Journal of Language and Literature

Vol. 23 No. 1, April 2023, pp. 13 – 24

DOI: 10.24071/joll.v23i1.4967

Available at https://e-journal.usd.ac.id/index.php/JOLL/index

Databases on the Indonesian Prefixes *PE-* and *PEN-*

Karlina Denistia
karlinadenistia@staff.uns.ac.id
English Diploma Program, Vocational School, Universitas Sebelas Maret, INDONESIA

Abstract

This paper provides the theoretical grounding in constituting databases related to PE- and PEN-, two Indonesian nominalizing prefixes, which have various meanings (e.g., patient, agent, or instrument). The first database contains the words with PE- and PEN- whereas the second database provides the cosine similarity between two words of interest. Using a written Indonesian corpus as the primary source (Leipzig Corpora Collection), the databases contain the following information: PE- or PEN- prefixes, allomorph of PEN-, base word, semantics role, morphological variation, cosine similarity, as well as the word frequency. Furthermore, this paper elaborates the theoretical consideration on how each information was cultivated. In building the databases, Indonesian morphological parser and Word to Vector were used to analyze the Indonesian morphological status and to put the words in the corpus into a vector. In addition, manual verification for the data against the Indonesian comprehensive dictionary was also conducted. In the end, the databases are available for free so that the data could be used as materials for a corpus-based analysis on Indonesian morphology. This research shed light to a careful and thorough classification of the open-access databases of PE- and PEN- from their allomorphs, base word, semantics role, and morphological variation. The information provided in this article is hoped to be contributive in Indonesian morphology specifically, and other linguistics fields (e.g., corpus linguistics and quantitative linguistics) in general.

Article information

Received:
5 May 2022

Revised:
25 November 2022

Accepted:
3 December 2022

***Keywords**: corpus data; morphology; prefixes; cosine similarity*

Introduction

PEN- and *PE-* are two nominalizing prefixes to create an agent, an instrument, or a patient. Several studies related to the prefixes' form, meaning and their corresponding verbs have been conducted to investigate *PEN-* and *PE-* (Dardjowidjojo, 1983; Ramlan, 1985; Chaer, 2008; Putrayasa, 2008; Sneddon et al., 2010; Subroto, 2012; Ermanto, 2016; Sugerman, 2016). *PEN-*, the first prefix, derives nouns from a process of affix substitution with *MEN*- verbal prefix (e.g., *pembaca* 'writer'-*membaca* 'to read'). *PE*-, the second prefix, derives nouns from a process of affix substitution with *ber*- or *di*- verbal prefixes (e.g., *pelari* 'runner'-*berlari* 'to run' and *pesapa* 'addressee'-*disapa* 'to be addressed').

From the semantics perspective, both forms might occur in a similar semantics role (Sneddon et al., 2010). *PEN*- expresses agent, instrument, or causer. For instance, from the base word *kasih* 'to love' an agent *pengasih* 'lover' is derived, *pemotong* 'cutter' is derived from *potong* 'to cut', as well as from the base word *sakit* 'to be sick' becomes a causer *penyakit* 'disease'. Words with *PE*-, meanwhile, express patient, agent, or instrument (e.g., *sapa* 'to address'-*pesapa* 'addressee', *lari* 'to run'-*pelari* `runner', *pekasih* 'love poison).

Nasalization in *PEN*-, denoted by 'N', shows that it has five nasalized allomorphs (e.g., *PEN*$_{pen}$-, *PEN*$_{pem}$-, *PEN*$_{peng}$-, *PEN*$_{peny}$-, *PEN*$_{penge}$-). There is only one allomorph that does not follow the nasalization rule, *PEN*$_{pe}$-, which is described as very similar to the invariant *PE*-. As a result, non-native Indonesian may find difficulty to differentiate *PE*- and *PEN*- as one of *PEN*- allomorph occasionally appears in the same phonological environment (see Table 1). For example, *pelari* 'runner' is *PE*-, whereas *pelukis* 'painter' is *PEN*- although both proceed a stem initialized by the lateral liquid /l/. The only way to differentiate *PEN*- and *PE*- in this circumstance is by relating them to the corresponding verb.

The overlapping issue on these two prefixes is not yet well addressed until now. What makes it more difficult to distinguish *PE*- and *PEN*- is because there has not been a consensus whether these formations are derived from one or two prefixes (Denistia, 2018). What might be the reason of this inconclusive finding of *PE*- and *PEN*- is due to a few numbers of observations. Therefore, a set of databases are needed to explore this phenomenon from the quantitative perspective.

