The Indonesian prefixes PE- and PEN-: A study in productivity and allomorphy

Karlina Denistia1 · R. Harald Baayen1

Received: 20 April 2018 / Accepted: 24 January 2019 / Published online: 7 February 2019
© The Author(s) 2019

Abstract This study examines two nominalizing prefixes in Indonesian: PE- and PEN-, which derive nouns from verbs with a range of meanings similar to that found in -er suffix in English. The prefix PE- is form-invariant, whereas PEN- has several nasal allomorphs. Given their similarity in form and function, the question arises of whether PE- and PEN- are allomorphs. We conducted a corpus-based analysis of their productivity, using the written Indonesian corpus in the Leipzig Corpora Collection. In this corpus, PEN- is apparently more productive than PE-. Interestingly, the frequency of words with PEN- correlates significantly with the productivity of the corresponding base verbs. In addition, PEN- is more integrated into the verbal system; verbs that have PEN- are part of larger verb families. PEN- attaches almost exclusively to verbs and creates nouns denoting agents and instruments. By contrast, PE- creates nouns denoting agents and patients and attaches not only to verbs but also to nouns and adjectives. For derived words with PE-, there is no significant correlation between the frequency of the nominalization and the frequency of its base. PE- also does not participate in the linearity of the productivity of the allomorphs of base and derived words that characterizes PEN-. Words with PE- are also more often input to further reduplication and inflectional variants than is the case for PEN-. This corpus-based research thus illustrates that affixes can have different qualitative and quantitative properties, although at first blush they look like allomorphs. Our analyses justify their treatment in the Indonesian literature as separate prefixes.

Keywords Indonesian · Productivity · Productivity paradox · Allomorphy · Paradigmatics

✉ K. Denistia
karlina.denistia@student.uni-tuebingen.de; karlinadenistia@gmail.com

R.H. Baayen
harald.baayen@uni-tuebingen.de

1 Eberhard Karls Universitaet Tuebingen – Quantitative Linguistics Department, Wilhemstrasse 19, 72074 Tuebingen, Baden Wuerttemberg, Germany
1 Introduction

The question addressed in this study is whether two phonologically very similar prefixes of Indonesian are allomorphs or rather independent prefixes. According to the classical definition of allomorphy, variants of a morpheme which have the same underlying form, which share the same meaning, and are in complementary distribution, are classified as allomorphs (Bloomfield 1933; Alber 2011). When two different affixes express roughly the same semantics, they are referred to not as allomorphs but as rival affixes (Aronoff and Anshen 2017). Conversely, when the same form signifies completely different semantic functions, as in the case for English -s (third person singular vs. plural inflection vs. third person genitive, . . . Plag et al. 2017), we have affix homonymy. Less clear-cut are cases where formatives are obviously similar in form as well as in meaning, without the form similarity being phonologically conditioned. For instance, Peters (2004) argued that English -er and -eer are allomorphs where the choice of -eer is semantically conditioned on the referent being from the semantic field of war. Baayen et al. (2013) discussed the Russian prefixes pere- and pre-, which are etymologically related but express subtly different semantics. Endresen (2014) provides detailed discussion of the limitations of the classical definition of allomorphy. She points out that there are counterexamples where other parameters should be taken into account, such as subtle differences in meaning as exhibited by the Russian affix pairs s- vs. so- ‘together’, o- vs. ob- ‘around’, pere- vs. pre- ‘across’, vz- vs. voz- ‘up’, and vy- vs. iz- ‘out of’. The two Indonesian prefixes that are the subject of this study likewise raise the question of whether these prefixes are allomorphs, given their phonological similarity, or separate prefixes. As pointed out by Denistia (2018), Indonesian linguists mainly have described the two morphs as independent prefixes (Ramlan 2009; Sneddon et al. 2010), but there are also studies that take them to be allomorphs (Darwowidjojo 1983; Kridalaksana 2008). Since these two prefixes are similar in form, but not phonologically conditioned, and since they are similar, but not identical in meaning, the classical criteria for allomorphy are only approximately satisfied. Thus, the present study is a corpus based investigation into what Endresen (2014) refers to as non-standard allomorphy. Specifically, we examine in detail the differences in the semantics of PE- and PEN-, the differences in their productivity, and the differences in the extent to which derived words with PE- and PEN- are input to further inflection. In our analyses, the paradigmatic relations between base words and derived words are especially informative.1

In what follows, we first introduce some basic aspects of Indonesian verb morphology and deverbal nominalization. In the next section, we introduce the databases that inform our analyses. We then present our analyses and conclude with a discussion of the results obtained.

2 Indonesian verb morphology and deverbal nominalization

The morphology of Indonesian is characterized by productive processes of affix substitution. In this study, we are interested in two prefixes that create nouns

1For discussion of paradigmatic relations in derivational morphology, see (Marle 1986; Stekauer 2014).
The Indonesian prefixes \textit{PE-} and \textit{PEN-} from verbs through affix substitution, and which express a range of semantic functions (e.g. agent, instrument, patient, Sneddon et al. (2010, pp. 30–33)). One prefix, henceforth \textit{PEN-}, forms nouns from verbs with the prefix \textit{MEN-} (e.g. \textit{penari} ‘dancer’ – \textit{menari} ‘to dance’). In what follows, for notational clarity, we write prefixes in upper case and their allomorphs as subscripts. \textit{PEN-} and \textit{MEN-} have six allomorphs: \textit{PENpeng-}, \textit{PENpen-}, \textit{PENpe-}, \textit{PENpe-}, \textit{PENpeny-}, and \textit{MENmeng-}, \textit{MENmen-}, \textit{MENmen-}, \textit{MENme-}, \textit{MENmeny-} and \textit{MENmenge-}. Sukarno (2017), Ramlan (2009) and Sugerman (2016) summarized the phonological conditioning of these allomorphs as follows:

- \textit{PENpeng-}, \textit{MENmeng-} occurs with base words beginning with a vowel or a velar obstruent /g/, /k/, /h/, or /kh/,
- \textit{PENpen-}, \textit{MENmen-} occurs with base words beginning with a alveolar or palatal obstruent /d/, /t/, /c/, /j/, /sy/, or /z/,
- \textit{PENpem-}, \textit{MENmem-} occurs with base words beginning with a labial consonant /b/, /p/, or /f/,
- \textit{PENpe-}, \textit{MENme-} occurs with base words beginning with a nasal, a semivowel, or a liquid /m/, /n/, /ng/, /ny/, /w/, /j/, /r/, or /l/,
- \textit{PENpeny-}, \textit{MENmeny-} occurs with base words beginning with /s/, and
- \textit{PENpenge-}, \textit{MENmenge-} occurs with monosyllabic base words.

The nasal allomorphy of Indonesian \textit{MEN-} and \textit{PEN-} is an example of classical phonologically conditioned allomorphy.

