studies.

Lambrecht, Knud. 1994. *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Nasanius, Yassir, Peter Cole & Gabrielle Hermon. 2016. The formation of relative clauses in Jakarta Indonesian: Data from adults and children. *NUSA* 61. 1–18.

Nomoto, Hiroki, Shiro Akasegawa & Asako Shiohara. 2018a. Building an open online concordancer for Malay/Indonesian. Paper presented at the 22nd International Symposium on Malay/Indonesian Linguistics (ISMIL 22), 11–12 May 2018.

Nomoto, Hiroki, Shiro Akasegawa & Asako Shiohara. 2018b. Reclassification of the Leipzig Corpora Collection for Malay and Indonesian. *NUSA* 65. 47–66. doi:10.15026/92899.

Nomoto, Hiroki, Hannah Choi, David Moeljadi & Francis Bond. 2018. MALINDO Morph: Morphological dictionary and analyser for Malay/Indonesian. In Kiyoaki Shirai (ed.), *Proceedings of the LREC 2018 workshop "The 13th Workshop on Asian Language Resources"*, 36–43. http://lrec-conf.org/workshops/lrec2018/W29/pdf/8_W29.pdf.

Nomoto, Hiroki & Kartini Abd Wahab. 2014. Person restriction on passive agents in Malay: Information structure and syntax. In Siaw-Fong Chung & Hiroki Nomoto (eds.), *Current trends in Malay linguistics*, 31–50. *NUSA* 57. Jakarta and Tokyo: Universitas Katolik Indonesia Atma Jaya and Tokyo University of Foreign Studies.

Shibatani, Masayoshi. 1985. Passives and related constructions: A prototype analysis. *Language* 61(4). 821–848.

Sneddon, James Neil, 2006. *Colloquial Jakarta Indonesian*. Canberra: Australian National University.

Sneddon, James Neil, K. Alexander Adelaar, Dwi N. Djenar & Michael Ewing (eds.). 2010. *Indonesian: A comprehensive grammar* (*2nd edition*). London: Taylor and Francis.