Table 1. Words with PE- and PEN- that have similar phonological condition

Word	Prefix	Noun Translation	*PEN*- Allomorph	Base Word	Base Translation	Base Word Class	Semantic Role
pelari	*PE*-	runner		lari	to run	v	agent
pelukis	*PEN*-	painter	pe	lukis	to paint	v	agent
pemusik	*PE*-	musician		musik	music	n	agent
pemasak	*PEN*-	cooker	pe	masak	to cook	v	instrument
perenang	*PE*-	swimmer		renang	to swim	v	agent
perokok	*PEN*-	smoker	pe	rokok	cigarette	n	agent
pewisata	*PE*-	traveler		wisata	to travel	v	agent
pewawancara	*PEN*-	interviewer	pe	wawancara	interview	n	agent

Recent studies on these prefixes conducted analyses based on corpus data (Denistia & Baayen, 2019, 2022a, 2022b, Denistia et al., 2022). Their research focused on investigating whether *PE*- and *PEN*- are allomorphs from their productivity, computational learning, and semantics distribution respectively. One of their significant findings concluded that *PE*- and *PEN*- should be treated as two different prefixes due to their different productivity and semantics. *PEN*- is found more productive than *PE*-. In addition, although both *PE*- and *PEN*- creates agents; *PEN*- is productive in creating instruments, while *PE*- is productive in creating patients. Moreover, the number of derived words with *PEN*- (and all of its allomorphs) is linearly dependent on the number of base words for *MEN*- allomorphs. *PE*-, however, is an outlier in the linearity of the base words' productivity. Apart from productivity analysis, using semantics distribution (Mikolov et al., 2013), Denistia et al. (2022) measured the similarity of all possible combination between *PE*- and *PEN*-. They found that *PE*- and *PEN*- are semantically discriminable. *PE*- and *PEN*- cosine similarity is significantly different only across prefixes. Furthermore, compared to derived words with *PEN*-, words starting with *PE*- have meanings that are more similar to their noun bases.

This paper provides a detailed explanation of the materials and database used in Denistia & Baayen (2019) and Denistia et al. (2022). Theoretical grounding on how the information in database were classified (e.g., the classification of *PE*- and *PEN*-, allomorph of *PEN*-, semantics role, cosine similarity, tokens frequency in the corpus) is described. The tools used to generate two database of *PE*- and *PEN*- are also elaborated in this paper. The information and explanation provided in this paper are structured in a way that I hope to be generally contributive in both corpus and quantitative linguistics analysis.

In what follows, I first introduce the main corpus and tools. In the next section, I present the databases. Finally, I conclude the study in the final section. Along with this paper, two databases are made available for public and can be downloaded at http://bit.ly/PePeNProductivity and http://bit.ly/PePeNSemVector.

Methodology

Leipzig Corpora Collection

The Leipzig Corpora Collection corpus, which includes a range of Indonesian textual registers from 2008 to 2012, including newspapers, the web, and Wikipedia (Goldhahn et al., 2012), was used to create the PePeN Database. This corpus contains 36.608.669 word-tokens that belong to 112.025 different word types and appear in 2.759.800 sentences.

Started by the *Projekt Deutscher Wortschatz* since 15 years ago, now, the Leipzig Corpora Collection has developed into 136 monolingual corpora including Indonesian (Goldhahn et al., 2012; Quasthoff et al., 2006). It uses available online newspapers to crawl as a method for gathering text data [http://www.abyznewslinks.com]. In addition, it uses a framework for parallel Web crawling utilizing http://www.httrack.com as the Web site copier. Another way that was conducted to collect the corpus is by crawling the World Wide Web randomly, utilizing FindLinks [http://wortschatz.uni-leipzig.de/findlinks/] (Heyer & Quasthoff, 2004). Besides, UDHR [http://www.ohchr.org] and Wikipedia [http://sourceforge.net/projects/wikiprep/] were also used as its resource, resulting in more texts in various languages that are covered for this corpora. The text data in the corpora has been preprocessed using the HTML-Stripping in order to take the data containing the well-formed sentences, LangSepa created by Pollmächer (2011) so that each language would be clustered separately, and www.sonderzeichen.de to generate the sentence boundary. To sidestep the copyright issue and to make it impossible to recreate the original material, the phrases were jumbled. The Indonesian Leipzig Corpora Collection corpus is made available online at https://corpora.uni-leipzig.de/en?corpusId=ind_mixed_2013.