A second prefix, henceforth \textit{PE-}, forms nouns from verbs with the prefix \textit{BER-}, again through affix substitution (e.g. \textit{petani} ‘farmer’ – \textit{bertani} ‘to farm’), see Ramlan (2009), Ermanto (2016), Sneddon et al. (2010), Putrayasa (2008), Darwowidjojo (1983), Benjamin (2009). \textit{BER-} has \textit{BERbe-} and \textit{BERbel-} as infrequent allomorphs. \textit{BER-} primarily creates verbs expressing reciprocity, reflexivity, or stativity (see Kridalaksana (2007), Ramlan (2009), Putrayasa (2008), Chaer (2008), Sneddon et al. (2010 for other meanings). \textit{BERbe-}occurs with stems beginning with /r/ or with stems the first syllable of which ends with /r/, as in \textit{risiko} ‘risk’, \textit{berisiko} ‘to run the risk’ and \textit{kerja} ‘work’, \textit{bekerja} ‘to work’. \textit{BERbel-} only occurs with the base word \textit{ajar} ‘to teach’, \textit{belajar} ‘to study’ (Sugerman 2016). If \textit{PE-} is regarded as an allomorph of \textit{PEN-}, its conditioning is not phonological, as for the allomorphs of \textit{PEN-}, but morphological: \textit{PEN-} is paradigmatically related to verbs with \textit{MEN-} and \textit{PE-} is paradigmatically related to verbs with \textit{BER-}.\footnote{We use the term paradigmatically related to denote systematic relationships between elements in absenta. Although in derivation, paradigmatic relations are less tightly knit compared to typical inflectional paradigms such as Latin or Estonian (Dressler 1989), derivation also can show paradigmatic organisation (Stekauer 2014). For the importance of paradigmatic organisation for linking elements in compounds, as well as for stress assignment in compounds, see Krott et al. (2009) and Plag (2006) respectively.}

The base words for the verbs and their nominalizations can be verbs, nouns, and adjectives. There is no consistent difference in lexical meaning between simple base verbs and derived verbs (e.g. \textit{buri} ‘to hunt’ – \textit{berburi} ‘to hunt’), although the derived forms may show different syntactic and aspectual behaviour (e.g. \textit{buri} ‘to hunt’ – \textit{memburi} ‘to hunt continuously’) (Nuriah 2004). The simple verb is typically used in imperatives.
Verbs with *MEN-* can be extended with the suffixes *-i* and *-kan*. *MEN-* typically renders a verb explicitly transitive. The suffixes *-i* and *-kan* add a further argument, either a beneficiary or a causer, while often at the same time expressing intensification or iteration (Arka et al. 2009; Sutanto 2002; Tomasowa 2007; Kroeger 2007; Sneddon et al. 2010).

1. transitives and ditransitives
   (a) *tulis* ‘to write’, *menulis* ‘to write something’, *menulisi* ‘to write something on something’
   (b) *tulis* ‘to write’, *menulis* ‘to write something’, *menuliskan* ‘to write something on behalf of someone’

2. causatives
   (a) *panas* ‘hot’, *memanas* ‘to become hot’, *memanasi* ‘to heat up something’
   (b) *panas* ‘hot’, *memanas* ‘to become hot’, *memanaskan* ‘to apply heat to something’

3. transitives and beneficiaries
   (a) *ajar* ‘to teach’, *mengajar* ‘to teach something’, *mengajari* ‘to teach someone something’
   (b) *ajar* ‘to teach’, *mengajar* ‘to teach something’, *mengajarkan* ‘to teach something to someone’
   (c) *kirim* ‘to send’, *mengirim* ‘to send something’, *mengirimi* ‘to send something to someone’

4. iteration and intensification
   (a) *lempar* ‘to throw’, *melempar* ‘to throw something’, *melempari* ‘to throw something repeatedly at something’
   (b) *pukul* ‘to hit’, *memukul* ‘to hit something’, *memukuli* ‘to hit something hard over and over again’

Verbs with *BER-* do not combine with the *-i* suffix, but are found with *-kan* or *-an* to express possession (5, 6) and reciprocity (7, 8):

5. *dasar* ‘base’, *berdasarkan* ‘be grounded in’
6. *alamat* ‘address’, *beralamatkan* ‘to have an address’
7. *gandeng* ‘to hold hands’, *bergandengan* ‘to hold hands with each other’
8. *cium* ‘to kiss’, *berciuman* ‘to kiss each other’

Derived nouns with *PEN-* do not carry the *-i* or *-kan* suffixes, even though they may correspond to verbs with these suffixes. For instance, *penerbang*, ‘pilot’, is paradigmatically related to *menerbangkan* ‘to fly an aircraft’ rather than to the verb *menerbangi*, ‘to fly in something’, with the suffix *-i* marking location. Importantly, the verb *menerbang* does not exist but only the verbs *terbang*, ‘fly’, *menerbangkan* and *menerbangi*.

Occasionally, one finds both *PEN-* and *PE-*.. There are 5 cases in which the form with *PE-* semantically refers to a profession and the form with *PEN-* does not, as listed in (9). There are also some cases in which the form with *PEN-* expresses agent, causer, or instrument and the form with *PE-* expresses patient or agent. In this case, 7 instances are attested in our database, as listed in (10).
(9) **PEN-** and **PE-** formations that both express agents
(a) *tembak* ‘to shoot’, *penembak* ‘someone who shoots’ and *petembak* ‘shooter’ (athlete)
(b) *tinju* ‘to punch’, *peninju* ‘someone who punches’ and *petinju* ‘boxer’ (athlete)
(c) *terjun* ‘to sky dive’, *penerjun* ‘someone who sky dives’ and *peterjun* ‘sky diver’ (athlete)
(d) *selam* ‘to dive’, *penyelam* ‘someone who dives’ and *peselam* ‘diver’ (athlete)
(e) *dayung* ‘to paddle’, *pendayung* ‘someone who paddles’ and *pedayung* ‘paddler’ (athlete)

10. **PEN-** and **PE-** formations expressing different semantic roles
(a) *ajar* ‘to teach’, *pengajar* ‘teacher’ (agent) and *pelajar* ‘student’ (patient)
(b) *kasih* ‘to love’, *pengasih* ‘lover’ (agent) and *pekasih* ‘love poison’ (instrument)
(c) *sakit* ‘to be sick’, *penyakit* ‘disease’ (causer) and *pesakit* ‘a person with disease’ (patient)
(d) *sapa* ‘to greet’, *penyapa* ‘a person who greets’ (agent) and *pesapa* ‘a person who is greeted’ (patient)
(e) *siar* ‘to announce/to sail’, *penyiar* ‘radio announcer’ (agent) and *pesiar* ‘a cruise ship’ (instrument)
(f) *tanda* ‘sign’, *penanda* ‘a sign’ (agent) and *petanda* ‘a hint’ (patient)
(g) *tempur* ‘to combat’, *penempur* ‘armament’ (agent) and *petempur* ‘combatant’ (instrument)

We compiled a database containing 3090 words with **PE-** and **PEN-**. Since **PEN**- and **PE-** share the form *pe-*, the question arises of how to assign occurrences of the form *pe-* to either **PEN-** or **PE-**. In 235 out of 240 potentially ambiguous forms, inspection of the paradigmatic relation with the corresponding base verb, either a verb with **MEN**- or a verb with **BER**-, the noun can be unambiguously assigned to be **PEN-** or **PE-**. Five words remain ambiguous: *pewushu*, ‘wushu athlete’, *perindang*, ‘provider of shadow’, *pemagang*, ‘probationer’, *pemuda*, ‘young male’, and *pemudik*, ‘homecomer’. The semantics of *pewushu* clarify that it belongs to **PE-**, the prefix that is used to denote professional athletes. The remaining 4 words are truly ambiguous, but are most likely, given their semantics, belong to the class of **PEN-** formation. For instance, *perindang* realises a causative reading, which, as we shall see below, is predominantly expressed by **MEN-**.

The goal of this paper is to clarify the morphological status of **PE-** and **PEN-**, allomorphs or separate prefixes, through a quantitative survey of their productivity, their paradigmatic relations with their base verbs, and the extent to which these derived nouns are input for further inflection. Indonesian inflection comprises several bound morphs: *-ku*, *-mu*, and *-nya* for first, second, and third person singular possessives or objects, *ku- and kau-* for first and second person subjects (Sneddon et al. 2010). In the Indonesian literature, these bound morphs are referred to as clitics, as they are phonologically reduced forms of free pronouns (Kridalaksana 2008). There are also two suffixes that attach to verbs or nouns to express emphasis (*-lah* and *-pun*) or questioning (*-kah*). In what follows, we will refer to these morphs as inflectional, as they do not give rise to new onomasiological units but rather modify existing words much in
the same way as adverbs modify verbs in English. Indonesian also has reduplication, which is used to express the plural for nouns and realizes various semantics function on verbs and adjectives, including intensification and iteration (Sugerman 2016; Chaer 2008; Rafferty 2002; Dalrymple and Mofu 2012). Following Booij (1996), we distinguish between inherent and contextual inflection. Agreement marking on verbs (e.g., *ku*- and *kau*) exemplifies contextual inflection, which is syntactically governed. Inherent inflection is more similar to word formation and hence in some languages can feed derivation and compounding. For instance, in Dutch, plural nouns can appear as left constituents in compounds (Schreuder et al. 1998). Reduplication in Indonesian is inherent inflection: it is not governed by syntactic context (*marah-marah ‘very angry’, anak-anak ‘children’, berhenti-berhenti ‘to stop repeatedly’*), and can feed further inflection, as in *memukul-mukuli* ‘to hit intensively over and over again’, which has as parse [[[me N + [pukul] N] V + pukul] V + i] V. We shall see below that derived words with *PE-* are more often input to these inflectional processes than derived words with *PEN-. We will argue that the joint quantitative evidence justifies to analyse *PE-* and *PEN-* as two distinct prefixes rather than as allomorphs. In the next section, we introduce the databases that we derived from a 36 million token corpus of written Indonesian (Goldhahn et al. 2012).