Indonesian Morphological Parser (MorphInd)

The MorphInd parser (Larasati et al., 2011), which has an overall accuracy of 84.6%, was used to perform morphological analysis on the words in the PePeN Database. It was run in non-compound mode. Before starting the parser, I manually fixed 200 words beginning with PE- or PEN- that had typos (see Table 2 for illustrations) and added the frequency of the typos to the frequency of the words. Additionally, using the dictionary as the gold standard manual verification, MorphInd's recall for detecting *PE*- and *PEN*- was 0.82 and its precision for doing so was 0.98.

The R open-source programming language, version 3.3.3, was used to process the data in R Studio(R Team, 2015). R is an open source that can be downloaded at https://cran.r-project.org for free (available for Windows, Mac, and Linux users).

Table 2. Typo example entries in the database

Word	Translation	Frequency	Typo Revision	Freq of Typo
pelukis	painter	321	pelunis	1
pemusik	musician	208	pemuzik	7
penulis	writer	5312	pemnulis,pemulis,pengnulis,penulia, penulih,penulsi,penults,penulus,peulis	1,1,1,1, 1,1,1,1,1
perokok	smoker	671	peerokok,peroko,perokor	1,1,1

Table 3. The MorphInd parser output examples

Word	Parser	Noun Translation	Allomorph	Base	Base Translation
pencipta	peN+cipta_NSD	creator	pen	cipta	to create
pendaki	peN+daki_NSD	climber	pen	daki	climbing
peninju	peN+tinju_NSD	puncher	pen	tinju	punch
petinju	peN+tinju_NSD	boxer		tinju	boxing
petani	petani_NSD	rice farmer		tani	farming
peternak	peternak_X–	farmer		ternak	cattle
pengelas	peN+kelas_NSD	welder	penge	las	to weld
pengusut	peN+kusut_NSD	investigator	peng	usut	to investigate

Table 3 shows the sample output of MorphInd parser. From Table 3, one can see that MorphInd correctly parses *pencipta*, *pendaki*, and *peninju*. However, MorphInd parser is not accurate in identifying *PE-* in *petinju*, *petani*, and *peternak.* In several instances, MorphInd is unable to accurately detect single-syllable base words. *Pengelas* 'welder', for instance, MorphInd identifies the base word is *kelas* 'classroom', thus the morphological process is [*PENpeng-* + *kelas*]. The base word of *pengelas* 'welder' is *las* 'weld' and thus the morphological analysis is [*PENpenge-* + *las*]. Also, the base identified by the parser is not accurate as in *pengusut* 'investigator'. *Pengusut* 'investigator' is supposedly formed from the base word *usut* 'to investigate', but MorphInd identifies its base as *kusut* 'tangled'. Thus, the correct parsing is [*PENpeng-* + *usut*] and not [*PENpeng-* + *kusut*]. Due to some misidentification, the online version of *Kamus Besar Bahasa Indonesia*, a comprehensive dictionary of Indonesian, (http://kbbi.kemdikbud.go.id; viewed on June 2016), which was published in 2012, and comprises more than 90,000 lemmas, was used to manually examine and correct MorphInd output as needed (Alwi, 2012).

Word to Vector

I lemmatized the Leipzig Corpora Collection corpus based on MorphInd's morphological analyses output. Prior to the lemmatization, all the word in the corpus were lower-cased, numbers were excluded, punctuation marks and 15 highest frequency stop words were removed. I removed stop words *itu* 'that', *ini* 'this, *dan* 'and', *yang* 'which', *pada* 'of', *di* 'in', *dengan* 'with', *akan* 'will', *juga* 'also', *dari* 'from', *untuk* 'to/for', *dalam* 'inside', *ke* 'to', *karena* 'because', and *tidak* 'not'. MorphInd lemmatizes *anti-*, *pra-*, *pasca-, non-*, *ku-* 'I', *-ku* 'my', *kau-* 'you', *-mu* 'your', *-nya* 'his/her/its', *se-* 'one', *per-* 'per', and particles (e.g. *-lah* to show emphasize, *-kah* to ask question) as bound morphemes. The suffix *-nya* was marked to indicate its function as a question word by *nya-WH*. However, *antar* was not separated from its base because it has a different reading when this word occurs in a form of a simple word (e.g., *antar paket itu* 'pick that package up' versus *antaragama* 'among religions`). Finally, hyphenated words were kept in the form of reduplication as the way they are (e.g., iteration, intensification, or plurality; Rafferty (2002), Chaer (2008), Dalrymple & Mofu (2012), Sugerman (2016)).