3 Materials

We created a database from the Indonesian corpus that is part of the Leipzig Corpora Collection at http://corpora2.informatik.uni-leipzig.de/download.html, accessed in April 2016. This corpus comprises a variety of written registers (the web, newspapers, Wikipedia) dating from the years 2008–2012 (Goldhahn et al. 2012). There are 112,025 different word types in this corpus, that occur in 2,759,800 sentences, to a total of 36,608,669 word tokens.

The words in the corpus were morphologically analyzed using the MorphInd parser, which has an overall accuracy of 84.6% (Larasati et al. 2011) and it was run in single word mode, i.e., compounds were not parsed. Prior to running the parser, the 200 words with *PE-* or *PEN-* that contained a typo were corrected manually. The MorphInd parser’s results for *PE-* and *PEN-* were checked and corrected manually against the online version of *Kamus Besar Bahasa Indonesia* (hereafter called the dictionary), a comprehensive dictionary of Indonesian (http://kbbi.kemdikbud.go.id; accessed on June 2016), to verify the morphological status and semantics of the *PE-* and *PEN-* words. We made use of the fourth edition, published in 2012, which has more than 90,000 lemmas (Alwi 2012). The language it records is formal; it omits words that are considered slang or foreign. Where the dictionary and the MorphInd are in conflict, we followed the dictionary. Where the dictionary does not provide information on the word category of the base, we followed the MorphInd parser. The precision of the parser for these words was 0.98 and its recall was 0.82, using the dictionary as the gold standard complemented with manual verification for out-of-vocabulary words.

Sample output of the parser is shown in Table 1: a morphological segmentation is provided where available, as well as a word category label. Table 1 shows that Mor-
Table 1 Examples of the output of the MorphInd parser

| Word       | Translation | Allomorph | Base      | Parser                          |
|------------|-------------|-----------|-----------|--------------------------------|
| pemerintah | government  | pem       | perintah  | peN+perintah<n>_NSD            |
| pemain     | player      | pe        | main      | peN+main<v>_NSD                |
| petugas    | officer     | tugas     | petugas   | petugas<n>_NSD                 |
| pekebun    | gardener    | kebun     | pekebun   | pekebun<x>_X–                  |
| pengusut   | investigator | peng      | usut      | peN+kusut<a>_NSD               |
| pengelas   | welder      | penge     | las       | peN+kelas<n>_NSD               |

phInd identifies pemerintah and pemain correctly. However, it is not able to identify PE- in petugas and pekebun. In some cases, the base identified by the parser is incorrect. For instance, pengusut is formed from usut (to investigate) [PENpeng-+usut], but MorphInd identifies its base as kusut (tangled) [PENpeng-+kusut]. MorphInd also is not always able to accurately identify single syllable base words. In the above examples, this is illustrated by pengelas (welder) which derives from las (weld), [PENpeng-+las], and not kelas (classroom), [PENpeng-+kelas]. Therefore, the output of the parser was manually checked and corrected when necessary.

We processed the data using the R (version 3.3.2) programming language (R Team 2008) in R Studio (R Team 2015). The databases and the R scripts used to construct these databases are available online at http://bit.ly/PePeNProductivity. In what follows, we first present the database with Indonesian verbs, and then proceed to the database with derived nouns with PE- and PEN-.

3.1 The database of Indonesian verbs

Indonesian has deverbal morphology for active, passive, causative, and transitive semantics among others, see Table 2 for examples. From the corpus, we retrieved all verbs recognized by the MorphInd parser and brought these together in a database. The total number of types in the database is 26996. Table 3 illustrates that for each verb, we provide information on the derived word’s frequency in the corpus, the parse provided by MorphInd, the base word, the word category of the base, and the affix or affixes in the verb. When particles (e.g. -lah, -kah, -pun) or affixes (e.g. ku-, -ku, kau-, -mu, -nya) are found attached to a verb (Sneddon et al. 2010; Sugerman 2016), this form is listed with its own entry.3

The database comprises 2489 simple verbs and 24507 affixed verbs (3665 verbs with suffixes, 11562 verbs with prefixes, and 9280 verbs with both prefix and suffix). We observed 27 verb constructions of which 13 are reported in the literature ((Hidajat 2014; Fortin 2006; Sneddon et al. 2010; Benjamin 2009; Arka et al. 2009; Sudaryanto 1993; Kridalaksana 2007)). In our corpus, there are 2 attested verb constructions (e.g. terke-/-an and terper-/-an) that are not productive (1 token and 8 tokens respectively). Table 2 lists the 25 productive constructions.

3The column MorphologicalVariation specifies the related particles or affixes. English translations in the tables of this paper are provided for convenience but are not part of the databases.
Table 2  Examples of simple and complex verbs in Indonesian, and affix combinations in complex verb as attested in the corpus

| Base translation | Translation | Derived verb | Affix semantics | Base word |
|------------------|-------------|--------------|-----------------|-----------|
| base             | to have the base | berdasarkan | having X | dasar |
| to meet          | to meet each other | bertemu | reciprocative | temu |
| to continue      | to countinuously continue | berkelanjutan | continuous – intransitive | lanjut |
| to assume        | to have assumption | beranggapan | having X | anggap |
| tools            | having many kinds of tools | berperalatan | having many kinds of X | alat |
| to help          | to help | membantu | active-transitive | bantu |
| to weight        | to consider | mempertimbangkan | active-transitive-causative | timbang |
| to teach         | to teach someone | mengajarkan | active-transitive-beneficiary | ajar |
| to remember      | to commemorate | memperingati | active-transitive-causative | ingat |
| easy             | to make something easier | mempermudah | active-transitive-causative | mudah |
| stop             | to make someone stop | memberhentikan | active-transitive-causative | henti |
| to hit           | to hit repeadly | memukuli | active-transitive-iterative | pukul |
| to find          | to be found | ditemukan | passive | temu |
| to carry         | to be carried | dibawa | passive | bawa |
| to agree         | to be agreed | disetujui | passive | setuju |
| wide             | to be made wider | diperluas | passive-causative | luas |
| to fight         | to be fought for | diperjuangkan | passive | juang |
| weapon           | to be equipped with weapon | dipersenjatai | having many kinds of X | senjata |
| to sleep         | to sleep unintentionally | tertidur | accidental | tidur |
| to be separated  | can be separated | terpisahkan | intransitive | pisah |
| to beat          | can be defeated | tertandingi | intransitive | tanding |
| laughter         | to laugh intensively at something | tertawaan | intransitive | tawa |
| to look          | to make something visible | kelihatan | adversative | lihat |
| to say           | to make someone say something | katakan | causative | kata |
| to meet          | to make someone meet someone else | temui | causative | temu |