Word to Vector was made used to convert all the lemmatized words in the corpus into a vector. Each word in the corpus was encapsulated in high-dimensional vectors so that a vector will represent a word (Turney & Pantel (2010)). Cosine similarity, which is length-normalized and is equal to the inner product of the vectors, was used to calculate the degree of semantic similarity between two lemmas, based on the distributional information of the words (their co-occurrences with other words in huge corpora). The similarity of the cosine of the angle θ is cosine similarity between $\vec{V}$ and $\vec{W}$.

$$sim(V,W) = cos(\theta) = \frac{\mathbf{V}.\mathbf{W}}{\|\mathbf{V}\|\|\mathbf{W}\|} = \frac{\sum_{i=1}^{n} V_i W_i}{\sqrt{\sum_{i=1}^{n} V_i^2}\sqrt{\sum_{i=1}^{n} W_i^2}}$$

In the PePeN CosSim Database, the results of computing the cosine similarity value for each conceivable pair combination of words from the set of *PE*-, *PEN*-, and their base words were stored. Lemma1, Lemma2, Cosine similarity (the cosine similarity value between Lemma 1 and Lemma 2), and Derived-Base Cosine Similarity (cosine similarity measure of the derived word with its base word) are all included in the database. Finally, I collected a total of 358224 permutation of derived words with *PEN*- and 59810 permutations of derived words with *PE*- together with their cosine similarity to their base words (see Table 4 for list of example entries of this database). Words with a token frequency less than 5 were not included in this database.

Table 4. PePeN CosSim Database's examples of entries

Lemma1	L1 English	Lemma 2	L2 English	CosSim L1L2	Prefix L1	Prefix L2	Base L1	B1 English	Base L2	B2 English	Cossim L1-B1	Cossim L2-B2
pencipta	creator	pelatih	trainer	0.022	peN	peN	cipta	to create	latih	to train	0.358	0.192
pencipta	creator	pemilih	voter	-0.111	peN	peN	cipta	to create	pilih	to vote	0.358	0.364
pencipta	creator	pencari	seeker	0.092	peN	peN	cipta	to create	cari	to search	0.358	0.147
peternak	farmer	pengupas	peeler	-0.03	pe	peN	ternak	to farm	kupas	to peel	0.717	0.212
peternak	farmer	penyapu	sweeper	0.03	pe	peN	ternak	to farm	sapu	broom	0.717	0.374
peternak	farmer	penyemprot	sprayer	0.005	pe	peN	ternak	to farm	semprot	to spray	0.717	0.646

Results and Discussion

PePeN Database includes a total of 3090 words; 2818 words with *PEN*-, 267 words with *PE*-, and 4 words with the unproductive variant *PER*-, Benjamin (2009). The latest prefix is not discussed in this paper. For the sake of the quantitative analysis, both PePeN Database and PePeN CosSim Database provide the information on how many times the words with *PE*- or *PEN*- and their base words occur in the corpus; usually called as 'token frequency' (see Table 5). The mentioned frequencies are the word's overall frequency and are not segmented by meaning.

Table 5. Sample entries of PePeN Database

Noun Word	Noun Translation	Frequency	PE-	Allomorph	Base Word	Base Translation	Base Word Class	Base Frequency	Semantic Role
pelari	runner	358	T		lari	to run	v	2312	agent
pelukis	painter	321	F	pe	lukis	to paint	v	282	agent
pemasak	cooker	6	F	pe	masak	to cook	v	1070	instrument
pemusik	musician	208	T		musik	music	n	9799	agent
perenang	swimmer	296	T		renang	to swim	v	821	agent
perokok	smoker	671	F	pe	rokok	cigarette	n	3619	agent
pewawancara	interviewer	101	F	pe	wawancara	interview	n	3015	agent
pewisata	traveler	1	T		wisata	to travel	v	7371	agent

Classifying PE- and PEN-

There are two ways to differentiate *PE-* and *PEN-*. The first one is by applying the phonological condition on *PEN-* and its six allomorphs: *PENpen-*, *PENpeng-*, *PENpem-*, *PENpeny-*, *PENpe-*, and *PENpenge-*. The phonological context influences the nasal allomorphy of *PEN-*. The phonological conditioning of *PEN-* allomorphs is summarized by Ramlan (1985), Sugerman (2016), and Sukarno (2017) as follows:

1. *-N* is lost with base words initialized by /ny/, /w/, /r/, /m/, /n/, /ng/, or /l/
   a. *peN-* + *nyanyi* 'to sing' to be *penyanyi* 'singer'
   b. *peN-* + *wangi* 'good smell' to be *pewangi* 'instrument to give a good smell'
   c. *peN-* + *rusak* 'broken' to be *perusak* 'destroyer'
   d. *peN-* + *mabuk* 'drunk' to be *pemabuk* 'who gets drunk'
   e. *peN-* + *lukis* 'to paint' to be *pelukis* 'painter'
2. *-N* becomes *-n* with base words initialized by /t/, /d/, /j/, /c/, /sy/, or /z/
   a. *peN-* + *tulis* 'to write' to be *penulis* 'writer'
   b. *peN-* + *daki* 'to climb' to be *pendaki* 'climber'
   c. *peN-* + *jelajah* 'to explore' to be *penjelajah* 'explorer'
   d. *peN-* + *cuci* 'to wash' to be *pencuci* 'instrument to wash/agent who wash'
3. *-N* becomes *-ng* with base words initialized by a vowel or /k/, /h/, /g/, or /kh/
   a. *peN-* + *ingat* 'to remember' to be *pengingat* 'reminder'
   b. *peN-* + *ganti* 'replacement' to be *pengganti* 'who/which replaces'
   c. *peN-* + *halang* 'block' to be *penghalang* 'barrier'
   d. *peN-* + *kuasa* 'power' to be *penguasa* 'ruler'
4. *-N* becomes *-m* with base words initialized by /p/, /b/, or /f/
   a. *peN-* + *buat* 'to make' to be *pembuat* 'maker'
   b. *peN-* + *picu* 'trigger' to be *pemicu* 'trigger'
   c. *peN-* + *fitnah* 'to sander' to be *pemfitnah* 'slander'
5. *-N* becomes *-ny* with base words initialized by /s/
   a. *peN-* + *saring* 'to filter' to be *penyaring* 'who filters'
6. *penge-* occurs in monosyllabic base words
   a. *peN-* + *cek* 'to check' to be *pengecek* 'checker'

There are some exceptions of these phonological condition given by Sneddon et al. (2010). If the stem is borrowed from other languages, some bases with initial /k/, /s/, /t/, /p/ are not lost. Thus, the derived words as a result of borrowing becomes more accepted as an Indonesian word as in the stem *klasifikasi* 'classification' to be *pengklasifikasi* 'classifier'.

Table 6 shows the second way to distinguish *PE-* and *PEN-,* which is by process of affix substitution. In this case, the prefix *PEN*- changes verbs with the *MEN*- prefix into noun. *MEN*- also has 6 allomorphs (*MENmeng-, MENmen-, MENmem-, MENme-, MENmeny-, and MENmenge-*). Again by affix substitution, the prefix *PE*- creates nouns from verbs with the prefix *BER*- (Dardjowidjojo, 1983; Ramlan, 1985; Putrayasa, 2008; Benjamin, 2009; Sneddon et al., 2010; Tjia, 2015; Ermanto, 2016). However, it should be noted that Ramlan (1985) acknowledged only several verbs with *BER*- correlates to *PE*-.

Table 6. *Examples of the corresponding PEN- with MEN- and PE- with BER-.*

Noun Word	**Noun Translation**	**PE-**	**Base Word**	**Base Translation**	**Corresponding Verb**	**Verb Translation**
pelari	runner	TRUE	lari	to run	berlari	to run
pelukis	painter	FALSE	lukis	to paint	melukis	to paint
pemusik	musician	TRUE	musik	music	bermusik	to play music
pemasak	cooker	FALSE	masak	to cook	memasak	to cook
perenang	swimmer	TRUE	renang	to swim	berenang	to swim
perokok	smoker	FALSE	rokok	cigarette	merokok	to smoke
pewisata	traveler	TRUE	wisata	to travel	berwisata	to travel
pewawancara	interviewer	FALSE	wawancara	interview	mewawancarai	to interview