As our specific interest is in nouns with PE- and PEN-, we extracted from this database all the verbs that correspond to these nominalizations and that carry the prefix BER- or MEN-. To this new database, henceforth the MeBer Database,
Table 3  Examples of entries in the verb database

| Word       | Translation | Frequency | MorphInd     | Base word | Base word class | Morphological variation | Affix |
|------------|-------------|-----------|--------------|-----------|-----------------|-------------------------|-------|
| menjadi    | become      | 15475     | meN+jadi     | jadi      | a               | meN-                    |       |
| bersama    | be together | 32200     | ber+sama     | sama      | 3rdSingPronoun  | ber-                    |       |
| terkait    | be connected to | 25561  | ter+kait     | kait      | v               | ter-                    |       |
| ujarnya    | his/her saying | 22957 | ujar<v>_VSA+ dia<p>_PS3 | ujar      | v               | simple                  |       |
| turun      | go down     | 18970     | turun<v>_VSA | turun     | v               | simple                  |       |
| diduga     | be suspected | 11240    | di+duga      | duga      | v               | di-                     |       |

we added information on the frequency of the base words of these complex verbs, whether the verbal prefix is MEN- or BER- and also the allomorph of MEN- (see Table 4). Whereas all nominalizations with PEN- have a corresponding verb with MEN-, there is one simple verb, sohor ‘to be famous’, that has a corresponding nominalization with PE-, pesohor ‘a famous person’, without having a corresponding verb with BER-. This verb-noun pair is not in the MeBer Database, but in a separate database (SimpleWords) which also specifies the frequency of the base verb and the frequency of the derived noun (see Sneddon et al. (2010) for discussion of such exceptional pairs).

All the data in MeBer Database were compiled computationally from the output of the MorphInd and subsequently checked manually using the dictionary. In total, there are 8484 words with the MEN- prefix and 3582 words with the BER- prefix. These counts include forms with the suffixes -i, -kan or -an. To this database, we added some words such as beserta ‘to be together with’, belajar ‘to study’, beternak ‘to farm’, bekerja ‘to work’, and beterbangan ‘to fly randomly’ and their inflectional variants, forms which MorphInd did not recognize but that we happened to identify in the course of this study. The MorphInd parser also does not recognize verbs with the allomorph menge-. For the 18 nominalizations with PENpenge-, we manually searched for the occurrences of the corresponding verbs and added these together with their frequency counts to the MeBer database. Finally, a total of 297 verbs with MEN- and 14 verbs with BER- were not recognized by the parser, and were corrected manually on the basis of the dictionary.4

3.2 The PePeN database

We brought together the PE- and PEN- words in a lexical database, henceforth the PePeN database. This database also includes the noun with PE- that have a simple verb as the base. In this way, we obtained a total of 3090 words, 267 with PE-, 2818

4We suspect that the base words of the MEN- and BER- verbs were not in MorphInd’s dictionary.
| Word       | Translation | Frequency | MeN | MeN allomorph | Base word | Base frequency | Base word class | MorphInd | Morphological variation |
|------------|-------------|-----------|-----|----------------|-----------|----------------|-----------------|----------|------------------------|
| mengatakan| to say      | 119115    | TRUE | meng           | kata      | 151552         | n               | meN+kata<n>++kan_VSA |
| melakukan  | to do       | 76116     | TRUE | laku           | 1689      | adj             | meN+laku<a>++kan_VSA |
| memiliki   | to have     | 62317     | TRUE | milik          | 12427      | v               | meN+milik<v>++kan_VSA |
| membuat    | to make     | 46242     | TRUE | buat           | 8431       | v               | meN+buat<v>++kan_VSA |
| memberikan | to give     | 43803     | TRUE | beri           | 1295       | v               | meN+beri<v>++kan_VSA |
| berada     | to be       | 36567     | FALSE| ada            | 196988     | adj             | ber+ada<a>++kan_VSA |
| bersama    | to be together | 32206   | FALSE| sama           | 49719      | adj             | ber+sama<a>++kan_VSA |
| berdasarkan| to be the basis | 19248 | FALSE| dasar          | 13180      | n               | ber+dasar<n>++kan_VSA |
| berbeda    | to be different | 17895  | FALSE| beda           | 1861       | adj             | ber+beda<a>++kan_VSA |
| berlangsung| to be on going | 17558  | FALSE| langsung       | 37225      | adj             | ber+langsung<a>++kan_VSA |
| melakukan  | to do it    | 2800      | TRUE | laku           | 1689       | adj             | meN+laku<a>++kan_VSA+dia<p>++PS3 |
| membuatnya | to make it  | 2775      | TRUE | buat           | 8431       | v               | meN+buat<v>++kan_VSA+dia<p>++PS3 |
| meningkatnya | the increase | 2311   | TRUE | men            | tingkat    | 22098          | n               | meN+tingkat<n>++kan_VSA+dia<p>++PS3 |
| berkurangnya | the decrease | 638    | FALSE| kurang         | 18761      | adj             | ber+kurang<a>++kan_VSA+dia<p>++PS3 |
| bertambahnya | the addition | 593    | FALSE| tambah         | 5457       | v               | ber+tambah<v>++kan_VSA+dia<p>++PS3 |
with PEn-, and 4 words with the unproductive variant PER- (Benjamin 2009). There are 34 words that the MorphInd parser did not analyze.

All derived words were annotated manually for semantic role (agent, instrument, causer, patient, and location), and checked (for at least one token) against both the dictionary and usage in the corpus. As in English, where -er nominalizations may express multiple semantic roles (Booij 2010; Booij and Lieber 2004) (e.g. printer, which has both an agent and instrument reading), Indonesian PE- and PEn- formations can have multiple interpretations (see Table 5). In this study, we did not distinguish between impersonal agent⁶ and instrument. Although it is well known that PEn- create agents, patients, and instruments (Sneddon et al. 2010), we observed a few cases of causer (e.g. penyakit ‘disease’) and location (e.g. penghujung ‘the end’) in our database. It is possible, even likely, that semantic roles are in use in the corpus without being registered in the database, as manual verification of all 579564 tokens with PE- or PEn- in the corpus was infeasible. In the database, words with more than one semantic role have multiple entries in the database, with one row for each role (cf. Table 5). The frequencies listed in rows of Table 5 are those of the overall frequency of the word and are not broken down by semantics.

The PePeN Database thus provides the following information:

1. Word frequency: the token frequency of the derived word in the corpus,
2. Allomorph: the form of the PEn- prefix; where the allomorph does not follow the rules as given in Chaer (2008), Sneddon et al. (2010), e.g. penglihat ‘seer’ is expected to be pelihat, this is marked in the ‘notes’ column of the database as AllomorphDeviation,
3. Base word,
4. Word category of the base word,
5. Base word frequency: the token frequency of the base word in the corpus,
6. MorphInd output as illustrated in Table 1,

5There are four such forms in our database, pertapa and petapa, and their reduplicated variants petapa-petapa, pertapa-pertapa.
6Booij (1986) uses the term impersonal agent for the meaning ‘radio station’ of the Dutch word zender which also has an agentive reading, ‘one who sends’, and an instrumental reading, ‘transmitter’.

| Word       | Translation     | Semantic role |
|------------|-----------------|---------------|
| pembanding | who compares    | agent         |
| pembanding | tool to compare | instrument    |
| pembanding | who is compared | patient       |
| penggerak  | tool to move    | instrument    |
| penggerak  | who moves       | agent         |
| pemici     | a trigger       | agent         |
| pemici     | who triggers    | instrument    |
| pewaris    | heir            | agent         |
| pewaris    | who gives inheritance | patient |
Table 6  Example entries in the PePeN database

| Word       | Translation         | Frequency | Allomorph | Base | Base frequency | Semantic role | Morphological variation |
|------------|---------------------|-----------|-----------|------|----------------|----------------|------------------------|
| pemerintah | government          | 78047     | pem       | perintah | n             | 4315          | agent                  |
| pelaku     | doer                | 17776     | pe        | laku   | n             | 1689          | agent                  |
| penyakit   | disease             | 12042     | peny      | sakit  | adj            | 20454         | causer                 |
| pengusaha  | entrepreneur        | 8053      | peng      | usaha  | n             | 18041         | agent                  |
| pendukung  | supporter           | 5960      | pen       | dukung | v             | 710           | agent                  |
| pelajar    | student             | 3421      | ajar      | n      | 729           | patient        |
| penasihat  | advisor             | 2050      | pe        | nasihat | n            | 386           | agent                  |
| penyebabnya| his/her causer      | 1106      | pen       | sebab  | n             | 18271         | agent Poss3rdSing      |