The base words for both the verbs with *MEN*- or *BER*- and their nominalizations with *PEN*- and *PE*- can be nouns, adjectives, and verbs. Verbs with *MEN*-, which ordinarily renders a transitive verb, can be added by *-i* and *-kan* suffixes. The suffixes *-i* and *-kan* typically signify intensification or iteration while also adding a location, a beneficiary, or a causer as a new argument (Arka et al., 2009; Kroeger, 2007; Sneddon et al., 2010; Sutanto, 2002; Tomasowa, 2007). In the same vein, verbs with *BER*-, which has infrequent allomorphs *be*- and *bel*-, essentially express reciprocity, stativity, or reflexivity. *BER*- are found with *-an* or *-kan,* but *PE*- does not combine with the suffixes (Chaer, 2008; Kridalaksana, 2007; Putrayasa, 2008; Ramlan, 1985; Sneddon et al., 2010). The verb structure with *BER-an* and *BER-kan* create respectively reciprocative (e.g., *peluk* 'to hug'-*berpelukan* 'to hug each other') or 'having X' (e.g., *dasar* 'base'-*berdasarkan* 'based on') (Sneddon et al., 2010).

Although derived nouns with *MEN*- can be further modified with the suffixes *-i* or *-kan*, derived nouns with *PEN*- do not. Nevertheless, the verbs with *MEN-/-i* or *MEN-/-kan* affixes may have semantics that are similar to the derived nouns. For instance, *pewawancara*, 'interviewer', is related to *mewawancarai* 'to interview someone'. Also, although the corresponding verbs with *BER*- can be extended by *-an* or *-kan* suffixes, derived nouns with *PE*- do not carry the suffixes.

Base Word of PE- and *PEN-*

Indonesian nouns, verbs, and adjectives can be monomorphemic or polymorphemic. Kridalaksana (2007) explained that nouns are classified into abstract or concrete, animate or inanimate, countable or uncountable, as well as collective or non-collective. In term of verbs, they can be characterized by adding *dengan* and adjective which function as an adverbial of manner (referring to the *-ly* suffix in English). For instance, *berlari* 'to run' can be modified into *berlari dengan cepat* 'to run fast'; therefore, *berlari* is a verb. Verb formations are classified into transitive or intransitive, active or passive or anti-active or anti-passive, reciprocal or nonreciprocal, reflective or nonreflective, copulative or equative, and performative or constant. With regards to adjectives, they could be indicated by *tidak* 'not' as the negation, premodifiers (e.g., *sangat* 'very', *agak* 'pretty', *lebih* 'more'), and that they could modify nouns. They are classified into predicative or attributive and gradual or nongradual adjectives.

Table 7 shows examples of the base word and base word category in the database. In PePeN Database and PePeN CosSim Database, the dictionary and MorphInd were used to decide what base word category of the *PE-* and *PEN-* nouns. There might be a conflict in determining the base word category between those two tools. Upon that case, I followed the base word category information provided by the Indonesian dictionary. However, in the case where the information on the word category of the base is not provided in the dictionary, I used the MorphInd parser identification. I did not provide a further classification on each type (such as whether the verb is transitive or intransitive, or whether the noun is animate or inanimate).

Table 7. *Examples of PePeN base word and base word category.*

Word	Noun Translation	PE-	Allomorph	Base Word	Base Translation	Base Word Class
pencipta	creator	F	pen	cipta	to create	n
pendaki	climber	F	pen	daki	climbing	v
peninju	puncher	F	pen	tinju	punch	n
petinju	boxer	T		tinju	boxing	n
petani	rice farmer	T		tani	farming	n
peternak	farmer	T		ternak	cattle	n
pengelas	welder	F	penge	las	to weld	n
pengusut	investigator	F	peng	usut	to investigate	v

Semantics Role of PE- and *PEN-*

Manual verification of all *PE-* and *PEN-* words was not doable. Therefore, I did a manual annotation for the semantic role for all derived words with *PE-* and *PEN-* and checked against the usage in the corpus for at least one token, as well as the dictionary (Alwi, 2012). One of the implications of this limitation is that the ambiguity in assigning a semantic role to *PE-* and *PEN-* words which express multiple semantic roles could not be resolved. Thus, it is possible that there are cases for which a semantic role was realized in the corpus with no semantic role registered in the database.

Table 8 shows various readings for *PE-* and *PEN-* formations. As in English, *-er* nominalizations may have a range of semantic roles (e.g., *printer*, which has both an instrument and agent reading) (G. Booij, 2010; G. Booij & Lieber, 2004). I did not distinguish between impersonal agent in this research. The term impersonal agent was introduced by Booij (1986) for 'radio station' of the Dutch word *zender* which also has both an instrumental interpretation, 'transmitter', and an agentive meaning, 'one who sends'. Although it is commonly known that *PEN-* create agents, patients, and instruments (Sneddon et al., 2010), the database contains a small number of instances of causer (e.g., *penyakit* 'disease') and location (e.g., *penghujung* 'the end'). Semantic roles that are not registered in the database may nonetheless be used in the corpus, which is plausible and perhaps likely.