Table 7  Example entries in the PePeN database illustrating spelling variants and typos (pemain is the second most frequent PEN- nominalizations in the database)

| Word       | Translation         | Frequency | TypoRevision | FreqOfTypo | FreeVariance |
|------------|---------------------|-----------|--------------|-------------|--------------|
| pebowling  | bowling player      | 23        | peboling     | 23          | TRUE         |
| pengonsumsi| consumer            | 12        | pengkonsumsi | 12          | TRUE         |
| pemain     | player              | 34704     | pemian,pemaen, pemasi,pemein, pemai,pemailn, pemainn,pemian, pemjain | 20,4, 2,2, 1,1, 1,1, 1 | FALSE         |

7. Semantic role of the derived noun with respect to the base word (agent, instrument, patient, ...),
8. Morphological variation: reduplications, particles (e.g. -lah, -pun, per-) or affixes (e.g. -ku, -mu, -nya), if present,
9. Typo: whether the form in the corpus had a spelling error (corrected in the database, frequency counts include the frequency of the corrected typos); when several spelling alternants are in use, this is indicated in the FreeVariance column of the database as illustrated in Table 7.

Entries of this database are listed in Table 6.

4 Analysis

4.1 Productivity of PE- and PEN- derived nouns

The PE- and PEN- prefixes differ in their productivity. As shown in the upper panel of Table 8, PEN- occurs with more tokens, more types, and more hapax legomena compared to PE-. Further detail is provided by the lower panel of Table 8, which shows the numbers of tokens, types, and hapaxes for PEN- allomorphs and PE-.
The Indonesian prefixes PE- and PEN-

Table 8  Counts of tokens, types, and hapaxes for PE- and PEN- (upper table) for the six allomorphs of PEN- (lower table)

| Prefix | Tokens | Types | Hapaxes |
|--------|--------|-------|---------|
| PEN    | 498484 | 2221  | 588     |
| PE     | 81083  | 184   | 45      |
|        |        |       |         |
|        |        |       |         |
| Prefix | Tokens | Types | Hapaxes |
|--------|--------|-------|---------|
| penge- | 535    | 18    | 6       |
| peny-  | 38533  | 244   | 75      |
| peng-  | 83515  | 628   | 173     |
| pen-   | 91985  | 546   | 142     |
| pe-    | 138165 | 417   | 103     |
| pem-   | 145696 | 364   | 89      |

Fig. 1  Rank-frequency curves for PE- and PEN- (left panel), and for PE- and sum of the allomorphs of PEN-’s frequency (right panel). PE- is less productive than PEN-, and it is also less productive than the allomorphs of PEN-, with the exception of PENpenge-, which is attested with only 18 types (Color figure online)

Figure 1 presents rank-frequency plots for PE- and PEN- (left panel), and for PE- and all allomorphs of PEN- (right panel), using logarithmic scales (Zipf 1935, 1949). The left panel clarifies that the highest ranked words with PEN- also exceed in frequency the highest ranked words with PE-. Nevertheless, the productivity index V1/N (Baayen 2009) remains greater for PEN- (0.00118) than for PE- (0.00055). The second panel of Fig. 1 shows that four of the six allomorphs of PEN- have rank-frequency curves that lie above the rank-frequency curve of PE-. The curve for PENpeny-, crosses the curve for PE- around rank 50, but still shows many more low-frequency formations. The only allomorph that is less productive than PE- is PENpenge-, an allomorph that attaches to monosyllabic words and which appears in the corpus with only 18 types.

Given the similarity of PE- and PEN- form, the question arises of whether it makes sense to consider PE- as a low productivity allomorph of PEN-. To address this question, we examined the counts of types and hapax legomena for PE- and the allo-
Fig. 2  Counts of types for base verbs (horizontal axis) and counts of types and hapaxes for PE- and PEN- (vertical axis); solid and dashed lines represent regression lines to the PEN-allomorphs for counts of types and counts of hapax legomena respectively.

Table 9 Cross-tabulation of PE- and the allomorphs of PEN- by semantic role. Upper table: counts of types; lower table: counts of hapax legomena

|       | Agent | Causer | Instrument | Location | Patient |
|-------|-------|--------|------------|----------|---------|
| PE-   | 170   | 0      | 3          | 0        | 15      |
| pe-   | 316   | 0      | 94         | 1        | 6       |
| pem-  | 271   | 0      | 91         | 0        | 2       |
| pen-  | 412   | 0      | 130        | 3        | 1       |
| peng- | 474   | 0      | 149        | 4        | 1       |
| penge-| 13    | 0      | 5          | 0        | 0       |
| peny- | 185   | 6      | 53         | 0        | 0       |

|       | Agent | Causer | Instrument | Location | Patient |
|-------|-------|--------|------------|----------|---------|
| PE-   | 39    | 0      | 0          | 0        | 6       |
| pe-   | 75    | 0      | 25         | 1        | 2       |
| pem-  | 62    | 0      | 27         | 0        | 0       |
| pen-  | 108   | 0      | 33         | 0        | 1       |
| peng- | 136   | 0      | 37         | 0        | 0       |
| penge-| 3     | 0      | 3          | 0        | 0       |
| peny- | 59    | 1      | 15         | 0        | 0       |

morphs of PEN- as a function of the number of base verbs with BER- and base verbs with allomorphs of MEN-. The panel of Fig. 2 shows that the rate at which base verbs give rise to derived nouns is the same (according to a regression model) for all allomorphs of MEN- and that PE- patterns as an outlier, both with respect to type counts and with respect to hapax legomena. It is remarkable that the rate at which hapaxes and types appear is so constant across the allomorphs of PEN- and MEN-. From this, we draw the conclusion that the outlier PE- is best understood as a formative in its own right. We note here that Indonesian PEN- and MEN- offer a remarkable window on the relation between base productivity and derived productivity.

Further evidence that PE- is not an allomorph of PEN- emerges when we take the semantic roles of the derived noun into account. Table 9 cross-tabulates PE- and the allomorphs of PEN- by the semantic roles of these nouns in our database;
The Indonesian prefixes \textit{PE-} and \textit{PEN-}

Fig. 3 provides the corresponding visualisation for the three roles that are most frequent: agent, patient, and instrument. Both \textit{PE-} and \textit{PEN-} create agent nouns. \textit{PE-} shows some productivity for patient nouns, of which there are proportionally very few among the nouns with \textit{PEN-}. (The numbers are small, but this asymmetry is significant according to a chi-squared test, $\chi^2_{(1)} = 81.32$, $p < 0.0001$; interestingly, the few patient nouns with \textit{PEN-} are realised with the allomorph \textit{pe-}, however, the proportion of patient hapaxes is much lower ($0.02$ for \textit{PEN-} and $0.13$ for \textit{PE-}, $p < 0.015$, proportion test). Conversely, \textit{PEN-} is productive for instruments, which are virtually absent for \textit{PE-}. This may be one of the reasons that \textit{PEN-} is more productive than \textit{PE-}. For \textit{PEN-}, a chi-squared test indicates that the ratios of agents to instruments are proportional across all allomorphs ($\chi^2_{(5)} = 1.01$, $p > 0.1$ and $\chi^2_{(5)} = 5.48$, $p > 0.1$ for both types and hapax legomena). The uniformity of semantic functions across the allomorphs of \textit{PEN-} is perfectly in line with the fact that these allomorphs are phonologically conditioned. Conversely, the lack of productivity for instruments that characterizes \textit{PE-}, and its (limited) productivity for patient nouns that is strongly attenuated for \textit{PEN-} is a further indication that \textit{PE-} is unlikely to be an allomorph of \textit{PEN-}. Thus, Indonesian \textit{PEN-} and \textit{PE-} show the kind of semantic specialisation that led Baayen et al. (2013) to conclude that Russian \textit{pere-} and \textit{pre-} are not allomorphs but independent prefixes.

The counts underlying Table 9 and Fig. 3 are based on a type definition that distinguishes between forms of the noun with different possessive suffixes or suffixes expressing emphasis, as well as noun plurals. When such variants are collapsed into a single type, the pattern of results on the ratios of agents to instruments across all allomorphs remains similar ($\chi^2_{(5)} = 0.75$, $p > 0.1$ and $\chi^2_{(5)} = 5.11$, $p > 0.1$ for both types and hapax legomena). However, the number of distinct types for patient nouns with \textit{PE-} reduces to 5, each of which occurs more than once. Thus, \textit{PE-} appears to be
Table 10 Counts of variants types for PE- and allomorphs of PEN-. The base represents the non-variant forms. Particles, possessive suffixes, and plural reduplications dominate the counts

| PE-   | pe-  | pem- | pen- | peng- | pense | peny- |
|-------|------|------|------|-------|-------|-------|
| base  | 84   | 247  | 218  | 325   | 396   | 15    | 160   |
| BoundMorpheme | 3    | 0    | 0    | 1     | 1     | 0     | 0     |
| Particle | 13   | 20   | 14   | 16    | 15    | 0     | 2     |
| Possession | 42   | 93   | 86   | 133   | 140   | 2     | 55    |
| Possession+Particle | 1    | 1    | 1    | 1     | 7     | 0     | 1     |
| Reduplication | 34   | 43   | 36   | 61    | 56    | 1     | 23    |
| Reduplication+Particle | 0    | 0    | 0    | 0     | 1     | 0     | 0     |
| Reduplication+Possession | 10   | 13   | 9    | 9     | 12    | 0     | 3     |

Fig. 4 Mosaic plot for the cross-classification of PE- and PEN- by type of inflection. The colour coding represents the Pearson residuals, which clarify where the observed counts are greater (purple) or smaller (pink) than the expected values. A chi-squared test confirms that PE- and PEN- distribute differently over inflectional types ($\chi^2(4) = 36.59, p < 0.0001$) (Color figure online)

well-entrenched for a handful of patient nouns, but does not show real productivity here.

Krott et al. (1999) reported the paradoxical finding that words with less productive affixes tend to be used more as base words for further word formation. A similar observation holds for PE- and PEN-, but now for inflection rather than word formation. Inflectional variation is well illustrated by the noun pengikut ‘follower’, which is attested in the corpus with 9 variants: pengikutku ‘my follower’, pengikutmu ‘your follower’, pengikutnya ‘his/her follower’; reduplication as in pengikut-pengikut ‘followers’; reduplication and affixes as in pengikut-pengikutmu ‘your followers’, pengikut-pengikutnya ‘his/her followers’; affixes and particles as in pengikut mun ‘your follower’ (contrastive your, i.e., your, not somebody else’s follower), pengikut mun ‘his/her follower’ (contrastive), pengikut yah ‘his/her follower’ (contrastive in imperative mood). Table 10 shows the counts of the different kinds of inflections types for PE- and PEN-. In our corpus, particles (e.g. -lah, -pun), possessive suffixes (e.g. -ku, -mu, -nya), and plural reduplications are used most often. Figure 4 presents a
mosaic plot for the cross-classification of \emph{pe} and \emph{PEN}- by type of inflection. The mosaic plot shows that inflected forms of \emph{PE}- are overrepresented for particles, plurals, and combinations of plurals and possessives. In other words, the less productive prefix, \emph{PE}-, is used more intensively as input for further inflection than is the case for \emph{PEN}-. This is likely to be due to the greater entrenchment of words with \emph{PE}- in the mental lexicon, which makes them more readily available for more further affixation. Thus, the same principles that Krott et al. (1999) reported for derivation in Germanic languages generalize to inflection in Indonesian.

4.2 The base verbs of \emph{PEN}- and \emph{PE}-: \emph{MEN}- and \emph{BER}-

Several studies call attention to the tight relation between \emph{PE}- and \emph{PEN}- and their verbal base words (Putrayasa 2008; Chaer 2008; Ramlan 2009; Kridalaksana 2007; Darwowidjojo 1983). We therefore inspected the productivity of verb formation, focusing on monomorphemic words as potential base words. In our database, a total of 5581 such monomorphemic words is attested, with 3617 simple nouns, 943 simple adjectives, and 1021 simple verbs. As shown in Table 2, a large number of affixes is available for creating verbs from nouns, adjectives, and verbs. For this study, the number of different complex verb forms will be referred to as a monomorphemic word’s verb family size. The verb family size measure includes inflectional variants of the verbs in its counts. Plots of this verb family size against base frequency show that, as expected, a higher base frequency predicts a greater verb family size. Interestingly, the functional form of this relation is different for base words that give rise to nouns with \emph{PEN}-, and those that do not. This is illustrated in Fig. 5 (see also Table 11), which present the results of a GAM (Generalized Additive Model, \textsc{mgcv} package version 1.8–17, Wood (2006, 2011)) with a poisson link fitted to the verb family size with centered log base frequency as the predictor. The increase of verb family size with base frequency is greater when \emph{PEN}- is present, as can be seen by comparing the right panel with the left. In the right panel, we see a linear increase, whereas in the left panel, there is no increase at all for the lowest frequency base words. For the larger part of the range of the base word frequencies, the verb family size is larger if the verb family has a noun with \emph{PEN}-. We also considered the base words with \emph{PE}- in the verb family, but as the resulting curve was not significantly different from that of base words with verb families that did not have either nominalization, the two sets were merged into one defined by the absence of \emph{PEN}- in the verb family. Apparently, base productivity and derived productivity are interacting for \emph{PEN}-, but independent for \emph{PE}-.

Figure 6\(^7\) presents mosaic plots for the cross-classification of word category and the presence of \emph{PE}- or \emph{PEN}- in a monomorphemic base word’s verb family. The mosaic plot in the left panel concerns base words that have at least one formation in their verb family (i.e. with neither \emph{PE}- and \emph{PEN}-, with \emph{PE}-, or with \emph{PEN}-). The plot shows that simple words that give rise to affixed verbs but not to any formations with \emph{PE}- or \emph{PEN}- are overrepresented for nouns, and that base words that have \emph{PEN}- in their verb family are overrepresented for verbs, unsurprisingly \((\chi^2_{(d)} = 839.97, p <\)

\(^7\)This plot is created using \textsc{vcd} package version 1.4.4 (Zeileis et al. 2007)
Table 11  GAM summary for partial effects for verb family size regressed on centered log base frequency, for morphological families including derived nouns with PEN- and without PEN- but possibly including nouns with PE-.

| A. parametric coefficients | Estimate | Std. Error | t-value | p-value |
|-----------------------------|----------|------------|---------|---------|
| intercept                   | 0.7790   | 0.0106     | 73.4636 | <0.0001 |
| type = PEN-                 | 0.9446   | 0.0158     | 59.6832 | <0.0001 |

| B. smooth terms             | edf      | Ref.df    | F-value | p-value |
|-----------------------------|----------|-----------|---------|---------|
| log base freq, type != PEN- | 3.7165   | 3.9521    | 982.0015| <0.0001 |
| log base freq, type = PEN-  | 1.0003   | 1.0005    | 512.5895| <0.0001 |

Fig. 5  Partial effects for verb family size regressed on centered log base frequency, for morphological families without nouns with PEN- but possibly including nouns with PE- (left panel) and for morphological families including derived nouns with PEN- (right panel).