Table 8. *Examples of PePeN semantic role.*

No	Word	Noun Translation	PE-	Base Word	Base Translation	Semantic Role
1	pembanding	who compares	F	banding	to compare	agent
2	pembanding	something to compare	F	banding	to compare	instrument
3	pembanding	something to be compared	F	banding	to compare	patient
4	pesiar	cruise	T	siar	to broadcast	instrument
5	pesiar	traveler	T	siar	to broadcast	agent
6	penyiar	radio announcer	F	siar	to broadcast	agent
7	penyelam	who dives	F	selam	to dive	agent
8	peselam	diver (athlete)	T	selam	to dive	agent
9	pengasih	who loves	F	kasih	love	agent
10	pekasih	love poison	T	kasih	love	instrument
11	penyakit	disease	F	sakit	to be sick	causer
12	pesakit	patient	T	sakit	to be sick	patient
13	penyapa	addressor	F	sapa	to address	agent
14	pesapa	addressee	T	sapa	to address	patient

Words with more than one semantic role have multiple entries in the database, one row per role (cf. Table 8, rows 1-6). Occasionally do the prefixes *PEN-* and *PE-* attach to the same base word; often, the form with *PE-* alludes to a profession in a semantic sense, whereas the word with *PEN-* does not (cf. Table 8, rows 7 and 8). In some instances, the form with the prefix *PEN-* expresses the agent, causer, or instrument, while the form with the prefix *PE-* expresses the patient or agent (cf. Table 8, rows 9-14).

Morphological Variation of PE- and *PEN-*

In Indonesian, there are bound morphs for possession of nouns, (first *-ku*, second *-mu*, and third person singular *-nya*), subject (first *ku-* and second person singular *kau-*) and object (first *-ku*, second *-mu*, and third person singular *-nya*) marking on verbs (Sneddon et al., 2010). These bound morphemes fulfill the contextual inflection, an inflection which is not dictated by syntax, proposed by Booij (1996). Additionally, there are two suffixes that can be added to verbs or nouns to indicate emphasize *(-lah)* or query *(-kah)*. Clitics are the term given to bound morphemes, which are phonologically condensed versions of free pronouns (Kridalaksana, 2008). Therefore, I will refer to these morphs as inflectional because they alter existing words rather than creating new ones, much to how English adverbs modify verbs.

Reduplication creates different semantic functions on verbs and adjectives, including intensification and iteration respectively, as well as to convey the plural for nouns. (Rafferty, 2002; Chaer, 2008; Dalrymple & Mofu, 2012; Sugerman, 2016). According to Booij (1996), reduplication as well as *-lah, -kah* and *-pun* instantiate inherent inflection. Although it may have syntactic relevance, inherent inflection is the kind of inflection that is not required by the syntactic context. In the database, reduplication is more like syntactic modification than to word formation. Hence, reduplicated forms were classified as inflectional because their semantics are still related to a plurality (e.g., intensifier or iterative). Some examples on the inflection are listed in Table 9.

Table 9. *Examples of inflection in PePeN database.*

Word	Translation	PE-	Allomorph	Base Word	Base Word Class	Inflection
pemerintahnya	his/her/its government	F	pem	perintah	n	Possession
pemerintahlah	government (emphasize)	F	pem	perintah	n	Particle
pemerintahpun	government (emphasize)	F	pem	perintah	n	Particle
pemerintah-pemerintah	government (plural)	F	pem	perintah	n	Reduplication
pelarinya	runners	T		lari	v	Possession
pelari-pelari	his/her/its runner	T		lari	v	Reduplication

Conclusion

Given the fact that there have been many qualitative descriptive about the Indonesian *PE-* and *PEN-* prefixes, some questions on how to discriminate them remain unanswered. *PEN-* has 5 allomorphs: *PENpen-, PENpem-, PENpeng-, PENpeny-, PENpenge-* that follow the nasalization rule and there is only one allomorph, *PENpe-*, that is not nasalized. A case arises when these two are in a contest, appearing in the same phonological environment. Moreover, there has been an inconclusive agreement among theories whether these nominalizing prefixes are one or two independent formations.