0.0001). These overrepresentations are indicated by the residuals (Zeileis et al. 2007). The right panel concerns monomorphemic base words for which the verb family size is zero. Again, we see that base words that have PEN- in their verb family are overrepresented for verbs ($\chi^2_{(4)} = 288.58, p < 0.0001$). No such overrepresentation is visible for PE-. Whereas the literature on PE- and PEN- holds that PEN- is derived from verbs with MEN-, our corpus data indicate that PEN- actually can attach to simple words that do not have a corresponding verb with MEN-, even though the total number of instances is small (45). It is possible that the relevant MEN- verbs are in use in the language, but not attested in our corpus. Alternatively, it is conceivable that these MEN- verbs only have a virtual existence as possible words.

We have seen that PEN- is more productive than PE- and more tightly integrated into the verbal system. This raises the question of whether the reduced productivity of PE- might be due to reduced productivity of the verbal prefix BER-. Indeed, verbs with MEN- are more productive overall than verbs with BER- (2714704 tokens with MEN- vs. 801052 tokens with BER-, 5174 types with MEN- vs. 2869 types with BER-, and 996 hapax legomena with MEN- vs. 760 hapax legomena with BER-); see
The Indonesian prefixes \textit{PE-} and \textit{PEN-}.

![Fig. 6](image)

Left panel: mosaic plot for the type counts of verbs derived from monomorphemic words cross-classified by the word category of the monomorphemic word and the presence of \textit{PE-} or \textit{PEN-} in its verb family. Right panel: corresponding mosaic plot for the type counts of monomorphemic words that do not have any derived verbs attested in the corpus. The colour coding represents the Pearson residuals, which clarify where the observed counts are greater (blue) or smaller (red) than the expected values (Color figure online).

![Table 12](image)

| Prefix | Tokens | Types | Hapaxes |
|--------|--------|-------|---------|
| MEN-: menge- | 1704 | 26 | 4 |
| MEN-: meny- | 187756 | 519 | 91 |
| MEN-: me- | 538078 | 1074 | 173 |
| MEN-: mem- | 558348 | 977 | 246 |
| MEN-: men- | 706181 | 1476 | 190 |
| MEN-: meng- | 722637 | 1102 | 292 |
| BER- | 801052 | 2869 | 760 |

also Table 12 and the rank-frequency plot for \textit{BER-} and \textit{MEN-} in the left panel of Fig. 7. However, when considering the allomorphs of \textit{MEN-} separately, it turns out that \textit{BER-} is more productive than any of these allomorphs, as shown in the right panel of Fig. 7. Although \textit{BER-} is more productive than any of the allomorphs of \textit{MEN-}, it is not the case that \textit{PE-} is proportionally more productive than any of the allomorphs of \textit{PEN-}. It follows that the modest productivity of \textit{PE-} is not a straightforward consequence of the lack of productivity of \textit{BER-}. This conclusion receives further support from the presence of a significant correlation between the frequency of the \textit{MEN-} base and the \textit{PEN-} nominalization ($r_s = 0.4397$, $p < 0.0001$) and the absence of such a correlation for \textit{BER-} and \textit{PE-} ($r_s = 0.1908$, $p = 0.1711$).

### 5 General discussion

We have presented a quantitative investigation of the use of two nominalizing prefixes of Indonesian: \textit{PE-} and \textit{PEN-}. Although quite similar in form, nouns with \textit{PE-} are described by literature as derived from verbs with the prefix \textit{BER-}. Conversely, nouns
with \textit{PEN-} typically originate from verbs with the prefix \textit{MEN-}, and show the same allomorphy in the same conditioning contexts as these prefixed verbs. In this paper, we addressed three questions. First, do \textit{PE-} and \textit{PEN-} differ with respect to their degree of productivity? Second, how does their productivity relate paradigmatically to the productivity of their base words? Third, given the similarity in form of \textit{PE-} and \textit{PEN-}, should they be taken to be allomorphs? To answer these questions, we examined the use of these nominalizations and their base words in a corpus of written Indonesian.

With regards to their productivity, \textit{PEN-} is clearly more productive than \textit{PE-} by any measure of productivity. In fact, \textit{PE-} is less productive than any of the allomorphs of \textit{PEN-}, with as only exception the allomorph \textit{PEN\_penge-}, for which only 18 words are attested. \textit{PEN-} is productive for agents and instruments, whereas \textit{PE-} is productive for agent nouns and to some small extent for patient nouns. Nouns with \textit{PE-} and \textit{PEN-} reveal the same productivity paradox that was reported by Krott et al. (1999) for derivation and compounding. Krott et al. observed that less productive morphological categories are used more intensively as input for further word formation. In our data, we likewise find that the less productive prefix, \textit{PE-}, appears with more variants compared to \textit{PEN-}.

Whereas words with \textit{PE-} are more readily accessible for further inflection compared to \textit{PEN-} (see Fig. 4), words with \textit{PEN-} emerge as paradigmatically more entrenched. Verbs to which \textit{PEN-} attaches tend to allow for more verbal affixation than is the case for verbs to which \textit{PE-} attaches (see Fig. 5). Furthermore, the productivity of the allomorphs of \textit{PEN-} mirrors the productivity of the allomorphs of their base words with \textit{MEN-} (see Fig. 2). The proportionalities that govern the types and hapaxes of the allomorphs of \textit{MEN-} and \textit{PEN-} does not extend to \textit{BER-} and \textit{PE-}. In fact, \textit{PE-} is surprisingly uncommon with base verbs with \textit{BER-}, which is not what standard descriptions in the literature—\textit{PEN-} is derived from \textit{MEN-}, \textit{PE-} is derived from \textit{BER-} (Chaer 2008; Ramlan 2009; Ermanto 2016; Sneddon et al. 2010; Putrayasa 2008; Darwowidjojo 1983; Benjamin 2009) — would lead one to expect.
It is well known that the productivity of an affix can vary depending on the structure of its base words (Aronoff 1976; Baayen and Renouf 1996). Nevertheless, it is surprising to see an almost perfect linear relation between the productivity of the allomorphs of MEN- and the productivity of the allomorphs of PEN-, both with respect to types and with respect to hapax legomena. This linear relationship strongly supports analyses according to which the variant forms of PEN- and MEN- are allomorphs. Our examination of the use of PE- and PEN- in written Indonesian revealed some novel uses that have not been noted in the preceding literature on allomorphy.

This raises the question of whether PE- should be considered to be yet another allomorph of PEN-. Several observations argue against this possibility. First, PE- does not participate in the linear dependence that characterizes the productivity of the allomorphs of MEN- and PEN-. Second, our data indicate that PEN- has a strong preference for verbs as base words, but PE- does not show such a preference. Third, a monomorphemic base word’s verb family tends to be larger when this verb family gives rise to a nominalization with PEN-, but no such tendency is present for PE-. Fourth, the frequencies of words with PEN- enter into a significant correlation with the frequency of the base words, but no such correlation is present for PE-: the formations with PE- have become independent of their base words. Finally, PE- is proportionally overrepresented for patient nouns, whereas PEN- creates primarily instruments in addition to agents.

That allomorphy is to some extent a matter of degree is well known (Baayen et al. 2013; Endresen 2014). Obviously, PE- is highly similar in form to PEN-, in fact, it is identical to one of its allomorphs (although it is possible that phonetically the two are different, see Plag et al. (2017) for durational differences between the realisations of English -s depending on the semantics functions expressed). Yet, even though PE- and PEN- are largely in complementary distribution, they differ substantially in their productivity, both quantitatively and qualitatively, as well as in their entrenchment in the verbal system of Indonesian.