This paper provides detailed information on two databases, namely PePeN Database and PePeN CosSim Database, as the contribution to a quantitative approach for Indonesian linguistics. Taken from *Leipzig Corpora Collection*, I used several tools and programming language to classify the database from its prefix, allomorph, base word, base word class, semantics role, inflection, as well as cosine similarity. These databases could be used to conduct a further study on *PE-* and *PEN-* formations.

This study, however, is limited to only two nominalizing prefixes, *PE-* and *PEN-*. Indonesian has other nominalizing affixes (e.g., *-an* as in *luar* `outside' to *luaran* `outcome', *Makmur* `prosperous' to *ke-/-an* as in *kemakmuran* `prosperity'). In addition, *PEN-* could also attach to the suffix *-an* to form *peN-/-an* circumfixes (e.g., *tinggal* `stay' to *peninggalan* `heritance'). Another noun could also be derived from *per-/-an,* such as *unbah* `to change' to *perubahan* `a change'. Therfore, some explanation on databases of other nominalizing affixes would be useful for further research.

Acknowledgment

Lembaga Pengelola Dana Pendidikan, the endowment fund for education in Indonesia, provided funding for this study under grant number PRJ-1610/LPDP/2015.

References

Alwi, H. (2012). *Kamus Besar Bahasa Indonesia* (fourth). 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) (pp. 85–105). CSLI Publications.

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

Booij, G. (2010). Construction morphology. *Language and Linguistics Compass*, *4*(7), pp. 543–555.

Booij, G. E. (1986). Form and Meaning in Morphology: The Case of Dutch Agent Nouns. *Linguistics*, *24*, pp. 503–517.

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

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

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

Dalrymple, M., & Mofu, S. (2012). Plural Semantics, Reduplication, and Numeral Modification in Indonesian. *Journal of Semantics*, *29*(2), pp. 229–260. https://doi.org/10.1093/jos/ffr015

Dardjowidjojo, S. (1983). *Some Aspects of Indonesian Linguistics*. Jakarta: Djambatan.

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

Denistia, K., & Baayen, R. H. (2019). The Indonesian prefixes PE- and PEN-: A study in productivity and allomorphy. *Morphology*, pp. 1–23. https://doi.org/10.1007/s11525-019-09340-7

Denistia, K. & Baayen, R. H. (2022a). Affix substitution in Indonesian: A computational modelling approach. *Linguistics.* https://doi.org/10.1515/ling-2020-0191

Denistia, K., and Baayen, R. H. (2022b). The morphology of Indonesian: Data and quantitative modeling. In Shei, C., and Li, S. (Eds.) *The Routledge Handbook of Asian Linguistics*, 605-634. London: Routledge.

Denistia, K., Shafaei-Bajestan, E., & Baayen, H. (2022). Exploring semantic differences between the Indonesian prefixes PE- and PEN- using a vector space model. *Corpus Linguistics and Linguistic Theory*, *18*(3), pp. 573–598. https://doi.org/10.1515/cllt-2020-0023

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

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

Heyer, G., & Quasthoff, U. (2004). Calculating Communities by Link Analysis of URLs. *Proceedings of IICS-04*.

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

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

Kroeger, P. R. (2007). *Architectures, Rules, and Preferences: Variations on Themes of Joan Bresnan* (A. Zaenen, J. Simpson, T. H. King, G. Jane, J. Maling, & C. Manning, Eds.; pp. 229–251). CSLI Publications.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, pp. 3111–3119.

Pollmächer, J. (2011). *Separierung mit FindLinks gecrawlter Texte nach Sprachen* [Master's Thesis]. University of Leipzig.

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

Quasthoff, U., Richter, M., & Biemann, C. (2006). *Corpus Portal for Search in Monolingual Corpora*. pp. 1799–1802.

R Team, S. (2015). *RStudio: Integrated Development for R. RStudio*. RStudio, Inc. 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. (1985). *Morfologi: Suatu Tinjuan Deskriptif*. Yogyakarta: CV Karyono.

Sneddon, J. N., Adelaar, A., Djenar, D. N., & Ewing, M. C. (2010). *Indonesian: A Comprehensive Grammar* (second). London: Routledge.

Subroto, E. (2012). *Pemerian Morfologi Bahasa Indonesia: Berdasarkan Perspektif Derivasi dan Infleksi Proses Afiksasi*. Surakarta: Yuma Pressino.

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), pp. 48–52.

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

Tjia, J. (2015). Grammatical relations and grammatical categories in Malay: The Indonesian prefix meN- revisited. *Wacana*, *16*(1), pp. 105–132.

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