Acknowledgement This study was funded by Indonesia Endowment Fund for Education (Lembaga Pengelola Dana Pendidikan / LPDP) (No. PRJ-1610/LPDP/2015).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References
Alber, B. (2011). Studies on German-language islands (pp. 33–64). Amsterdam: John Benjamins Publishing Company. chap Past participles in Mòcheno: Allomorphy, alignment and the distribution of obstruents, 123
Alwi, H. (2012). Kamus Besar Bahasa Indonesia, 4th edn. Jakarta: Gramedia Pustaka Utama.
Arka, I. W., Dalrymple, M., Mistica, M., & Mofu, S. (2009). A linguistic and computational morphosyntactic analysis for the applicative -i in Indonesian. In M. Butt & T. H. King (Eds.), International Lexical Functional Grammar Conference (LFG), CSLI Publications (pp. 85–105).
Aronoff, M. (1976). Word formation in generative grammar. Cambridge, Mass: MIT Press.
Aronoff, M., & Anshen, F. (2017). The handbook of morphology (pp. 237–247). Hoboken: John Wiley & Sons, Inc. chap Morphology and the lexicon: lexicalization and productivity.
Baayen, R. (2009). *Corpus linguistics. An international handbook* (pp. 900–919). Berlin: De Gruyter. 

Baayen, R. H., & Renouf, A. (1996). Chronicling the times: Productive lexical innovations in an English newspaper. *Language, 72*, 69–96.

Baayen, R., Janda, L. A., Nessel, T., Dickey, S., Endresen, A., & Makarova, A. (2013). Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian Linguistics, 37*, 253–291.

Benjamin, G. (2009). Affixes, Austronesian and iconicity in Malay. *Bijdragen tot de Taal. Land- en Volkenkunde, 165*(2–3), 291–323.

Bloomfield, L. (1933). *Language*. London: George Allen & Unwin Ltd.

Booij, G. E. (1986). Form and meaning in morphology: the case of Dutch agent nouns. *Linguistics, 24*, 503–517.

Booij, G. E. (1996). Inherent versus contextual inflection and the split morphology hypothesis. In G. E. Booij & M. Jv (Eds.), *Yearbook of morphology 1995* (pp. 1–16). Dordrecht: Kluwer Academic Publishers.

Booij, G. (2010). *Construction morphology*. Oxford: OUP.

Cook, R., & Lieber, R. (2004). On the paradigmatic nature of affixal semantics in English and Dutch. *Linguistics, 42*, 327–357.

Chaer, A. (2008). *Morfologi Bahasa Indonesia (Pendekatan Proses)*. Jakarta: PT Rineka Cipta.

Dalrymple, M., & Moof, S. (2012). Plural semantics, reduplication, and numeral modification in Indonesian. *Journal of Semantics, 29*(2), 229–260. [https://doi.org/10.1093/jos/ffr015].

Darwowidjojo, S. (1983). *Some aspects of Indonesian linguistics*. Jakarta: Djambatan.

Denistia, K. (2018). Revisiting the Indonesian prefixes PEN-, PE2-, and PER-. *Linguisitik Indonesia, 36*(2), 145–159.

Dressler, W. (1989). Prototypical differences between inflection and derivation. *Zeitschrift für sprachwissenschafter und kommunikationsforschung, 42*, 3–10.

Endresen, A. (2014). *Non-standard allomorphy in Russian prefixes: Corpus, experimental, and statistical exploration*. PhD thesis, Faculty of Humanities, Social Sciences and Education, The Artic University of Norway.

Ermano (2016). *Morfologi Afiksasi Bahasa Indonesia Masa Kini: Tinjauan dari Morfologi Derivasi dan Infleksi*. Jakarta: Kencana.

Fortin, C. R. (2006). Reconciling meng- and NP movement in Indonesian. *Berkeley Linguistics Society and the Linguistic Society of America, 2*, 47–58.

Goldhahn, D., Eckart, T., & 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* (pp. 1799–1802).

Hidajat, L. (2014). A distributed morphology analysis of Indonesian ke-/l-an verbs. *Linguisitik Indonesia, 32*(1), 11–31.

Kridalaksana, H. (2007). *Kelas Kata dalam Bahasa Indonesia* (2nd ed.). Jakarta: Gramedia Pustaka Utama.

Kridalaksana, H. (2008). *Kamus linguistik* (4th ed.). Jakarta: PT Gramedia Pustaka Utama.

Kroeger, P. R. (2007). Architectures, rules, and preferences: Variations on themes of Joan Bresnan. In CSLI lecture notes (Vol. 184, pp. 229–251). Stanford, California: CSLI Publications. chap Morphosyntactic vs. morphosemantic functions of Indonesian –kan

Krott, A., Robert, S., & Baayen, R. (1999). Complex words in complex words. *Linguistics, 37*, 905–926.

Krott, A., Robert, S., & Baayen, R. (2009). Semantic influence on linkers in Dutch noun-noun compounds. *Folia Linguistica, 36*, 7–22.

Larasati, S., Kuboń, V., & Zeman, D. (2011). Indonesian morphology tool MorphInd: Towards an Indonesian corpus. In C. M & M. P (Eds.), *Systems and frameworks for computational morphology* (Vol. 100, pp. 119–129). Berlin: Springer.

Marle, J. (1986). The domain hypothesis: the study of rival morphological processes. *Linguistics, 24*, 601–627.

Nuriah, Z. (2004). The relation of verbal Indonesian affixes men- and -kan with argument structure. Master’s thesis, Utrecht University, Netherland.

Peters, P. (2004). *The Cambridge guide to English usage*. Cambridge: Cambridge University Press.

Plag, I. (2006). The variability of compound stress in English: structural, semantic and analogical factors. *English Language and Linguistics, 10*(1), 143–172.

Plag, I., Homann, J., & Kunter, G. (2017). Homophony and morphology: The acoustics of word-final S in English. *Journal of Linguistics, 53*(1), 181–216.
The Indonesian prefixes PE- and PEN-

Putrayasa, I. B. (2008). Kajian Morfologi: Bentuk Derivasional dan Infleksional. Bandung. PT Refika Aditama.

R Team DC (2008). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org. ISBN 3-900051-07-0.

R Team S (2015). RStudio: integrated development for R. RStudio. Boston, MA: RStudio http://www.rstudio.com/.

Rafferty, E. (2002). Reduplication of nouns and adjectives in Indonesian. Papers from the Tenth Annual Meeting of the Southeast Asian Linguistics Society (pp. 317–332).

Ramlan, M. (2009). Morfologi: Suatu Tinjauan Deskriptif. Yogyakarta: CV Karyono.

Schreuder, R., Neijt, A., Van der Weide, F., & Baayen, R. H. (1998). Regular plurals in Dutch compounds: linking graphemes or morphemes? Language and cognitive processes, 13, 551–573.

Sneddon, J. N., Adelaar, A., Djenar, D. N., & Ewing, M. C. (2010). Indonesian: a comprehensive grammar (2nd ed.). New York: Routledge.

Stekauer, P. (2014). The Oxford handbook of derivational morphology (pp. 354–369). Oxford: Oxford University Press. chap Derivational paradigms.

Sudaryanto (1993). Metode dan aneka teknik analisis Bahasa: Pengantar Penelitian Wahana Kebudayaan secara linguistis. Yogyakarta: Duta Wacana University Press.

Sugerman (2016). Morfologi Bahasa Indonesia: Kajian ke Arah Linguistik Deskriptif. Yogyakarta: Penerbit Ombak.

Sukarno (2017). The behaviours of the general nasal /N/ in Indonesian active prefixed verbs. International Journal of Language and Linguistics, 4(2), 48–52.

Sutanto, I. (2002). Verba berkata dasar sama dengan gabungan afiks men-i atau men-kan. Makara, Sosial-Humaniora, 6(2), 82–87.

Tomasowa, F. H. (2007). The reflective experiential aspect of meaning of the affix -i in Indonesian. Linguistik Indonesia, 25(2), 83–96.

Wood, S. (2006). Generalized Additive Models: An Introduction with R. Boca Raton: Chapman and Hall/CRC.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B), 1(73), 3–36.

Zeileis, A., Meyer, D., & Hornik, K. (2007). Residual-based shadings in visualizing (conditional) independence. Journal of Computational and Graphical Statistics, 16(3), 507–525.

Zipf, G. (1935). The psycho-biology of language. Boston: Houghton Mifflin.

Zipf, G. (1949). Human behaviour and the principle of the least effort. An introduction to human ecology. New York: Hafner.